Prilagodba sadržaja mrežnih stranica za internetske tražilice pomoću strojnoga učenja i obrade prirodnoga jezika

Matošević, Goran

Doctoral thesis / Disertacija

2019

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Organization and Informatics / Sveučilište u Zagrebu, Fakultet organizacije i informatike**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:211:750631

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-29**

Repository / Repozitorij:

Faculty of Organization and Informatics - Digital Repository



Sveučilište u Zagrebu

Fakultet organizacije i informatike

Mr. sc. Goran Matošević

PRILAGODBA SADRŽAJA MREŽNIH STRANICA ZA INTERNETSKE TRAŽILICE POMOĆU STROJNOGA UČENJA I OBRADE PRIRODNOGA JEZIKA

DOKTORSKI RAD

Zagreb, 2019



Sveučilište u Zagrebu

Fakultet organizacije i informatike

Mr. sc. Goran Matošević

PRILAGODBA SADRŽAJA MREŽNIH STRANICA ZA INTERNETSKE TRAŽILICE POMOĆU STROJNOGA UČENJA I OBRADE PRIRODNOGA JEZIKA

DOKTORSKI RAD

Mentori:
prof.dr.sc. Jasminka Dobša, Sveučilište u Zagrebu
prof.dr.sc. Dunja Mladenić, Institut Jožef Stefan, Ljubljana

Zagreb, 2019



University of Zagreb

Faculty of organization and informatics

Goran Matošević

WEB PAGE CONTENT ADJUSTMENT FOR SEARCH ENGINES USING MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING

DOCTORAL THESIS

Supervisors:

Prof. Jasminka Dobša, PhD, University of Zagreb
Prof. Dunja Mladenić, PhD, Jožef Stefan Institute, Ljubljana

Zagreb, 2019

PODACI O DOKTORSKOM RADU

I. AUTOR

Ime i prezime	Goran Matošević
Datum i mjesto rođenja	27.12.1977. Pula, RH
Naziv fakulteta i datum diplomiranja	Ekonomski fakultet Rijeka, 2003 Ekonomski fakultet Zagreb, 2009
Sadašnje zaposlenje	Sveučilište J. Dobrile u Puli

II. DOKTORSKI RAD

Naslov	PRILAGODBA SADRŽAJA MREŽNIH STRANICA ZA INTERNETSKE TRAŽILICE POMOĆU STROJNOGA UČENJA I OBRADE PRIRODNOGA JEZIKA
Broj stranica, slika, tabela, priloga, bibliografskih podataka	141 stranica, 21 slika, 28 tabela, 7 priloga, 90 bibliografskih podataka
Znanstveno područje i polje iz kojeg je postignut akademski stupanj	Informacijske i komunikacijske znanosti
Mentor i voditelj rada	Prof. dr. sc. Jasminka Dobša Prof. dr. sc. Dunja Mladenić
Fakultet na kojem je rad obranjen	Fakultet organizacije i informatike Varaždin
Oznaka i redni broj rada	155

III. OCJENA I OBRANA

Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena tema	16.05.2017.
Datum predaje rada	03.04.2019.
Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena pozitivna ocjena rada	19.09.2019.
Sastav Povjerenstva koje je rad ocijenilo	Prof. dr. sc. Danijel Radošević, predsjednik Izv. prof. dr. sc. Jan Šnajder, član Izv. prof. dr. sc. Markus Schatten, član
Datum obrane	03.10.2019.
Sastav Povjerenstva pred kojim je rad obranjen	Prof. dr. sc. Danijel Radošević, predsjednik Izv. prof. dr. sc. Jan Šnajder, član Izv. prof. dr. sc. Markus Schatten, član
Datum promocije	

INFORMACIJE O MENTORIMA

prof. dr. sc. Jasminka Dobša

Jasminka Dobša rođena je 28. kolovoza 1971. godine u Čakovcu, gdje je završila osnovnu školu i srednju školu matematičko-informatičkog usmjerenja. Diplomirala je 1995. godine na Matematičkom odjelu Prirodoslovno-matematičkog fakulteta u Zagrebu, smjer Matematička statistika i informatika. Iste godine je upisala poslijediplomski studij matematike i zaposlila se na Matematičkom odjelu PMF-a kao znanstveni novak.

Na Fakultetu organizacije i informatike radi od 1997. godine. U zvanje docenta izabrana je 2007. godine, u zvanje izvarednog profesora 2013. godine, a u zvanje redovitog profesora 2019. godine. U okviru prediplomskog i diplomskog studija na Fakultetu organizacije i inoformatike izvodi nastavu na kolegijima Statistika, Statistička analiza podataka, Statističke metode za informatičare i Metode pretraživanja i klasifikacije informacija. U okviru doktorskog studija informatike na Fakultetu organizacije i informatike izvodi kolegije Odabrana poglavlja primjene informacijske tehnologije i Odabrana poglavlja statističkih metoda u informacijskim znanostima.

Magistrirala je 1999. godine na Matematičkom odjelu PMF-a u Zagrebu s temom Oblik atraktora dinamičkog sustava. Doktorirala je 2006. godine na Fakultetu elektrotehnike i računarstva u Zagrebu s temom Dubinska analiza teksta uporabom konceptnog indeksiranja.

Tokom 2003. godine tri je mjeseca boravila na Odjelu za inteligentne sustave Instituta Jožef Štefan u Ljubljani u svrhu znanstvenog usavršavanja. Vodila je dva bilateralna hrvatsko-slovenska projekta: projekt Sustav za automatsku klasifikaciju stranica hrvatskog i slovenskog Interneta (2005./2006), te projekt Reprezentacija velikih podataka (2016-2017.) u suradnji s Institutom Jožef Šštefan u Ljubljani. Uključen je u rad na četiri znanstvena projekta i jednom projektu primjene informacijske tehnologije na nacionalnom nivou te jednom međunarodnom projektu. Koautorica je udžbenika statistike Statistika (deskriptivna i inferencijalna) i vjerojatnost publiciranog 2008. godine.

Aktivno radi kao recenzentica za časopise Journal or Information and Organizational Sciences, Journal of Computing and Information Technology, Environmental Modelling and Software, Journal of Information and Knowledge Management i Informatica (Ljubljana). Od rujna 2014. godine je pridružena urednica časopisa Journal of Information and Knowledge

Management. Dobitnica je nagrade za posebna postignuća u znanstvenom i nastavnom radu Fakulteta organizacije i informatike 2015. godine.

Članica je Hrvatskog biometrijskog društva i Hrvatskog statističkog društva. Jedna je od suosnivača Laboratorija za generativno programiranje i strojno učenje koji djeluje na FOI-u od kraja 2011. godine. Članica je programskih odbora konferencije Central European Conference on Information and Intelligent Systems i konferencije International Statiatical Conference in Croatia.

Područja znanstvenog interesa su joj dubiska analiza podataka s posebnim naglaskom na tekstualne podatke te primijenjena statistika.

Udana je i majka dvoje djece.

prof. dr. sc. Dunja Mladenić

Dunja Mladenić diplomirala je računalstvo i informatiku na Fakultetu za računalstvo i informatiku na Sveučilištu u Ljubljani 1990. god. pod mentorstvom prof.dr. Ivana Bratka. Na istom fakultetu obranila je magisterij (1995.god., mentori prof.dr. Ivan Bratko i prof.dr. Nada Lavrač) i doktorat (1998. god., mentori prof.dr. Ivan Bratko i prof.dr. Tom Mitchell). Njeno područje istraživanja je analiza i razvoja sustava za strojno učenje i otkrivanje znanja iz podataka s posebnim naglaskom na analizi teksta.

Od 1992. god. zaposlena je na Institutu "Jožef Stefan", najprije kao mladi istraživač, od 1998. kao voditelj projekata, od 2007. kao viši znanstvenik djelatnik, a od 2011. kao voditelj Samostalnog laboratorija za umjetnu inteligenciju. Od 2008. dodatno je zaposlena na Međunarodnoj postdiplomskoj školi Jožefa Stefana na kojoj predaje od 2005. god. i nositelj je nekoliko kolegija. Bila je na stručnom osposobljavanju u inozemstvu na Carnegie Mellon University, SAD, 1996-1997 i 2000-2001. Sudjelovala je na više nacionalnim i međunarodnim projektima, posljednih godina kao i voditelj projekta 6.OP (SEKT, PASCAL, ALVIS, CEC-WYS, NeOn, SWING, TAO, IMAGINATION, SMART, WS-DEBATE), 7.OP (XLike, RENDER, PlanetData, ALERT, Multilingual Web, MetaNet, ACTIVE, PASCAL2, GENDERA, ENVISION, SIS-Catalyst, SYMPHONY), H2020 (AquaSmart). Osim toga koordinirala je međunarodni projekt 5OP Sol-Eu-Net (2000-2003). Obnašala je dužnost i financijske tajnice Odsjeka za tehnologije znanja Instituta "Jožef Stefan" (2007-2010).

Prof.dr.sc. Dunja Mladenić je predstavnica Slovenije u strukovnim skupinama Europske komisije (Enwise - promocija znanstvenica u državama srednje i istočne Eurepe te balkanskih država). Članica je Komisije za žene u znanosti pri Ministrstvu obrazovanja, znanost i šport Slovenije (2014 - 2018). Od 2000.god. sudjeluje pri Europskoj komisiji kao evaluator projekata za područje informacijskih tehnologija. Godine 2001. sudjelovala je kao evaluator NSF projeka u SAD-u iz područja informacijskih tehnologija. Od 1996. god. članica je izvršnog odbora Slovenskog društva za umjetnu inteligenciju SLAIS, 2010-2014 je bila predsjednica tog društva. Od 2001. god. članica je nadzornog odbora Slovenskog društva ACM. Članica je programskih odbora više međunarodnih konferencija, suorganizirala je više međunarodnih konferencija, bila je voditelj jedne od glavnih međunarodnih konferencija iz područja strojnog učenja ECML PKDD 2009, i suvoditelj programskog odbora međunarodne konferencije ECML 2007. Bila je članica ECML/PKDD Steering Committee (2007-2012) i jedna od voditelja programskog odbora više glavnih međunarodnih konferencija: ESWC-

2013, ECIR-2012, WSDM-2012, ECAI-2010, KDD-2009, ICML-2009, KDD-2007, ICML-2006, ITI (2006-2013), ICDM-2003. Dunja Mladenić je u uredničkom odboru više međunarodnih časopisa: Machine Learning Journal (od 2010. god.), Knowledge and Information Systems (od 2014. god.), ECML/PKDD 2013-2015 Guest Editorial Board, Journal of Computing and Information Technology (2008-2012), Journal of Information and Organizational Sciences (od 2008. god.), International Journal of Computational Intelligence (od 2008 god.).

SAŽETAK

Optimizacija mrežnih stranica za tražilice (engl. *Search engine optimization, SEO*) podrazumijeva tehnike pomoću kojih autor mrežnih stranica provodi nad svojim stranicama kako bi one što bolje rangirale u organskim (prirodnim) rezultatima pretraživanja na internetskim tražilicama za odabrane ključne riječi. Taj proces između ostalog uključuje i optimizaciju sadržaja, odnosno prilagodbu sadržaja mrežnih stranica prema preporukama za optimizaciju mrežnih stranica za tražilice (u daljem tekstu SEO preporukama). Ovim istraživanjem ispituje se mogućnost upotrebe strojnog učenja za klasifikaciju mrežnih stranica u tri predefinirane klase s obzirom na stupanj prilagodbe sadržaja SEO preporukama. Pomoću strojnoga učenja izgrađeni su klasifikatori koji su naučili svrstati nepoznati uzorak (mrežnu stranicu) u predefinirane klase, te utvrditi značajne faktore (varijable) koje utječu na stupanj prilagodbe. Također izgrađen je sustav ispravka „neprilagođenih“ stranica upotrebom tehnika iz domene obrade prirodnog jezika. Rezultati su pokazali da se pomoću strojnog učenja može ocijeniti stupanj prilagođenosti stranice SEO preporukama, da se strojno učenje može koristiti za utvrđivanje značajnih faktora, te da se izgrađeni sustav prilagodbe može koristiti za ispravak tj. poboljšanje mrežnih stranica koje su u prethodnim fazama klasificirane kao "neprilagođene".

Ključne riječi: optimizacija mrežnih stranica, SEO mrežnih stranica za tražilice, on-page optimizacija, klasifikacija, strojno učenje, obrada prirodnoga jezika

SUMMARY

Search engine optimization (SEO) involves techniques by which the author of the website customizes the website so that it ranks higher in organic (natural) search results on popular Internet search engines for selected keywords. This process includes, among others, the optimization of content (text) to fit SEO recommendations. This study examines the possibility of using machine learning tecniques to classify web pages into three predefined classes related to the degree of content adjustment to the SEO recommendations. Using machine learning algorithms, classifiers are built and trained to classify an unknown sample (web page) in the predefined classes and to identify important factors that affect the degree of adjustment. In addition, using algorithms from the domain of natural language processing a system for correction is built and tested. Results show that machine learning can be used to predict the degree of adjustments of web pages to SEO recommendations, for identifying important SEO factors and that the proposed correction system can be used to correct pages which were classified as "misfits" in prior stages.

Keywords: Search engine optimization, SEO, on-page optimization, classification, machine learning, natural language processing

PROŠIRENI SAŽETAK

U ovom radu istražuje se upotreba tehnika strojnog učenja i obrade prirodnog jezika u domeni optimizacije mrežnih stranica za internetske tražilice (engl. *Search engine optimization*, SEO). Rad je strukturian tako da se najprije opisuju osnove SEO, strojnog učenja i obrade prirodnoga jezika, te se zatim izlaže metodologija istraživanja i naposlijetku se prezentiraju rezultati i zaključci.

Proces optimizacije mrežnih stranica dijeli se na dva glavna dijela: optimizacija sadržaja (engl. *on-page optimization*) i optimizacija poveznica (engl. *off-page optimization*). Ovaj rad se bavi optimizacijom sadržaja koja uključuje tehnike rada na sadržaju stranice (HTML-kodu i tekstu) s ciljem upotrebe ključnih riječi na određenim mjestima u strukturi stranice. Cilj istraživanja je razviti modele klasifikatora koji će ispravno svrstavati mrežne stranice u tri predefinirane klase s obzirom na stupanj prilagođenosti stranica SEO preporukama koje objavljuju današnje internetske tražilice. Za treniranje klasifikatora koristi se znanje SEO stručnjaka koji su uzorak od 600 mrežnih stranica prethodno označili tj. svrstali u predefinirane klase. Metodama validacije utvrđuje se zatim uspješnost (točnost) modela klasifikatora u predviđanju klasne pripadnosti. Modeli razvijenih klasifikatora koriste se i za utvrđivanje značajnih SEO faktora sadržaja koji utječu na stupanj prilagođenosti. Ti faktori, zajedno sa skupom mrežnih stranica koje pripadaju najlošijoj klasi (koje imaju najmanji stupanj prilagođenosti), predstavljaju ulaz u sustav prilagodbe koji koristi tehnike iz domene obrade prirodnoga jezika s ciljem ispravka stranica tj. poboljšanja njihovog stupnja prilagođenosti. Učinkovitost takvog sustava testira se ponovnom klasifikacijom pomoću modela razvijenih u prethodnim fazama.

Na osnovu rezultata točnosti modela klasifikatora zaključujemo da se tehnike strojnog učenja mogu uspješno upotrijebiti u svrhu detekcije stupnja prilagođenosti sadržaja mrežne stranice SEO preporukama, te u svrhu detekcije značajnih faktora koji na to utječu. Rezultati su pokazali da sustav za prilagodbu tj. ispravak loše prilagođenih stranica koji se predlaže u ovom radu također uspješno može predložiti izmjene potrebne da bi mrežna stranica poboljšala stupanj prilagodbe.

Ključne riječi: optimizacija mrežnih stranica, SEO, on-page optimizacija, klasifikacija, strojno učenje, obrada prirodnoga jezika

EXTENDED ABSTRACT

This dissertation investigates the use of machine learning and natural language processing in the domain of search engine optimization (SEO). We first provide the description of SEO process, machine learning techniques and methods of natural language processing, following is research methodology and presentation of the evaluation results and conclusions.

The process of search engine optimization includes two main parts: „on-page“ optimization and „off-page“ optimization. This research explores only „on-page“ optimization which includes techniques for working on the content of the page (HTML code and text) with the goal of using targeted keywords on specific places in web page structure. The goal of this research is to develop classification models which will correctly classify web pages into three predefined classes based on content adjustments to SEO guidelines which are published by today search engines. The knowledge of SEO experts is used to train the classificators. SEO experts marked a sample of 600 web pages in three predefined classes. The accuracy of built models is validated by using standard validation methods used in machine learning. These models are also used to identify important „on-page“ SEO factors that affect the degree of content adjustment to SEO guidelines. These factors, together with data set that consists of unadjusted web pages are input into the correction system that uses natural language processing techniques for the purpose of improving the degree of adjustment. The efficiency of the proposed system is tested by using a classification model built in the first phase of this research.

Based on the classification accuracy of the constructed models, we conclude that the applied machine learning techniques can be successfully used to detect the degree of web page adjustments to SEO guidelines for „on-page“ SEO, and for detecting relevant factors. Results also show that the proposed correction system for modifying unadjusted web pages can successfully propose proper correction modifications so that the web page will be better adjusted to SEO guidelines.

Keywords: Search engine optimization, SEO, on-page optimization, classification, machine learning, natural language processing

SADRŽAJ

Popis slika I

Popis tablica II

1. UVOD 1

1.1. Predmet istraživanja 1

1.2. Motivacija 2

1.3. Ciljevi istraživanja i hipoteze 2

1.4. Metodologija istraživanja 3

1.5. Znanstveni doprinos 7

1.6. Struktura rada 7

2. Optimizacija mrežnih stranica za internetske tražilice 8

2.1. Faktori sadržaja 9

2.2. Vanjski faktori 12

2.3. Ključne riječi 13

2.4. Etički i neetički SEO 15

3. Strojno učenje 17

3.1. Metode klasifikacije 18

3.1.1. Stabla odlučivanja 18

3.1.2. Naivni Bayesov klasifikator 21

3.1.3. Metoda najbližih susjeda 22

3.1.4. Metoda potpornih vektora 24

3.1.5. Logistička regresija 26

3.2. Mjere uspješnosti i evaluacija klasifikatora 28

3.3. Optimizacija hiperparametara i regularizacija 30

4. Obrada prirodnog jezika 35

4.1. Tokenizacija, normalizacija i segmentacija teksta 36

4.1.1. Lematizacija 37
4.1.2. Svođenje na korijen riječi 38
4.2. Označavanje vrste riječi 38
4.3. Matrica riječ-dokument i težinske funkcije 39
4.4. Leksička baza WordNet 41
4.5. Sažimanje teksta 42
5. Metodologija istraživanja 44
5.1. Odabir uzorka 44
5.2. Označavanje uzorka 46
5.3. Formiranje nezavisnih varijabli i skupa podataka 48
5.4. Ekstrakcija faktora sadržaja 51
5.5. Razvoj sustava za prilagodbu 53
5.5.1. Obogaćivanje teksta ključnim riječima 55
5.5.2. Algoritam sažimanja 56
5.6. Analiza predloženih promjena 59
6. Rezultati istraživanja 60
6.1. Rezultati klasifikacije mrežnih stranica 60
6.1.1. Stabla odlučivanja 60
6.1.2. Naivni Bayesov klasifikator 63
6.1.3. KNN klasifikator 64
6.1.4. Metoda potpornih vektora 66
6.1.5. Logistička regresija 67
6.2. Usporedba rezultata klasifikacije 70
6.4. Karakteristike značajnih faktora sadržaja 77
6.5. Sustav za prilagodbu 79
6.6. Ponovna klasifikacija ispravljenih stranica 84
7. ZAKLJUČAK 86

8. POPIS LITERATURE 89
9. Prilozi 97

POPIS SLIKA

Slika 1.1 Shema istraživanja i predloženog sustava 5
Slika 3.1. Primjer stabla odlučivanja nad skupom podataka "Weather" 19
Slika 3.2 Primjer KNN metode sa K=5 23
Slika 3.3 Primjer mogućih razdvajajućih pravaca u 2-dimenzionalnom prostoru kod binarne klasifikacije 24
Slika 3.4 Primjer potpornih vektora (točke na isprekidanim linijama) u dvodimenzionalnom prostoru. 25
Slika 3.5 Logistička funkcija 26
Slika 4.1 Matrica riječ-dokument 40
Slika 5.1 Shema sustava prilagodbe mrežnih stranica 54
Slika 6.1 Rezultat optimizacije hiperparametra C i M kod algoritma stabla odlučivanja pomoću metode mrežne pretrage 61
Slika 6.2 Rezultat detaljnije optimizacije parametra C i M kod algoritma stabla odlučivanja pomoću metode mrežne pretrage 61
Slika 6.3 Iteracije i točnost kod mrežne pretrage hiperparametara stabla odlučivanja 62
Slika 6.4 Rezultat optimizacije hiperparametra k i p kod KNN algoritma pomoću metode mrežne pretrage 64
Slika 6.5 Iteracije i točnost kod mrežne pretrage hiperparametara algoritma KNN 65
Slika 6.6 Rezultat optimizacije hiperparametra c i σ kod SVM algoritma pomoću metode mrežne pretrage 66
Slika 6.7 Iteracije i točnost kod mrežne pretrage hiperparametara algoritma SVM 66
Slika 6.8 Rezultat optimizacije hiperparametra *cost* kod logističke regresije pomoću metode mrežne pretrage 68
Slika 6.9 Iteracije i točnost kod mrežne pretrage hiperparametara *cost* logističke regresije ... 68
Slika 6.10 Usporedba točnosti klasifikatora kod 10-struke unakrsne validacije s baznom točnošću tj. najčešćom klasom (označeno crvenom linijom). 71
Slika 6.11 Histogrami distribucija točnosti po klasifikatorima 72
Slika 6.12 Podjela statističkih testova po problemima strojnog učenja i vrsti testa (Izvor: Japkowicz i Mohak, 2011) 73
Slika 6.13 Krivulja učenja odabranih klasifikatora 77

POPIS TABLICA

Tablica 3.1 Matrica zabune 30
Tablica 4.1 Kratice za najčešće korištene POS oznake 39
Tablica 5.1 Broj mrežnih stranica u uzorku po kategorijama DMOZ-a 45
Tablica 5.2 Distribucija oznaka u uzorku 46
Tablica 5.3 Izračunata Fleiss-ova Kappa statistika za ocjene tri stručnjaka nad 600 primjera uzorka 47
Tablica 5.4 Značenje vrijednosti Kappa statistike (Izvor: Landis i Koch, 1977) 48
Tablica 5.5 Weighted Kappa statistika za ocjene tri stručnjaka nad 600 primjera iz uzorka 48
Tablica 5.6 Nezavisne varijable korištene u istraživanju 49
Tablica 5.7 Matrica Pearsonove korelacije nezavisnih varijabli 50
Tablica 5.8 Analiza glavnih komponenti 51
Tablica 5.9 Važnost varijabli po različitim testovima 52
Tablica 5.10 Primjer jednog zapisa iz rezultata sumarizacije teksta mrežne stranice 57
Tablica 5.11 Ocjena eksperata o podobnosti sažetaka za „meta opis“ (prema Matošević, 2018) 58
Tablica 6.1 Točnost algoritma J48 sa različitim načinima validacije 62
Tablica 6.2 Točnost naivnog Bayesovog Bayes klasifikatora 63
Tablica 6.3 Točnost KNN klasifikatora sa k=45 i p=1 65
Tablica 6.4 Točnost SVM klasifikatora sa c=7.74e+03 i σ=0.000464 67
Tablica 6.5 Točnost klasifikacije pomoću logističke regresije sa parametrom *cost*=2 68
Tablica 6.6 Signifikantnost modela logističke regresije 69
Tablica 6.7 Rezultati McNemarovog testa usporedbe klasifikatora sa baznom točnošću 74
Tablica 6.8 Rezultati Wilcox signed rank testa za 10-struku unakrsnu validaciju 75
Tablica 6.9 Frekvencije vrijednosti značajnih faktora stranica (varijabli) u klasi „prilagođeno“ 77
Tablica 6.10 Gustoća ključnih riječi u značajnim faktorima 78
Tablica 6.11 Dužina teksta (broj riječi) na pozicijama značajnih faktora 79
Tablica 6.12 Frekvencije vrijednosti nekih varijabli stranica u klasi „neprilagođeno“ 79
Tablica 6.13 Broj predloženih (od strane algoritma) i prihvaćenih zamjena riječi (od strane stručnjaka) po pozicijama (varijablama) 83
Tablica 6.14 Broj poboljšanja po varijablama nakon izlaza iz sustava ispravka 83

Tablica 6.15 Rezultati ponovne klasifikacije ispravljenih stranica 84

1. UVOD

Današnji Web se sastoji od 4.26 bilijuna indeksiranih mrežnih stranica[1]. Vlasnici mrežnih stranica žele da njihove stranice budu posjećene kako bi ostvarili svoje ciljeve (konverzije), a za to je potrebno da stranice budu dobro rangirane na internetskim tražilicama. Taj cilj je s ovom veličinom Web-a teško ostvariv – konkurencija je velika, svi se žele probiti što više u rezultatima pretrage. Jedan od načina je plaćeno oglašavanje na tražilicama, no korisnici više cijene prirodne ili tzv. organske rezultate pretrage na kojima se rang tj. pozicija određuje na temelju algoritma same tražilice. Autori mrežnih stranica koji poznaju kako algoritam radi su u prednosti, jer mogu mrežnu stranicu učiniti takvom da podilaze algoritmu i time ostvaruju više pozicije. Internetske tražilice svoj algoritam čuvaju u tajnosti, no objavljuju osnovne smjernice kojih se mrežne stranice trebaju držati ako žele da ih njihov algoritam što bolje razumije pri rangiranju. Oko ovog procesa prilagodbe mrežnih stranica za algoritme internetskih tražilica razvila se cijela industrija koja se naziva *optimizacija mrežnih stranica* (engl. *Search engine optimization*).

Ovo istraživanje bavi se utvrđivanjem značajnosti nekih faktora koji utječu na rang mrežnih stranica pomoću strojnog učenja, te izgradnjom algoritama za polu-automatiziranu prilagodbu mrežnih stranica upotrebom tehnika iz domene obrade prirodnog jezika.

1.1. Predmet istraživanja

Rang mrežne stranice na internetskim tražilicama ovisi o mnogo faktora. Ti faktori su poznati (Gupta, i dr. 2016; Luh, Yang i Huang 2016; Zhu i Wu 2011; Zhang i Dimitroff 2005; Hussien 2014), ali manje je poznato u kojoj mjeri svaki od njih utječe na rang. Kako bi povećali šanse da neka mrežna stranica bude bolje rangirana na određene ključne riječi, web autori i/ili stručnjaci za optimizaciju stranica za tražilice (u daljnjem tekstu „SEO stručnjaci“) provode određene akcije nad mrežnim stranicama. Jedan dio tih akcija uključuje rad na sadržaju. Stručnjak na osnovu svog iskustva i znanja lako može procijeniti u kojoj mjeri je neka mrežna stranica prilagođena preporukama tražilica za optimizaciju (u daljnjem tekstu „SEO preporukama“) tj. internetskim tražilicama za određenu ključnu riječ. Ako stranica nije

[1] prema http://www.worldwidewebsize.com/, dana 31.03.2018.

prilagođena, stručnjak će na osnovu svojih uvjerenja i svog znanja poduzeti određene korake i izmijeniti sadržaj i strukturu stranice.

Pod sadržajem mrežne stranice podrazumijevamo prvenstveno tekstove, ali i HTML-kod koji ih omeđuje i s kojim se definiraju dijelovi sadržaja (naslovi, tijelo teksta, poveznice, slike, navigacija i sl.). SEO preporuke su opće smjernice koje izdaju popularne tražilice (ali i SEO stručnjaci) o tome kako mrežne stranice trebaju biti izgrađene kako bi ih same tražilice lakše razumjele i rangirale (Google Webmaster Guidelines n.d.; Bing Webmaster Guidelines n.d.).

Koje su to karakteristike sadržaja, koje više, a koje manje utječu na rang mrežnih stranica predmet je mnogih istraživanja u ovom području (Gupta, i dr. 2016; Luh, Yang i Huang 2016; Zhu i Wu 2011; Zhang i Dimitroff, 2005; Hussien 2014). Metode tih istraživanja baziraju se na analizi sadržaja visokorangiranih stranica i eksperimentima (praćenje promjena ranga ukoliko se promjeni određeni faktor).

1.2. Motivacija

Posao stručnjaka u SEO procesu uključuje razne postupke prilagodbe mrežne stranice odabranim ključnim riječima: mijenja se HTML-kod stranice, pišu se i modificiraju postojeći tekstovi, dodaju se ključne riječi na određene pozicije te radi se na popularnosti mrežne stranice izgradnjom pozadinskih hiperveza. Neki od tih poslova, pogotovo iz domene prilagodbe sadržaja (engl. *on-page SEO*), često su rutinske prirode (npr. dodavanje ključnih riječi u naslov, meta opis, obogaćivanje teksta ključnim riječima i sl.). Istraživanje u ovom radu motivirano je izgradnjom računalnog sustava za potporu SEO procesu kako bi se određeni postupci automatizirali ili barem polu-automatizirali. Takav sustav uvelike bi pomogao SEO stručnjacima u svakodnevnom radu – smanjilo bi se vrijeme prilagodbe mrežnih stranica te bi se fokus s rutinskih poslova mogao prebaciti na neke značajnije procese poput rad na hipervezama. Vjerujemo da rezultati ovog istraživanja mogu pomoći u izgradnji jednog takvog sustava.

1.3. Ciljevi istraživanja i hipoteze

Ciljevi ovog istraživanja su:

1) Korištenjem metoda strojnog učenja za automatsku klasifikaciju podataka razviti model klasifikatora kojim bi se mrežne stranice ispravno klasificirale prema stupnju prilagodbe sadržaja s obzirom na SEO preporuke. Za učenje klasifikatora koristi se znanje eksperata (SEO stručnjaka).

2) Utvrditi relevantne faktore sadržaja pomoću razvijenih modela klasifikatora.

3) Razviti sustav za prilagodbu mrežnih stranica s obzirom na relevantne faktore sadržaja korištenjem metoda iz domene prirodnog jezika i testirati funkcionalnosti takvog sustava.

Postavljene se sljedeće hipoteze:

H1: Upotrebom strojnog učenja za klasifikaciju mrežnih stranica s obzirom na stupanj prilagodbe sadržaja mrežnih stranica SEO preporukama, postići će se točnost klasfikatora veća nego osnovna klasifikacija u većinski razred.

H2: Upotreba algoritama iz domene obrade prirodnog jezika nad mrežnim stranicama s najlošijim stupnjem prilagodbe sadržaja SEO preporukama, poboljšati će stupanj prilagodbe većine tih stranica.

Istraživanje je napravljeno nad stranicama na engleskom jeziku, međutim predloženi pristup i metodologija nisu ograničeni na engleski jezik, te se iste metode mogu upotrijebiti za mrežne stranice na bilo kojem jeziku uz uvjet da postoji razvijena leksička baza za taj jezik. Pod leksičkom bazom podrazumijevamo digitalnu bazu riječi nekog jezika u kojoj su riječi grupirane u sinonimske skupove povezanim različitim semantičkim odnosima. Najpoznatija takva baza je WordNet[2].

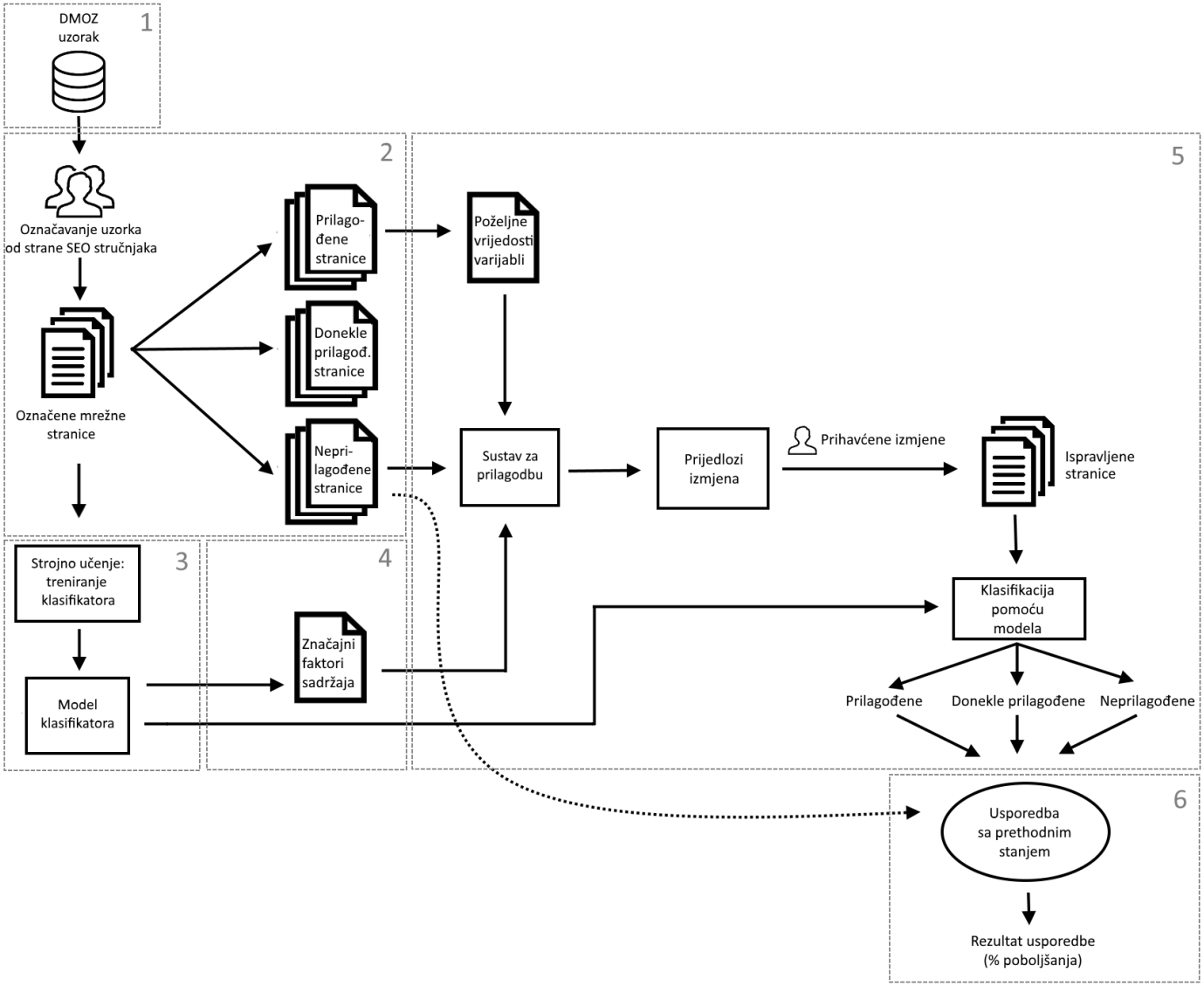
1.4. Metodologija istraživanja

U radu se koristite opće istraživačke metode deskripcije, analize i sinteze, komparacije (za usporedbu rezultata s prethodnim istraživanjima), metoda ispitivanja (za označavanje uzorka od strane stručnjaka i evaluaciju sustava prilagodbe) te metoda eksperimenta (za provjeru sustava prilagodbe). Metode iz domene strojnog učenja koriste se za postizanje ciljeva i potvrđivanje hipoteza. Hipoteza H1 će se prihvatiti ako rezultat modela klasifikatora tj.

[2] detaljnije u poglavlju 4.4.

njegova točnost bude veća od postotka dokumenata u većinskom razredu (klasi). Podaci za izgradnju modela uzeti su slučajnim uzorkovanjem iz DMOZ[3] direktorija na način da su ciljane ključne riječi izvučene iz naziva kategorije. Za potvrdu hipoteze H2 koristit se metoda eksperimenta i metoda ispitivanja za provjeru učinkovitosti izgrađenog sustava prilagodbe. Hipoteza H2 će se prihvatiti ako više od 50% stranica iz uzorka iz klase "neprilagođeno" nakon obrade u sustavu prijeđe u klasu "djelomično prilagođeno" ili "prilagođeno". Za izgradnju sustava prilagodbe u svrhu potvrđivanja hipoteze H2 koriste se tehnike iz domene obrade prirodnog jezika, poput tehnika zamjena sinonima, sažimanja teksta i strukturnih promjena u HTML-kodu mrežnih stranica. Sustav prilagodbe evaluira se na način da ljudski evaluator ocijeni predložene izmjene kao prihvatljive ili neprihvatljive, korigira izmjene, te se izmijenjena stranica ponovno klasificira radi utvrđivanja da li je poboljšana klasa stranice. Slika 1.1 prikazuje shemu istraživanja.

[3] https://dmoztools.net/, datum pristupa 3.5.2017.



Slika 1.1 Shema istraživanja i predloženog sustava

Istraživanje je podijeljeno u šest faza kako je označeno na slici 1.1:

1) Odabir slučajnog uzorka mrežnih stranica

2) Označavanje uzoraka od strane SEO stručnjaka u tri kategorije s obzirom na stupanj prilagođenosti na odabrane ključne riječi

3) Treniranje klasifikatora i evaluacija

4) Ekstrakcija značajnih faktora sadržaja iz rezultata klasifikacije

5) Razvoj sustava za prilagodbu mrežnih stranica

6) Analiza predloženih promjena

Prva faza uključuje odabir slučajnog uzorka od 600 mrežnih stranica iz direktorija DMOZ, pripremu skupa podataka na način da se za svaku stranicu definiraju ključne riječi iz naziva kategorije kojoj pripadaju. Nakon toga, u drugoj fazi, slijedi označavanje tj. ocjenjivanje uzorka stranica od strane SEO stručnjaka u tri predefinirane kategorije: „prilagođeno“, „djelomično prilagođeno“ i „neprilagođeno“ – ovisno o tome koliko je mrežna stranica prilagođena SEO preporukama tražilica. Skup podataka formiran u ovj fazi predstavlja ulazne podatke za klasifikaciju u trećoj fazi. Rezultati klasifikacije treće faze koriste se u četvrtoj fazi za utvrđivanje značajnih faktora. Sustav za prilagodbu razvija se u petoj fazi i u svojim algoritmima koristi značajne faktore utvrđene u prethodnoj fazi. Posljednja faza je analiza predloženih promjena koje je predložio sustav za prilagodbu, te utvrđivanje poboljšanja stupnja prilagodbe mrežnih stranica koje su u drugoj fazi klasificirane kao „neprilagođene“.

Faze su detaljnije opisane u 5. poglavlju.

Predloženi pristup prikladan je za većinu mrežnih stranica. Budući da se sustav ispravka bazira na tekstu stranica i HTML-kodu potrebno je naglasiti nekoliko ograničenja:

- Mrežne stranice koje nemaju tekst u tijelu neće biti moguće ispraviti
- Postoje poteškoće u ekstrakciji tekstova iz HTML-a za mrežne stranice koje sadržaj generiraju JavaScript-om ili sličnim tehnologijama
- Algoritam korekcije je primjenjiv samo na mrežnim stranicama na jeziku za kojeg postoji leksička baza.

Navedena ograničenja mogla bi se prevazići daljnjim istraživanjima i napretkom leksičkih baza.

1.5. Znanstveni doprinos

Izvorni znanstveni doprinos ovog doktorskog istraživanja očekuje se u:

a) razvoju novog modela za primjenu metoda strojnog učenja u cilju klasifikacije mrežnih stranica prema stupnju prilagođenosti SEO preporukama i utvrđivanje primjerenosti tog postupka

b) metodi za utvrđivanje relevantnosti faktora sadržaja temeljem razvijenih modela

c) razvoju sustava za prilagodbu mrežnih stranica SEO preporukama korištenjem metoda iz domene obrade prirodnog jezika u skladu s utvrđenim relevantnim značajkama sadržaja

d) skupu podataka s oznakama prilagođenosti SEO preporukama koji će se izraditi u radu i biti javno dostupan za buduća istraživanja u ovom području.

1.6. Struktura rada

Rad je strukturiran tako da se najprije daje kratki teorijski uvod u teme kojima se rad bavi (SEO, strojno učenje, obrada prirodnog jezika), nakon čega slijedi opis istraživanja i rezultata istraživanja.

Uvod u optimizaciju mrežnih stranica za internetske tražilice dan je u drugom poglavlju. Treće poglavlje opisuju metode strojnog učenja koje se koriste u radu. Četvrto poglavlje bavi se osnovama obrade prirodnog jezika. Metodologija istraživanja opisana je u petom poglavlju, a rezultati su predstavljeni u šestom poglavlju.

Na kraju slijedi zaključak i prilozi koji sadrže skup podataka izrađen i korišten u istraživanju.

2. OPTIMIZACIJA MREŽNIH STRANICA ZA INTERNETSKE TRAŽILICE

Optimizacija mrežnih stranica za internetske tražilice (engl. *Search Engine Optimization*, SEO) se u literaturi navodi kao jedna od tehnika internetskog marketinga. Radi se o procesu pomoću kojega SEO stručnjaci nastoje postići što bolju poziciju na rezultatima pretrage na internetskim tražilicama za odabrane ključne riječi. Kada govorimo o SEO onda govorimo o tzv. prirodnim ili organskim rezultatima pretrage – kako se nazivaju rezultati pretrage koji isključivo ovise o algoritmima tražilice – za razliku od plaćenih rezultata koje tražilice prikazuju pored prirodnih rezultata. Rangiranje mrežnih stranica u plaćenim rezultatima ovise između ostalog o cijeni koju su oglašivači (vlasnici mrežnih stranica) spremni platiti (Rutz i Bucklin 2016). Rang u prirodnim rezultatima pretrage nije moguće kupiti tj. na njega se ne može direktno utjecati, stoga je i puno teže postići dobru poziciju u prirodnim rezultatima. Područje internet marketinga koji se bavi plaćenim rezultatima se naziva „oglašavanje na tražilicama“ ili „PPC oglašavanje“ (vrsta naplate po kliku koja se najčešće koristi kod ove vrste oglašavanja), i njome se često bave marketing stručnjaci. Optimizacija mrežnih stranica za internetske tražilice, koja se bavi rangiranjem u prirodnim rezultatima, zahtijeva nešto dublje tehničko poznavanje načina rada tražilice i kako su same mrežne stranice građene (HTML-kod), te se ovime često bave stručnjaci koji imaju iskustva u izgradnji mrežnih stranica (tzv. „web masteri“ ili „web inženjeri“).

Dok za oglašavanje na tražilicama iste nude certifikate pomoću kojih se možete profilirati u stručnjaka za navedenu djelatnost (npr. Google Ads certifikat), za SEO ne postoje službeni načini certificiranja. Postoje tvrtke koje se bave SEO-om i nude svoje certifikate, ali budući da nisu izdane od strane samih tražilica ne možemo ih smatrati službenim. Zbog toga se pojam „SEO stručnjaka“, koji se koristi u ovom radu, ne može službeno definirati, ali može se reći da se radi o osobama sa iskustvom i znanjem u izradi mrežnih stranica, načinom rada internetskih tražilica te SEO procesima (Kilgore 2017).

Algoritmi internetskih tražilica pomoću kojih one pronalaze i rangiraju rezultate pretrage u svojim sučeljima koriste višestruke faktore (Hussien 2014). Većina tih faktora je poznata, međutim njihova uloga u funkcijama rangiranja nije javno poznata. SEO stručnjaci trebaju stalno pratiti promjene u načinu rada tražilica i eksperimentirati kako bi dokučili kako što bolje pozicionirati određenu mrežnu stranicu za određene ključne riječi. Internetske tražilice

javno objavljuju opće smjernice kako mrežne stranice trebaju izgledati kako bi ih algoritam tražilice što bolje razumio i dobro pozicionirao (Google Webmaster Guidelines n.d.; Bing Webmaster Guidelines n.d.). Faktori koji se uzimaju u obzir se često dijele u dvije glavne skupine: faktori sadržaja (engl. *on-page*) i vanjski faktori (engl. *off-page*)

2.1. Faktori sadržaja

Faktori sadržaja uključuju karakteristike same mrežne stranice tj. njenog tekstnog, slikovnog i video sadržaja, navigacije, URL-a, strukture i HTML-koda. Jasno je da su ti faktori pod direktnim utjecajem onoga tko izrađuje mrežnu stranicu, što znači da njihova uspješnost s aspekta optimizacije mrežnih stranica ovisi isključivo o znanju i vještini kreatora mrežne stranice. Internetske tražilice u svojim preporukama (Google Webmaster Guidelines i Bing Webmaster Guidelines n.d.) naglašavaju sljedeće faktore sadržaja:

- Kvaliteta teksta na stranici (podrazumijeva i kvalitetu informacija koja se pruža)
- Jasna navigacija
- Naslov stranice (HTML oznaka „title“)
- Meta opis stranice (HTML oznaka „meta description“)
- Upotreba H oznaka za naslove (H1, H2, H3 itd.)
- ALT atribut u slikama (kratki opis slike)
- Tekst poveznica (engl. *anchor text*)
- URL stranice (uključujući domenu)
- Brzina učitavanja stranice
- Omjer HTML-koda i sadržaja (treba biti u korist sadržaja)

Ovaj popis nije konačan, u faktore sadržaja mogu spadati i ostali manje značajni faktori (poput npr. starosti domene i sl.), no njima svima je zajedničko da na njih autor stranice može utjecati – za razliku od vanjskih faktora na koje ne možemo izravno utjecati.

Faktori sadržaja i njihov utjecaj na rang stranice u rezultatima pretrage tema su mnogih znanstvenih istraživanja. Budući da internetske tražilice objavljuju samo opći (kratki) popis najznačajnjih faktora sadržaja i opće smjernice o tome kako bi mrežna stranica trebala biti strukturirana, a ne i važnost pojedinih faktora, mnogi istraživači u tom području pokušavaju raznim metodama doći do detaljnjih podataka o faktorima sadržaja. Većina prethodnih

istraživanja se bavi utvrđivanjem značajnosti faktora, njihovim rangiranjem, ali i otkrivajem nekih novih faktora i njihove povezanosti s postojećim, te rangom stranica.

Slijedi kratki pregled najznačajnjih prethodnih istraživanja u ovom području.

U (Moreno i Martinez 2013) autori primjećuju podudarnosti ranking faktora s faktorima upotrebljivosti mrežne stranice – zaključujući kako sama upotreba određenih smjernica za optimizaciju mrežnih stranica će ujedno i povećati upotrebljivost stranice, a samim time i rang stranice na tražilicama. U (Su, i dr. 2014) autori istražuju gustoću ključnih riječi na pojedinim pozicijama na mrežnoj stranici i njihov utjecaj na poziciju u rezultatima pretrage (engl. *Search engine results pages*, SERP), te predlažu optimalnu frekvenciju ključnih riječi. Također autori tog istraživanja izgrađuju sustav za predviđanje ranga mrežne stranice s obzirom na utvrđene značajne faktore. Kao izvor podataka koriste mrežne stranice koje se pojavljuju na prvih 100 pozicija u SERP-u na odabrane ključne riječi te izgrađuju sustav koji uči nad tim podacima i izgrađuju model za predviđanje. U prosjeku u 78% slučajeva njihov sustav dobro predviđa rang stranice među prvih 10 pozicija. Korak dalje u utvrđivanju faktora otišli su autori članka (Zhu i Wu 2011) koji su primijenili principe obrnutog inženjeringa analizirajući visoko pozicionirane mrežne stranice kako bi došli do faktora koji su najvažniji.

Opsežno istraživanje karakteristika sadržaja tj. faktora sadržaja koji utječu na vidljivost mrežnih stranica u SERP-u objavljena su u (Zhang i Dimitroff, The impact of webpage content characteristics on webpage visibility in search engine results (Part I) 2005) i (Zhang i Dimitroff, The impact of metadata implementation on webpage visibility in search engine results (Part II) 2005).

U (Zhang i Dimitroff, The impact of webpage content characteristics on webpage visibility in search engine results (Part I) 2005) autori istražuju utjecaj frekvencije ključne riječi u naslovu stranice (HTML oznaka „title“) i tijelu stranice na rang stranice u različitim tražilicama. Eksperiment koji su proveli pokazao je da frekvencija ključnih riječi u naslovu utječe na rang te da je optimalna frekvencija 3. Stranice s većom frekvencijom od 3 imale su lošije performanse. Analiza frekvencija ključnih riječi u tijelu mrežne stranice (HTML oznaka „body“) pokazala je da mrežne stranice s većom frekvencijom rangiraju bolje na tražilicama. Autori su promatrali mrežne stranice s frekvencijom ključnih riječi u tijelu mrežne stranice do 5 i stranice s maksimalnom frekvencijom od 5 koje su bile više rangirane, dok su stranice s nižim frekvencijama ostvarivale niži rang. Uspoređujući rezultate u različitim tražilicama autori su utvrdili da postoje razlike kako pojedine tražilice rangiraju testne stranice. Također,

eksperiment je pokazao da frekvencija ključnih riječi u tijelu stranice ima jači utjecaj na rang od frekvencije riječi u naslovu stranice, te da stranice koje imaju ključne riječi na obje pozicije (u naslovu i tijelu) rangiraju bolje od stranica s ključnim riječima samo na jednoj poziciji (samo u naslovu ili samo u tijelu). Rezultati rangiranja po tražilicama su bili konzistentniji ako je mrežna stranica sadržavala ključne riječi na obje pozicije, za razliku od upotrebe samo jedne pozicije. Autori su također istraživali utjecaj oblikovanja (boje fonta, veličina fonta, velika-mala slova i sl.) na poziciju i došli do zaključka da te karakteristike nemaju bitnog utjecaj na rang.

U (Zhang i Dimitroff, The impact of metadata implementation on webpage visibility in search engine results (Part II) 2005) autori istražuju utjecaj postojanja meta-podataka unutar mrežnih stranica na poziciju (vidljivost) u SERP-u. Od meta-podataka autori su odabrali sljedeća tri: „meta title“, „meta subject“ i „meta description“. Rezultati pokazuju da mrežne stranice koje sadrže meta-podatke rangiraju bolje od onih koje ne sadrže. Također testirane su i različite kombinacije meta-podataka tj. broj elemenata meta-podataka te su rezultati pokazali da mrežne stranice koje sadrže sva tri elementa („title“, „subject“, „description“) rangiraju bolje od stranica koje sadrže samo jedan ili dva navedena elementa. Rezultati se također razlikuju po tražilicama – neki elementi su više stabilni (npr. „meta subject“) dok su drugi nestabilniji (npr. „meta description“). Upotreba ključnih riječi u meta-podacima značajno povećava rang stranice – ako se iste te ključne riječi pojavljuju u naslovu i tijelu mrežne stranice.

Nešto novije istraživanje (Su, i dr. 2014) proširuje eksperiment na više pozicija. Autori su dokazali da je na temelju nekoliko varijabli moguće točno predvidjeti rang 7 od 10 stranica (78%) na tražilicama Google i Bing. Dominantan faktor u tom modelu je “PR” (PageRank) koji spada u vanjske parametre, nakon kojeg po važnosti slijede varijable “HOST” (domena), “TITLE” (naslov HTML dokumenta), “M_DES” (meta-opis HTML dokumenta), “PATH” (URL putanja s nazivom datoteke HTML dokumenta). Ostale varijable iz navedenog istraživanja su imale manju važnost. U istom članku autori istražuju procjene SEO eksperata. Anketiranjem 37 eksperata o faktorima važnim za rangiranje došli su do sljedećih faktora (poredano po važnosti – od više važnog prema manje važnom): ključna riječ u oznaci „title“, tekst dolazne poveznice (engl. *anchor text of inbound link*), globalna popularnost poveznice stranice, starost stranice, interna popularnost poveznice, tematska relevantnost dolaznih poveznica, popularnost poveznica unutar tematskih društvenih grupa, ključna riječ u tijelu stranice, globalna popularnost povezanih stranica, tematska vezanost povezanih stranica. Od navedenih faktora samo su tri faktora vanjska, dok su ostali faktori sadržaja. Proširujući

anketu na „polu-eksperte“ (osobe koje zanima SEO područje) došlo se do saznanja da su faktori koje oni smatraju bitnim bili uglavnom faktori sadržaja.

U (Zhu i Wu 2011) autori pomoću obrnutog inženjeringa dolaze do pet najvažnijih faktora ranga: dužina URL-a, ključna riječ u domeni, gustoća ključnih riječi u HTML oznaci H1, gustoća ključnih riječi u naslovu (oznaka “title”) i broj slojeva u URL-u. Analizom karakteristika 200.000 mrežnih stranica iz najviše rangiranih 20 na Google SERP-u za dane upite autori su došli do faktora i preporuka koje navode u svom radu. Slične rezultate o važnosti pojedinih faktora sadržaja (ključne riječi u oznaci “TITLE”, “H1” i tijelu stranice) potvrđeni su u (Sagot, Ostrosi i Fougères 2016; Giomelakis i Veglis 2016; Khan i Mahmood 2018; Buddenbrock 2016), čime je dokazano da su isti faktori sadržaja relevantni već godinama.

U (Matošević, 2015) autor predlaže metriku koja kombinira faktore sadržaja i ocjenjuje stranicu s obzirom na stupanj prilagođenosti SEO preporukama. Predložena metrika je izražena linearnom kombinacijom faktora i težinskih varijabli koje su određene empirijski (nisu naučene).

Iz analize prethodnih istraživanja o faktorima sadržaja možemo zaključiti da je popis faktora sadržaja koji je dan na početku poglavlja relevantan i poklapa se s preporukama internetskih tražilica. U većini istraživanja znanstvenici koriste rezultate pretrage kao početnu točku tj. promatraju najbolje rangirane stranice iz kojih pokušavaju utvrditi značajne faktore. Pristup u ovoj disertaciji je drugačiji – za utvrđivanje značajnih faktora koristi se znanje SEO stručnjaka na način da se metodama strojnoga učenja nad skupom mrežnih stranica koje su stručnjaci označili kao najboljim (najbolje prilagođenim SEO preporukama) izvuku značajni faktori sadržaja. Ovaj pristup omogućava da se faktori sadržaja tj. njihove poželjne vrijednosti promotre objektivnije nego kada bi se istraživanje temeljilo na stranicama iz rezultata pretrage koje mogu sadržavati stranice koje su dospjele tamo zahvaljujući utjecaju vanjskih faktora.

2.2. Vanjski faktori

Vanjski faktori ili faktori koji se ne nalaze na samoj stranici nego ovise o nekim vanjskim utjecajima čine srž algoritama za rangiranje današnjim tražilica. To su faktori na koje autor mrežne stranice nema potpunu kontrolu kao što je slučaj s faktorima sadržaja. Tu uglavnom spadaju faktori povezani s hiperpoveznicama (engl. *links*) kao npr. broj dolaznih poveznica

(engl. *inbound links*), broj odlaznih poveznica (engl. *outbound links*), kvaliteta povezanih stranica, faktori povezani s društvenim mrežama i sl. Prvi algoritmi rangiranja zasnovani na poveznicama bilu su PageRank (Brin i Page 1998) i HITS (Kleinberg, i dr. 1999). Oni za rangiranje tj. utvrđivanje relevantnosti stranice za određeni upit, koriste poveznice kao „glasove“ – što više dolaznih poveznica ima neka stranica to se pretpostavlja da je i značajnija. Veze se u tom slučaju promatraju kao „preporuke“ ili „glasovi“ za tu stranicu koje ostavljaju drugi korisnici web-a. Pojavom web 2.0 i društvenih mreža to je još više izraženo jer svatko svojim mišljenjem može doprinijeti tom „glasovanju“.

U (Zhang i Cabage 2017) autori istražuju učinkovitost procesa stvaranja pozadinskih poveznica (engl. *back links*), odnosno dolaznih hiperveza i upotrebe društvenih medija kako bi se povećao rang na tražilicama. Rezultati njihovog istraživanja ukazuju na to da upotreba društvenih medija (dijeljenje sadržaja) utječe pozitivno na rang stranice u SERP-u, ali je proces dobivanja dolaznih hiperveza (engl. *link building*) dugoročnije i stabilnije ulaganje u SEO odnosno u „off-page“ optimizaciju. Takvi rezultati su očekivani budući da je srž rangiranja PageRank algoritam, ali dokazuje i da pojavom novih alata na mreži (dijeljenje, „lajkanje“ i sl.) tražilice prate trendove i ažuriraju svoje algoritme novim ranking faktorima.

Vanjski faktori nisu predmet ovog istraživanja, stoga neće biti detaljnije opisivani.

2.3. Ključne riječi

SEO proces uvijek počinje ključnim riječima, zato što korisnici Interneta, tj. internetskih tražilica, upotrebljavaju riječi i fraze koje upisuju u sučelje internetske tražilice s ciljem pronalaska dokumenata (mrežnih stranica) koje će zadovoljiti njihovu potrebu za informacijom. Te riječi ili fraze se nazivaju ključne riječi ili upiti. Zadaća tražilica je da na osnovu upita pokušaju dokučiti korisnikovu namjeru – što on zapravo želi pronaći, da li traži odgovor na neko pitanje, da li traži informacije o nekom proizvodu, da li traži podatke o nekom objektu ili nešto drugo. Mogućnosti su beskonačne, a upit je u većini slučajeva kratak i najčešće se sastoji od 2-3 riječi. Upit može biti formuliran i u obliku pitanja – na što tražilice danas dosta dobro odgovaraju, međutim upotreba ključnih riječi tj. fraza je i dalje dominantno u upotrebi tražilica (White, Richardson i Yih 2015). Jasno je kako je zadatak tražilica težak i da dobar odabir ključnih riječi može značajno utjecati na prikaz relevantnih stranica. Osnovni problem je strojno razumijevanje upita zbog različitih značenja koje riječi i fraze mogu imati

u različitim jezicima. Riječi se ne promatraju kao skup slova nego je nužno promatrati ih semantički uzimajući u obzir kontekst u kojem se pojavljuju. Pod kontekstom se ne smatraju samo ostale riječi tj. tekst koji ga okružuje već se za utvrđivanje značenja mogu koristiti i logovi prethodnih upita korisnika (Korayem, i dr. 2015). Sve su to problemi kojima se bavi područje obrade prirodnoga jezika (engl. *natural language processing*) koje je detaljnije opisano u poglavlju 4.

Dužina ključnih riječi tj. upita je važna kada govorimo o pretraživanju informacija pomoću internetskih tražilica. Upiti koji se sastoje od jedne riječi često su teški za obradu – skoro da je nemoguće dokučiti korisnikovu namjeru na osnovu upita od samo jedne riječi. To su preopćeniti upiti za koje često postoji jako puno dokumenata koji su potencijalno relevantni. Druga krajnost su upiti koji se sastoje od više od 5 riječi. Oni su usko specijalizirani, korisnikova namjera je detaljno navedena u upitu, međutim broj dokumenata koji zadovoljavaju sve kriterije iz upita može biti vrlo mali, pa često korisnik može dobiti oskudne rezultate. Najčešće korisnici koriste 2-3 riječi i to se smatra i najboljim pristupom (Zhang i Dimitroff, The impact of webpage content characteristics on webpage visibility in search engine results (Part I) 2005).

Kako bi se određena mrežna stranica pojavila u prirodnim rezultatima pretrage za određene ključne riječi, potrebno je da ih ona koristi u faktorima sadržaja i vanjskim faktorima. Osnovna ideja SEO-a je pronalazak ključnih riječi koji će mrežnu stranicu uspjeti dovesti na što bolju poziciju u rezultatima pretrage. Bolja pozicija će se lakše postići za one ključne riječi za koje ne postoji mnogo dokumenata tj. gdje je konkurencija manja. S druge strane, stranica će dobiti više posjeta ako je dobro pozicionirana za ključne riječi koje se često koriste. To je dvosjekli mač – potrebno je pronaći ključne riječi za koje postoji veliki interes korisnika, a istovremeno mala konkurencija na rezultatima tražilica (mali broj dokumenata).

SEO stručnjacima su na raspolaganju razni softverski alati pomoću kojih mogu doći do dobrih ključnih riječi. Alati mogu prikazivati broj mjesečnih pretraga za određenu riječ (kao npr. „Google trends“), broj dokumenata u indeksu tražilica koji sadrže tu riječ (za procjenu konkurencije) te preporuke sličnih riječi i fraza. Stručnjaci na taj način biraju ključne riječi na koje žele ciljati, te nakon toga započinju proces optimizacije. Proces uključuje postavljanje ključnih riječi na određene pozicije u HTML-kodu mrežne stranice (za optimizaciju sadržaja) i upotrebu ključnih riječi kod vanjskih faktora (npr. poveznicama, društvenim mrežama i sl.).

Neki alati i algoritmi bave se ekstrakcijom ključnih riječi iz teksta stranice. To je dobar pristup ako nam je cilj saznati ključne riječi na koje određena (po mogućnosti optimizirana) stranica cilja (npr. kod analize konkurencije). No, kada govorimo o SEO prilagodbi za nove ili neoptimizirane stranice onda izvlačenje ključnih riječi iz teksta u većini slučajeva nije primjereno – cilj SEO-a je najprije odabrati najbolje ključne riječi (koje će polučiti najbolje pozicije na tražilicama i dovesti najviše organskih posjeta na stranicu) i onda prilagoditi stranicu (faktore) tim riječima.

Kada govorimo o prilagodbi sadržaja odabranim ključnim riječima onda mislimo na ubacivanje i upotrebu ključnih riječi u naslovu stranice (HTML oznaka „title“), meta opisu („meta description“ oznaka), tijelu stranice (sam tekst na stranici), ALT oznaci (slike), nazivima datoteka, nazivu domene, u naslovima na stranici (HTML oznake „h1“ do „h6“) i sl. Ključne riječi se trebaju ubacivati na navedene pozicije na način da ne narušavaju smisao teksta i samu kvalitetu stranice. Potrebno je uvijek imati na umu da se stranice rade za korisnike, a ne za robote i internetske tražilice.

2.4. Etički i neetički SEO

Etički SEO (engl. *white hat SEO*) se definira kao SEO proces koji je u potpunosti u skladu s preporukama tražilica. Sve ostale tehnike pomoću kojih SEO stručnjaci pokušavaju „prevariti“ tražilice nazivaju se neetičke (engl. *black hat SEO*). Jedan od primjera neetičkog SEO-a je prekomjerno korištenje ključnih riječi (engl. *keyword stuffing*), skrivanje sadržaja od korisnika dok je on istovremeno vidljiv robotima tražilica, preusmjeravanje korisnika na druge stranice bez njegovog znanja (koje se ne bave temom iz upita) itd. (Agrawal, Somani i Chhabra 2016). U literaturi se često te tehnike nazivaju „varanje tražilica“ ili engl. *spamdexing*. Tražilice imaju razvijene svoje filtere pomoću kojih pokušavaju otkriti takve stranice nakon čega one bivaju kažnjene izbacivanjem iz indeksa ili gubitkom pozicije.

Algoritmi za otkrivanje tzv. „spam stranica“ ili neetičkog SEO-a česta su tema znanstvenih istraživanja (Roul, Asthana i Kumar 2016). Širenje lažnih vijesti i varanje tražilica uoči značajnih događaja (npr. izbora) također je pojava koja predstavlja izazov za algoritme tražilica. U (Metaxas i Pruksachatkun 2017) autori istražuju pojavu lažnih vijesti u rezultatima pretrage u slučaju američkih izbora za kongres 2016. godine za svakog pojedinačnog kandidata, te promjenu ranga stranica kandidata tokom izborne kampanje. Rezultati su

pokazali da je manipulacija rezultatima pretrage itekako moguća, s time da se neke tražilice više a neke manje uspješno brane od takvih napada. Ovo je primjer neetičkog SEO-a i manipuliranja tražilica koje je drugačije od klasičnog varanja ključnim riječima.

3. STROJNO UČENJE

Strojno učenje je grana umjetne inteligencije koja se bavi oblikovanjem algoritama koji svoju učinkovitost poboljšavaju na temelju empirijskih podataka. To je interdisciplinarno područje koje uključuje statistiku, znanost o podacima, umjetnu inteligenciju, baze podataka i ostala povezana područja iz domene informacijskih znanosti. Cilj strojnog učenja je učenje iz podataka kako bi stroj mogao predviđati vrijednosti određenih varijabli, utvrđivati interesantne strukture u podacima ili raspoznavati uzorke. Obično se dijeli na dvije osnovne skupine: nadgledano učenje (engl. *supervised learning*) i nenadgledano učenje (engl. *unsupervised learning*), ovisno da li su podaci na temelju kojih se uči označeni, tj. imamo vrijednost zavisne varijable danu od strane eksperata (Witten, i dr. 2016).

Nadgledano učenje je učenje na temelju podataka u kojima je poznata vrijednost zavisne varijable koja se pokušava naučiti. To nazivamo označenim podacima – svaki uzorak podatka ima vrijednost zavisne varijable koju je cilj predvidjeti u budućnosti. Zavisna varijabla se obično izvodi kao kombinacija nezavisnih varijabli i težinskih faktora. Cilj je predvidjeti vrijednost zavisne varijable za novi podatak dosad neviđen u povijesnim podacima s obzirom na vrijednosti nezavisnih varijabli. Ako je zavisna varijabla kontinuirana (numerička) onda govorimo o problemu regresije, a ako je ona kategorijska (nominalna vrijednost) onda se radi o klasifikaciji. Npr. kod linearne regresije se to matematički može prikazati kao:

$$y=x_1w_1+x_2w_2+...+x_nw_n+w_0 \quad (1)$$

gdje je:

y zavisna varijabla,

x_1, x_2, ..., x_n nezavisne varijable

w_0, w_1, ..., w_n težinski faktori.

Kod nenadgledanog učenja varijabla *y* nije u podacima poznata, tj. podaci nisu označeni. Cilj nenadgledanog učenja nije predviđanje varijable *y* nego utvrđivanje interesantnih struktura u podacima. Najbolji primjer nenadgledanog učenja su algoritmi grupiranja (klasteriranja) čiji je cilj postojeće podatke grupirati u određene grupe na osnovi njihovih karakteristika (nezavisnih varijabli).

Bez obzira koju metodu strojnog učenja primjenjivali, da bismo mogli evaluirati rezultate izgrađenog modela strojnog učenja, moramo imati skup za treniranje i skup za validaciju, koji

moraju biti različiti. Treniranje je proces učenja tj. izgradnje modela i on koristi podatke iz skupa za treniranje. Izgrađeni model se zatim testira na posebnom skupu za validaciju. Važno je da skup za validaciju sadrži podatke koji nisu korišteni za treniranje, tj. da je različit od skupa za treniranje. Ukoliko ne postoji skup za validaciju onda se može koristiti unakrsna validacija. Više o načinima i mjerama evaluacije u poglavlju 3.2.

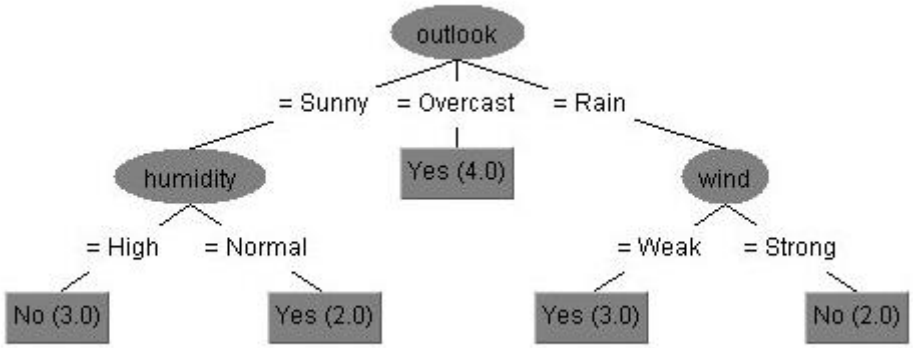
U ovom radu istražujemo stupanj prilagođenosti mrežnih stranica SEO preporukama. Definirana su tri stupnja prilagodbe: „prilagođeno“, „djelomično prilagođeno“ i „neprilagođeno“. Podaci su označeni od strane SEO stručnjaka i budući da su to kategorijske vrijednosti, za ovu vrstu problema prigodne su tehnike klasifikacije. U nastavku se daje pregled najčešće korištenih algoritama klasifikacije (Wu, i dr. 2008) koji se koriste u ovom istraživanju.

3.1. Metode klasifikacije

3.1.1. Stabla odlučivanja

Stabla odlučivanja (engl. *decision trees*) su jedan od jednostavnijih klasifikatora koji daju interpretativne rezultate. Zasnivaju se na metodi „podijeli pa vladaj“ što znači da se inicijalni skup podataka u više koraka dijeli na osnovu vrijednosti nezavisnih varijabli kako bi se izgradilo optimalno stablo. Stabla se sastoje od čvorova odluka na kojima se ispituje vrijednost odabrane varijable, grana koje predstavljaju moguće ishode čvorova, te listova kao krajnjih elemenata stabla, koji predstavljaju konačnu klasifikaciju tj. na njima se određuje pripadnost klasi. Izgrađeno stablo koristi se kao model za klasifikaciju novih uzorka podataka.

Stabla odlučivanja se mogu osim na klasifikaciju primijeniti i na regresiju – onda govorimo o regresijskim stablima. Slika 3.1 prikazuje primjer stabla odlučivanja nad skupom podataka „Weather“ – popularnim skupom koji se koristi u literaturi za objašnjavanje stabla odlučivanja (Witten i dr. 2011). Radi se o primjeru modela klasifikacijskog stabla za predviđanje da li će se igrati neka igra ovisno o vremenskim uvjetima. Stablo na slici 3.1 je izgrađeno nad podacima za treniranje, a model se može upotrijebiti nad novim podacima na način da se slijedi put kroz stablo od korijena prema listu (od gore prema dolje) ovisno o vrijednostima atributa u čvorovima. Na listu stabla predviđena je klasa.



Slika 3.1. Primjer stabla odlučivanja nad skupom podataka "Weather" (Izvor: Witten, i dr. 2016)

Najpoznatiji algoritmi stabla odlučivanja su ID3, C4.5, CART, CHAID i QUEST (Witten, i dr. 2016). Osnovni problem kod izgradnje modela stabla odlučivanja je odabir atributa s kojim će stablo započeti, tzv. korijenskog atributa koji se postavlja u čvor na vrhu stabla, tzv. korijen stabla. Kako odabrati najbolji atribut za taj početni čvor odluke? Odgovor leži u entropiji – mjeri čistoće skupa danoj formulom (1).

$$\boldsymbol{Entropija}\ (\boldsymbol{S}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} -\boldsymbol{p_i}\ \boldsymbol{log_2}\ \boldsymbol{p_i} \tag{2}$$

gdje je *S* skup, p_i proporcija klase *i* u skupu *S*, a *c* broj klasa.

Kako bi stablo bilo što točnije i što kraće u korijen se treba odabrati atribut (*A*) koja najbolje razdvaja skup podataka (S) prema odabranoj klasi. Isti princip se primjenjuje kod svakog sljedećeg čvora. Mjera bazirana na entropiji koja se pri tome uzima u obzir je informacijski dobitak (2) koji predstavlja očekivanu redukciju entropije uzrokovanu razdvajanjem primjera na osnovi određenog atributa.

$$\boldsymbol{Inf.Dobitak}\ (\boldsymbol{S},\boldsymbol{A}) = \boldsymbol{Entropija}\ (\boldsymbol{S}) - \sum_{\boldsymbol{v}\ \in \boldsymbol{vri}(\boldsymbol{A})} \frac{|\boldsymbol{S_v}|}{|\boldsymbol{S}|}\ \boldsymbol{Entropija}\ (\boldsymbol{S_v}) \tag{3}$$

Gdje je *A* atribut u skupu *S*, *vri(A)* skup svih mogućih vrijednosti atributa *A*, S_v podskup od *S* za koji atribut *A* ima vrijednost *v* (Gamberger i Šmuc 2001).

Nakon izračuna informacijskog dobitka za svaki atribut, odabire se onaj s najvećom vrijednošću, i tako kod svakog sljedećeg čvora odluke.

Prvi je informacijski dobitak kao mjeru za odabir atributa upotrijebio Quinland (1986) u svom algoritmu ID3 (Quinlan, Induction of decision trees 1986). To je algoritam koji radi samo s kvalitativnim varijablama, poprilično je jednostavan, a kao nedostaci navode se nepostojanje metoda skraćivanja stabla, nemogućnost rada s nedostajućim vrijednostima, laka pojava pretreniranosti (dobri rezultati na skupu za treniranje, a loši na stvarnim vrijednostima), činjenica da algoritam ne garantira dobivanje optimalnog rješenja i dr. Quinlan je stoga 1993 objavio poboljšani algoritam nazvavši ga C4.5 (Quinlan, C4. 5: programs for machine learning 2014). Glavna prednost je mogućnost skraćivanja stabla unatrag – uklanjanju grana koje ne pridonose točnosti modela, te podrška za kontinuirane i nedostajuće vrijednosti. C5.0 je komercijalna verzija ovog algoritma, a postoji još i Java implementacija koja se označava s J48 koja se koristi u alatu Weka (Witten, i dr. 2016).

Breiman je 1984. predstavio CART algoritam (Breiman, 2017) koji generira binarno stablo, koristi Gini čistoću kao mjeru za odabir atributa razdvajanja te koristi nešto kompleksnije postupke za skraćivanje stabla. Ostali nešto manje poznati algoritmi su CHAID (Kass 1980) i QUEST (Loh i Shih 1997).

Skraćivanjem stabla tj. rezanjem grana i njihovom zamjenom za listove dobivamo jednostavnije stablo koje je lakše interpretirati i koristiti. Jedna od glavnih prednosti stabla odlučivanja je mogućnost točnog utvrđivanja zašto je određeni uzorak (primjer) svrstan u određenu klasu. Primjena stabla odlučivanja nema nikakvih pretpostavki vezanih za distribuciju i strukturu podataka.

Nedostatak stabla odlučivanja je što pate od visoke varijance, odnosno od visokog postotka pogreške kada se primjenjuju na testne podatke (nekorištene u treniranju). Taj problem se može riješiti povećanjem uzorka za treniranje, što često nije moguće jer su dostupni podaci ograničeni. U tom slučaju pomoći nam mogu neke od tehnika ponovnog uzorkovanja. Najpoznatije su treniranje nad poduzorcima skupa za učenje (engl. *bagging*) (Breiman, 1996) i slijedno učenje algoritama na temelju pogrešaka prethodnih algoritama (engl. *boosting*) (Freund i Schapire 1996). Cilj obje tehnike je konstruirati više različitih modela iz istog skupa za učenje na način da se iz skupa za učenje kreira N uzoraka s ponavljanjem, što znači da jedna instanca može biti izabrana u više uzoraka. Rezultat je N modela i N predikcija, a kao konačan rezultat se uzima njihov prosjek. Ta tehnika podrazumijeva dakle upotrebu više slabijih modela čijom kombinacijom se na kraju dobije jači model. Razlika između treniranja nad poduzorcima i slijednog učenja je u tome što treniranje nad poduzorcima paralelno razvija

N modela nezavisno jedan od drugog, dok slijedno učenje za svaki sljedeći model od *N* uzima u obzir i rezultate prethodnog modela (Breiman, 1996; Freund i Schapire 1996).

U kontekstu stabla odlučivanja tehnike ponovnog uzorkovanja koriste se u algoritmu slučajnih šuma (engl. *random forest*) (Liaw i Wiener 2002). Algoritam slučajnih šuma je vrlo sličan treniranju nad poduzorcima uz jednu razliku: u kreiranju *N* uzoraka ne uzima u obzir sve nezavisne varijable (atribute), nego samo dio. Time se smanjuje utjecaj snažnih prediktora (koji bi bili izabrani u čvorove u većini *N* stabala) i generiraju manje korelirana stabla koja na kraju daju bolji prosječni rezultat.

3.1.2. Naivan Bayesov klasifikator

Naivan Bayesov klasifikator počiva na Bayesovoj teoriji uvjetne vjerojatnosti, a polazi od pretpostavke da su svi atributi (varijable) međusobno neovisni i jednako važni (Larose i Larose 2015). Često se koristi u domeni pretraživanja informacija. Bayesovo pravilo glasi:

$$\mathbf{P}(\boldsymbol{A}|\boldsymbol{B}) = \frac{\mathbf{P}(\boldsymbol{B}|\boldsymbol{A})\,\mathbf{P}(\boldsymbol{A})}{\mathbf{P}(\boldsymbol{B})} \qquad (4)$$

gdje se P(*A*|*B*) čita kao uvjetna vjerojatnost događaja *A* uz date (poznate) dokaze *B*.

Primijenimo li ovo pravilo na klasifikaciju onda *A* predstavlja pripadnost određenoj klasi, a *B* je vektor nezavisnih varijabli *x*, pa Bayesovu formulu možemo prikazati:

$$\mathbf{P}(\boldsymbol{y}|\boldsymbol{X}) = \frac{\mathbf{P}(\boldsymbol{X}|\boldsymbol{y})\,\mathbf{P}(\mathbf{y})}{\mathbf{P}(\boldsymbol{X})} \qquad (5)$$

gjde je *y* zavisna varijabla (pripradnost klasi), a *X* vektor nezavisnih varijabli (atributa).

Cilj je dakle izračunati vjerojatnost pripadnosti određenoj klasi ako su poznate vrijednosti ostalih varijabli tj. atributa. Po tome se i ovaj klasifikator razlikuje od ostalih – kao rezultat daje vjerojatnost pripadnosti klasi i odabire onu klasu s najvećom vjerojatnošću. Kod binarne klasifikacije (dvije klase) Bayesov klasifikator će predvidjeti prvu klasu ako je

$P(Y=1|X=x_0)>0{,}5$

gdje je x_0 vektor testnog uzorka, a *Y* je zavisna varijabla kojom je definirana pripadnost klasi. Ako je vjerojatnost manja od 0,5 predvidjet će drugu klasu.

Bayesov pristup, kao što vidimo iz formule (5) zahtijeva poznvanje apriorne vjerojatnosti klase P(y), drugim riječima vjerojatnosti pripadnosti klasi bez poznvanja vektora nezavisnih

varijabli X. Apriorna vjerojatnost može biti poznata (izračunata iz podataka) ili nepoznata. U slučaju da je nepoznata može se pretpostaviti na osnovu znanja eksperata u području u kojem se istraživanje radi, ili postaviti na jednake vjerojatnosti po klasama (npr. kod binarne klasifikacije može se postaviti na 0,5 za svaku klasu). Nakon što su podaci promatrani (nezavisne varijable), apriorno znanje o distribuciji Y se može ažurirati što dovodi do aposteriorne distribucije $P(y|X)$ gdje je X vektor promatranih podataka tj. nezavisnih varijabli. Iz aposteriorne distribucije biramo y koji maksimizira $P(y|X)$ kako bismo primjer klasificirali u klasu s najvećom vjerojatnošću, što se naziva MAP metoda (engl. *maximum a posteriori*). U Bayesovoj formuli (5) $P(X)$ predstavlja normalizacijski faktor, koji je zapravo konstanta u određenom istraživanju za dane podatke, stoga se može zanemariti, što dovodi do izraza za MAP funkciju:

$$\boldsymbol{y_{MAP} = arg\ max_y\ p(X|y)p(y)} \tag{6}$$

gdje $p(X|y)$ predstavlja izglednost klase tj. vjerojatnost primjera u klasi što je funkcija umnoška. Problem je u izračunavanju $p(X|y)$ jer ovisi o dimenzionalnosti X. Potrebno je izračunati $p(X_1=x_1,X_2=x_2,\ldots,X_m=x_m|y)$ za sve moguće kombinacije varijable x. To znači da je za k klasa i m prediktora potrebno izračunati k^m vjerojatnosti što uvelike otežava proces. Riješenje tog problema je pojednostavljenje s pretpostavkom da su nezavisne varijable (prediktori) uvjetno nezavisni prema zavisnoj varijabli (klasi). U tom slučaju je broj vjerojatnosti koje je potrebno izračunati $k \cdot m$ umjesto k^m, što je znatno manje. MAP funkcija za Naivni Bayesov klasifikator se na osnovu pretpostavke uvjetne nezavisnosti prediktora može prikazati kao:

$$\boldsymbol{y_{NB} = arg\ max_y \prod_{i=1}^{m} p(X_i = x_i\,|y)p(y)} \tag{7}$$

Bez obzira na ovu naivnu pretpostavku da su svi prediktori nezavisni (što često nije točno), naivni Bayesov klasifikator u praksi pokazuje dobre rezultate (Larose i Larose 2015).

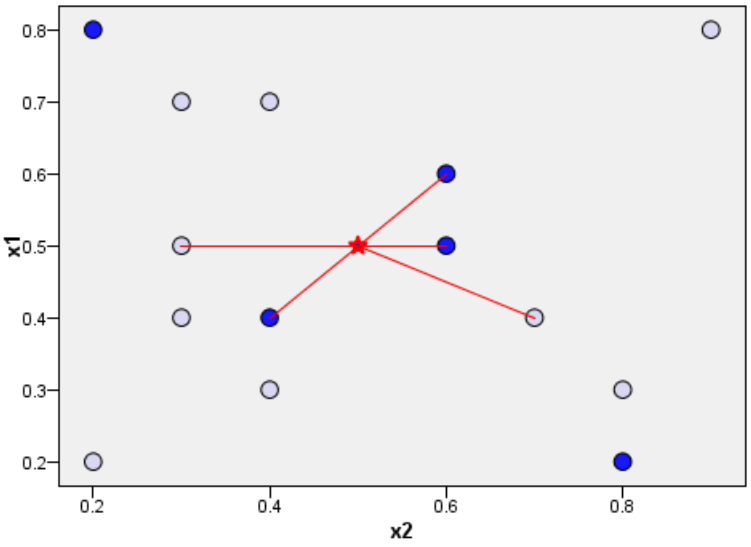
3.1.3. Metoda najbližih susjeda

Metoda najbližih susjeda ili KNN metoda (engl. *k-nearest neighbour*) je metoda klasifikacije koja kao rezultat ne daje model (Cover i Hart 1967). Ova metoda spada u tzv. „lijene algoritme“ jer se pripadnost klasi određuje u trenutku obrade novog podatka. Radi na principu

izračunavanja udaljenosti između novog podatka i svih ostalih podataka u *n*-dimenzionanom prostoru (gdje je *n* broj varijabli tj. atributa). Pri određivanju pripadnosti klasi u obzir se uzima *k* susjeda (najčešće *k* je između 3 i 10) – novi uzorak biti će klasificirana u onu klasu u kojoj je većina *k* susjeda tj. najbližih točaka. Budući da se ova metoda zasniva na izračunu udaljenosti od svake točke, taj proces je kod velikih skupova podataka izuzetno spor. Za korištenje ove metode potrebno je i odabrati mjeru udaljenosti koja će se koristiti. Najčešće je to euklidska udaljenost koja predstavlja drugi korijen iz sume razlike kvadrata između točaka *n*-dimenzionalnom prostoru dana formulom (8),

$$\boldsymbol{d(q,p)} = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{(q_i - p_i)^2}} \tag{8}$$

pri čemu su p i q uređene n-torke. Ostale mjere udaljenosti koje se mogu koristiti u KNN metodi su Manhattan udaljenost, Chebycheva, Minkowski, Mahalanobisova udaljenost i dr.



Slika 3.2 Primjer KNN metode s K=5

Slika 3.2 prikazuje primjer binarne klasifikacije s dvije nezavisne varijable (*x*1 i *x*2) KNN metodom gdje je *K*=5. Nova točka (označena crveno) tj. novi uzorak će se klasificirati u onu klasu tako što će se izračunati udaljenost od svih točaka i uzeti u razmatranje najbližih 5 susjednih točaka. Od tih 5 odabrat će se najčešća klasa. U primjeru na slici 3.2 to je plava klasa (3 plava susjeda naprema 2 iz bijele klase).

Odabir vrijednosti *K* je na istraživaču. Potrebno je testirati različite *K* vrijednosti kako bi pronašli onu koja daje najbolje rezultate. Uobičajeno je uzeti *K* neparan kako rezultat ne bi

bio izjednačen. Odabir optimalnog *K* tema je kojom se znanstvenici u ovom području bave. Jedan od posljednjih prijedloga je opisan u (García-Pedrajas, del Castillo i Cerruela-García 2017).

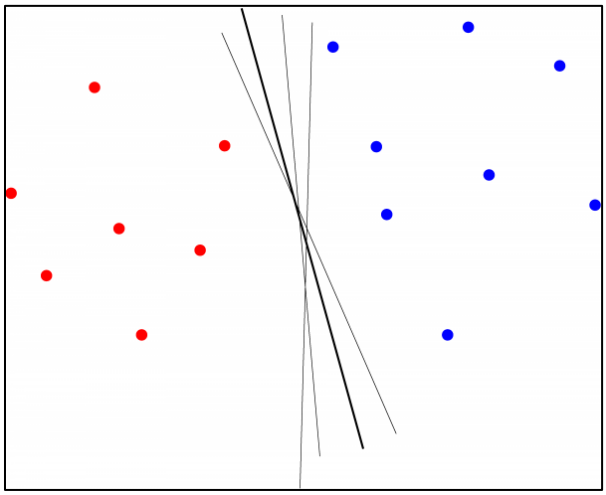
3.1.4. Metoda potpornih vektora

Problem klasifikacije u *n*-dimenzionalnom prostoru je zapravo problem pronalaska tzv. decizijske granice (linearne ili nelinearne) koja najbolje razdvaja uzorke iz skupa za učenje u predodređene klase. Decizijska granica je pravac u dvodimenzionalnom prostoru, ravnina u trodimenzionalnom, a hiperravnina u *n*-dimenzionalnom prostoru. Pronađemo li takvu funkciju koja predstavlja tu granicu, lako možemo odrediti klasu novog uzorka s obzirom s koje strane granice pada. Matematički hiperravninu možemo prikazati kao:

$$Y = w_0 + w_1 x_1 + w_2 x_2 \ldots + w_n x_n \qquad (9)$$

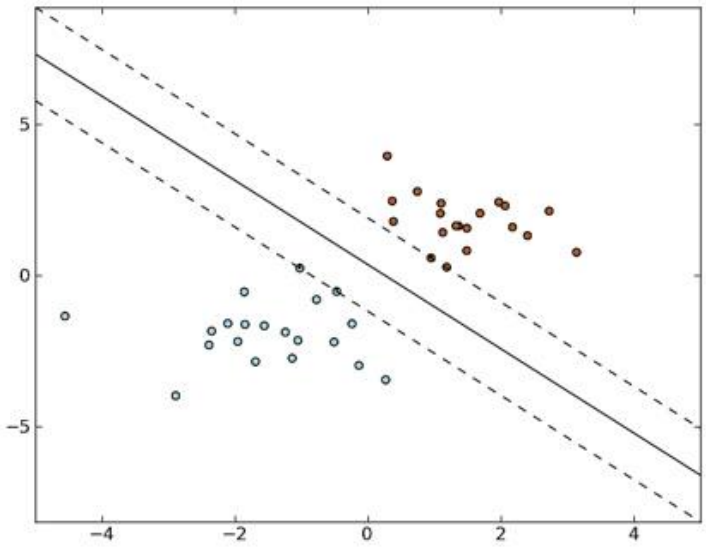
gdje su w_1, w_2, ..., w_n težinski parametri, a x_1, x_2, ..., x_n nezavisne varijable u *n*-dimenzionalnom prostoru. Ako je $Y>0$ onda ta točka pada s jedne strane hiperravnine i kod binarne klasifikacije može joj se dodijeliti pozitivna klasa. Ako je $Y<0$ onda ona pada s druge strane hiperravnine i pripada joj negativna klasa. Ako je $Y=0$ onda točka pada točno na hiperravninu.

Ako su klase linearno razdvojive u *n*-dimenzionalnom prostoru, onda postoji i više mogućih hiperravnina koje ih razdvajaju kao što je prikazano na slici 3.3.



Slika 3.3 Primjer mogućih razdvajajućih pravaca u 2-dimenzionalnom prostoru kod binarne klasifikacije

Metoda potpornih vektora (engl. *support vector machines* ili SVM) bavi se pronalaskom optimalne hiperravnine koja će najbolje razdvojiti uzroke iz skupa za učenje na način da udaljenost od nje i prvih susjednih točaka bude maksimalna. Te susjedne točke nazivaju se potporni vektori jer o njima zapravo ovisi pozicija granice i margina. Promjena u točkama iza potpornih vektora (margina) ne utječe na poziciju granice razgraničenja (Slika 3.4).



Slika 3.4 Primjer potpornih vektora (točke na isprekidanim linijama) u dvodimenzionalnom prostoru.

Na slici 3.4 margine su prikazane isprekidanom linijom, a razdvajajuća hiperravnina punom linijom. Može se vidjeti da svaka margina ima dva potporna vektora, a udaljenost između margina i hiperravnine je maksimalna. Ta metoda klasifikacije se još naziva metoda maksimiziranja margina (engl. *maximal margin classifier*). Njena uloga je odabir optimalne hiperravnine od mogućih hiperravnina, a logično je odabrati onu koja je najudaljenija od točaka iz skupa za učenje. Nove točke iz skupa za testiranje klasificiramo u onu klasu s obzirom na to s koje strane razdvajajuće hiperravnine padaju.

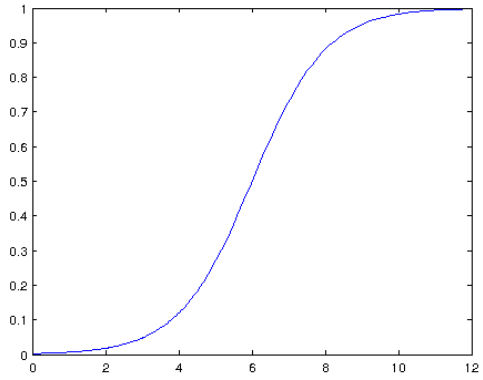
Ova metoda dobra je ako su uzorci iz skupa za učenje linearno razdvojivi tj. moguće je pronaći barem jednu hiperravninu koja razdvaja klase. Međutim, što ako to nije moguće, tj. ako ne postoji razdvajajuća hiperravnina? Tada možemo proširiti ovaj koncept na tzv. meke margine tj. hiperravnine koje ne razdvajaju savršeno klase iz skupa za učenje tj. dozvoljava se određeni postotak greške (netočne klasifikacije). Takav klasifikator se naziva klasifikator temeljen na potpornim vektorima (engl. *support vector classifier*). Treći slučaj klasifikacije je

kada razdvajajuća hiperravnina nije linearna, i takva vrsta klasifikacije se naziva metoda potpornih vektora (engl. *support vector machines*).

Za slučajeve s više od dvije klase metoda potpornih vektora se rjeđe koristi, ali je izvediva na dva načina: prvi način je usporedba klasa u parovima (tzv. „jedan naprema jedan“) gdje se uzorak svrstava u onu klasu u koju je svrstana najviše puta; drugi način je putem usporedbe jedne klase s ostatkom (tzv. „jedan naprema više“) (Hsu i Lin 2002).

3.1.5. Logistička regresija

Logistička regresija počiva na upotrebi logističke funkcije za klasifikaciju. Razlog upotrebe logističke funkcije leži u njenom „S“ obliku (u dvodimenzionalnom prostoru, Slika 3.5) koji omogućava kreiranje klasifikacijskog modela koji će dati vjerojatnost pripadnost klasi u rasponu od 0 do 1 (ako govorimo o binarnoj klasifikaciji).



Slika 3.5 Logistička funkcija

Naime, upotrijebimo li klasičnu linearnu regresiju za binarnu klasifikaciju na način da kodiramo prvu klasu s „0“, a drugu klasu s „1“, može se dogoditi da vjerojatnost (kao rezultat regresijske funkcije) bude manji od 0 ili veći od 1, što zapravo nema smisla, tj. teško je za interpretirati. Zbog toga nam treba funkcija koja će dati rezultate u rasponu od 0 do 1, a jedna od takvih je logistička funkcija (James, i dr. 2013, 131).

Klasični linearni regresijski model možemo prikazati kao:

$$p(x)=w_0+w_1x_1 \tag{10}$$

gdje su w_0 i w_1 težinski koeficijenti, a x_1 nezavisna varijabla. Rezultat ovakvog modela može

biti izvan raspona od 0 do 1 te je zbog toga za potrebe klasifikacije bolje koristiti logistički model:

$$l(x) = \frac{e^{w_0+w_1x_1}}{1 + e^{w_0+w_1x_1}} \tag{11}$$

Logistička regresija je klasifikacijski model koji za granicu razdvajanja točaka koristi sigmoidalnu (logističku) funkciju. Najčešće se koristi kada je zavisna varijabla dihotomna (može poprimiti dvije vrijednosti). U linearnom regresijskom modelu težinski parametar w_1 predstavlja prosječnu promjenu zavisne varijable ako se nezavisna (x_1) promjeni za jednu jedinicu – razlog tome leži u linearnoj zavisnosti između $p(x)$ i x. U logističkom modelu nema linearne zavisnosti između $l(x)$ i x, već njihova promjena ovisi o vrijednosti x. Ako je w_1 pozitivan, povećanje x dovesti će do povećanja $l(x)$, te ako je w_1 negativan, povećanje x dovesti će do smanjenja $l(x)$.

Težinski koeficijenti w_0 *i* w_1 u linearnom regresijskom modelu izračunavaju se tj. procjenjuju na temelju podataka iz skupa za učenje (postupak treniranja) pomoću metode najmanjih kvadrata. Ta metoda nije pogodna u logističkom modelu, te su umjesto nje koristi metoda maksimalne vjerodostojnosti koja pokušava procijeniti w_0 i w_1 na način da $l(x)$ za pozitivne klase bude što bliža vrijednosti 1, a za negativne klase bliža nuli (što kod linearnog modela nije moguće).

Binarni logistički model koji koristi više nezavisnih varijabli naziva se multipla logistička regresija koju možemo predstaviti jednostavnim proširenjem modela (11):

$$l(x) = \frac{e^{w_0+w_1x_1+\cdots+w_nx_n}}{1 + e^{w_0+w_1x_1+\cdots+w_nx_n}} \tag{12}$$

gdje se težinski parametri w_0, w_1, ... w_n procjenjuju metodom maksimalne vjerodostojnosti. Logistička regresija se može primijeniti i na situacije kada imamo više od dvije klase, kao što je slučaju o ovom radu – klase su: prilagođeno, djelomično prilagođeno i neprilagođeno. U tom slučaju modeliramo P(Y=prilagođeno|X) i P(Y=djelomično prilagođeno|X), s ostatkom P(Y=neprilagođeno|X) = 1 – P(Y=prilagođeno|X) – P(Y=djelomično prilagođeno|X).

3.2. Mjere uspješnosti i evaluacija klasifikatora

Nakon izgradnje modela klasifikatora potrebno je izračunati njegovu točnost tj. uspješnost. Model se gradi na skupu za učenje (treniranje), a testiranje tj. evaluacija se treba vršiti na zasebnom skupu koji se naziva skup za validaciju ili testiranje. Pri tome su nam na raspolaganju dvije tehnike:

- Metoda izdvajanja (engl. *hold out*) – prije učenja formiraju se dva skupa podataka, jedan za učenje i jedan za testiranje gdje se obično za učenje uzima dvije trećine, a za testiranje jedna trećina inicijalnog skupa podataka.
- Unakrsna validacija (engl. *cross-validation*) – gdje se validacija vrši koristeći skup za učenje na način da se on podijeli na *k* dijelova. Validacija modela se vrši *k* puta tako da se svaki put za učenje koristi *k-1* dijelova, a za validaciju 1 dio skupa podataka. Na kraju se izračuna prosjek mjera evaluacije za *k* klasifikacijskih modela kao konačna metrika za evaluaciju.

Bez obzira na odabranu metodu, nekoliko je metrika koje nam govore o tome koliko je određeni klasifikator dobar: točnost, preciznost, odaziv, F_1, MAE (srednja apsolutna pogreška) i MSE (srednja kvadratna pogreška).

Točnost (engl. accuracy) je udio točno klasificiranih primjera u skupu svih primjera:

$$Acc=(TP+TN)/(TP+TN+FP+FN) \quad (13)$$

gdje su:

TP = točno pozitivan (engl. *true positive*) – pozitivni primjeri točno klasificirani kao pozitivni

TN = točno negativan (engl. *true negative*) – negativni primjeri točno klasificirani kao negativni

FP = netočno pozitivan (engl. *false positive*) – negativni primjeri krivo klasificirani kao pozitivni

FN = netočno negativan (engl. *false negative*) – pozitivni primjeri krivo klasificirani kao negativni

Mjera suprotna točnosti je stopa pogrešne klasifikacije (engl. *misclassification rate*) što je 1-Acc, ili

$$Err= (FP+FN)/(TP+TN+FP+FN). \quad (14)$$

Preciznost (engl. *precision*) je udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera (ili drugim riječima kada je predikcija pozitivna, koliko često je to točno):

$$P=TP/(TP+FP) \tag{15}$$

Odaziv (engl. *recall*) je udio točno klasificiranih primjera u skupu svih pozitivnih primjera (ili drugim riječima koliko često se predviđa pozitivna klasifikacija kada ona stvarno jest pozitivna):

$$R=TP/(TP+FN) \tag{16}$$

F_1-mjera kombinira mjere preciznosti i odaziva, a računa se kao njihova harmonijska sredina:

$$F_1=2PR/(P+R) \tag{17}$$

F_1-mjera je pogodna za binarne klasifikacijske probleme gdje možemo odrediti pozitivnu i negativnu klasu. Kod višeklasne klasifikacije F-mjera i omjeri pozitivnih i negativnih klasa su mogući samo ako promatramo klase u paru ili promatramo jednu klasu naspram svih ostalih.

Klasifikatori se često validiraju i pomoću mjera srednje apsolutne pogreške (engl. *mean absolute error*, MAE) i srednje kvadratne pogreške (engl. *mean squared error*, MSE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)| \tag{18}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{19}$$

gdje je *n* broj uzoraka, *y* stvarna klasa, a *f(x)* predviđena klasa za primjer *x*.

Te mjere predstavljaju prosječnu (apsolutnu ili kvadriranu) razliku između predikcije i stvarne vrijednosti klase. Linearna korelacija između predikcije i stvarne klase također je mjera uspješnosti klasifikatora (visoka korelacija ukazuje na dobar klasifikator). U slučaju ordinalne klasne varijable, MAE i MSE su najčešće korištene metrike (Gaudette i Japkowicz 2009).

MSE će imati niske vrijednosti ako je klasifikator dobar, tj. ako je predviđena klasa blizu stvarnoj. Ovu je mjeru potrebno izračunavati nad podacima za testiranje (modelu prethodno neviđenim podacima) jer se tako osigurava njegova pouzdanost.

Kod analize rezultata klasifikatora često se koristi tzv. matrica zabune (engl. *confusion matrix*). Ona prikazuje broj točnih i netočnih klasifikacija (Tablica 3.1).

Tablica 3.1 Matrica zabune

		Predviđena vrijednost	
		Pozitivna	Negativna
Stvarna vrijednost	Pozitivna	Točno pozitivna (TP)	Netočno negativna (FN)
	Negativna	Netočno pozitivna (FP)	Točno negativna (TN)

Matrica zabune će imati onoliko redaka i stupaca koliko imamo klasa – u slučaju ovog istraživanja je to tri. Na lijevoj dijagonali matrice nalaze se točno klasificirani uzorci (TP i TN), a izvan dijagonala su pogreške. Matrica zabune nam omogućava da jednostavno vidimo koliko uzoraka je klasificirano u koju, točnu ili pogrešnu, klasu.

3.3. Optimizacija hiperparametara i regularizacija

Većina algoritama strojnog učenja imaju parametre koje istraživač može podešavati kako bi poboljšao performanse svog modela. Ti parametri se nazivaju *hiperparametri* kako bi se razlikovali od ostalih parametara koji se uče u samom algoritmu (npr. tu spadaju težinski parametri w i sl.). Podešavanje hiperparametara dio je procesa regularizacije koji se bavi sprječavanjem tzv. pretreniranosti (engl. *overfitting*) i podtreniranosti (engl. *underfitting*). Problem pretreniranosti označava model kod kojega vidimo jako dobre performanse na skupu za učenje, a loše na skupu za testiranje, tj. model je pretjerano istreniran i dobro pristaje podacima pomoću kojih je istreniran. To dovodi do loše generalizacije tj. primjene modela na drugim (stvarnim) podacima koji se koriste kod testiranja modela. Pretreniranost zapravo znači da je model previše dobro naučio sve šumove i nerelevantne detalje iz podataka za treniranje tako da nad stvarnim neviđenim podacima (iz iste domene) više griješi.

Podtreniranost se događa kada model ne predstavlja podatke za treniranje dovoljno dobro, ali i loše generalizira. Performanse su dakle loše i na podacima za treniranje i na podacima za testiranje tj. realnim podacima. Problem podtreniranosti je lakše detektirati jer nam to signaliziraju mjere evaluacije modela strojnog učenja.

Pretreniranost je puno veći i češći problem u domeni strojnog učenja. Može se riješiti ponovnim uzorkovanjem (učenjem i testiranjem modela na više uzoraka), testiranjem na posebnom skupu podataka (za oba slučaja nužno je imati više podataka, što je ponekad teško), upotrebom unakrsne validacije, kombiniranjem više modela različitih algoritama, optimizacijom hiperparametara algoritma i sl. Ponekad je dovoljno pojednostaviti model (npr. rezanjem grana kod stabla odlučivanja) kako bi umanjiili pretreniranost.

Kako bi spriječili pretreniranost većina algoritama ima hiperparametre koje istraživač može podešavati.

Kod stabla odlučivanja hiperparametri koji se najčešće optimiziraju su (Mantovani, i dr. 2016):

- parametar pouzdanosti (*cp*) – predstavlja pesimističnu gornju granicu postotka pogreške na razini lista ili čvora koja se koristi kod skraćivanja stabla. Što je vrijednost ovog parametra manja, skraćivanje je jače;
- minimalni broj uzoraka na listu (*minBucket*) – manja vrijednost ovog parametra generirat će veće (kompleksnije) stablo;
- maksimalna dubina stabla (*maxDepth*) – utječe na veličinu cjelokupnog stabla.

Stabla su sklona pretreniranosti, ali uklanjanjem grana koje idu previše u detalje i ne doprinose mnogo ukupnoj točnosti stabla taj problem se može smanjiti. Uklanjanje grana naziva se *rezanje stabla* (engl. *pruning*) kako bi ono postalo jednostavnije i bolje generaliziralo.

Kod algoritma potpornih vektora najčešći hiperparametri za optimiziraju su (Ben-Hur i Weston 2010):

- parametar kompleksnosti odnosno regularizacije C – manji C dovodi do povećanja margina na način da ignorira točke koje su bliže hiperravnini razdvajanja dok veći C postiže suprotni efekt - smanjuje margine;

- tip i parametri jezgrene (engl. *kernel*) funkcije – kod nelinearnih razdvajanja koristi se jezgrena funkcija određenog stupnja polinoma. Viši stupnji polinoma omogućuju fleksibilniju granicu razdvajanja.

Metoda najbližih susjeda podrazumijeva optimizaciju hiperparametra *k* tj. broja susjeda i metrike udaljenosti (euklidska, Manhattan ili dr.).

Logistička regresija koristi tzv. *parametre penalizacije* koji se još nazivaju i *regularizacijskim parametrima*. Dvije su vrste regularizacijskih tehnika u regresijskim modelima: L1 i L2. Model koji koristi L1 regularizaciju naziva se još Lasso regresija, a L2 Ridge regresija. Razlika je u parametru penalizacije. L1 dodaje apsolutnu vrijednost magnitude koeficijenta *funkciji pogreške* (engl. *cost function*) kao penalizaciju, dok L2 dodaje kvadriranu vrijednost. U linearnim modelima funkcija pogreške predstavlja razliku između predviđenih i stvarnih vrijednosti zavisne varijable *Y:*

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \tag{20}$$

gdje je *m* broj primjera (instanci), $h_\theta(x^{(i)})$ predviđena vrijednost za primjer *i*, a $y^{(i)}$ stvarna vrijednost. Cilj algoritma je minimizirati razlike između predviđene i stvarne vrijednosti *Y*, tj. pronaći koeficijente w_i koji minimiziraju funkciju pogreške kako bi dobili model sa što boljim performansama. Kako bi spriječili pretreniranost funkciji troška se dodaje regularizacijski izraz:

$$\lambda\sum_{j=1}^{p}|w_j|^q \tag{21}$$

gdje je λ regularizacijski faktor, *p* broj težinskih koeficijenata, w_i težinski koeficijenti, a *q* eksponent s kojim određujemo da li radimo L1 regularizaciju (*q*=1) ili L2 (*q*=2). Dodavanjem izraza (21) funkciji pogreške (20) dobivamo mogućnost da sa većim λ više kažnjavamo složenije modele jer oni imaju veće apsolutne iznose težinskih koeficijenata *w*. Kod logističke regresije se za spriječavanje pretreniranosti češće koristi L2 regularizacija.

Metode optimizacije hiperparametara su različite (Bergstra, i dr. 2011). Proces treniranja klasifikatora ne optimizira hiperparametre, stoga je potreban poseban proces za taj zadatak. Optimalni hiperparametri nisu isti za sve skupove podataka tj. jednom optimizirani

hiperparametri neće tj. ne moraju biti optimalni na drugim skupovima podataka s istim, sličnim ili različitim klasifikacijskim problemima. Zbog toga je optimizacija hiperparametara nužan proces. Rezultat tog procesa su vrijednosti parametara koji daju najbolje performanse modela (mjerene odabranom metrikom evaluacije). Sam proces se također zasniva na metodi treniranja i evaluacije (unakrsne ili izdvajajuće). Hiperparametri imaju svoj raspon vrijednosti koje mogu poprimiti, a kojeg proces optimizacije ispituje na način da mjeri performanse modela s različitim kombinacijama hiperparametara.

Dvije su najraširenije metode optimizacije hiperparametara: mrežna pretraga (engl. *grid search*) i slučajna pretraga (engl. *random search*). Mrežna pretraga radi na principu da ispituje sve moguće vrijednosti hiperparametara (i njihove kombinacije) u definiranom rasponu kojeg odredi istraživač (a koji treba biti unutar mogućeg raspona za dotični hiperparametar). Najčešće se najprije definira širi raspon da se utvrdi područje koje pokazuje dobre rezultate te se zatim detaljnije optimizira u rasponu tog područja. Budući da mrežna pretraga ispituje sve moguće vrijednosti i kombinacije njezina glavna mana je sporost. Metoda koja može pomoći u rješavanju tog problema je metoda slučajne pretrage. Ta metoda ne ispituje sve vrijednosti u definiranom rasponu nego samo određene, slučajno odabrane. To znači da je moguće da rezultat slučajne pretrage, tj. vrijednosti hiperparametara koje on predloži nisu stvarno optimalni jer je moguće da je proces zbog slučajnog odabira ispustio ispitati stvarno optimalne vrijednosti.

Optimizacija hiperparametara dakle uključuje sljedeće korake:

- odabir hiperparametara koji će se optimizirati
- odabir raspona vrijednosti za svaki odabrani hiperparametar
- optimizacija (mrežna pretraga ili slučajna pretraga ili neki drugi algoritam)
- upotreba rezultata optimizacije tj. ponovno treniranje modela strojnog učenja s dobivenim vrijednostima hiperparametara te evaluacija

Testiranje hiperparametara i modela se vrši pomoću ugniježđene unakrsne validacije koja radi nad tri skupa i ima dvije ugniježđene petlje. Skupovi su skup za učenje (engl. *training set*), skup za validaciju (engl. *validation* set) i skup za ispitivanje (engl. *test set*). Vanjska petlja služi za učenje i testiranje modela *k* puta, dok unutarnja testira hiperparametre *n* puta kako bi za svaku *k* iteraciju odabrala najbolje hiperparametre. Drugim riječima, u svakoj od *k* iteracija skup podataka za učenje se dodatno dijeli na *n* dijelova gdje se *n-1* dijelova koristi za treniranje u potrazi za optimalnim hiperparametrima, a 1 dio za validaciju odabranih

parametara (skup za validaciju). Nakon odabira modela isti se uči na skupu za učenje i validaciju, a testira se na skupu za testiranje. Odabir modela vrši se na temelju prosjeka pogreške.

Navedeni proces upotrijebljen je i u ovom istraživanju.

4. OBRADA PRIRODNOG JEZIKA

Obrada prirodnog jezika (engl. *natural language processing, NLP*) se najčešće koristi za opisivanje funkcije softvera ili hardvera računalnog sustava koji analizira ili sintetizira jezik u govoru ili pismu (Jackson i Moulinier, 2002). Termin *prirodni jezik* koristi se kako bi se razlikovali od ostalih oblika jezika kao npr. programskih jezika, logičkih i matematičkih izraza i sl. Govorimo dakle o sustavima za obradu jezika kojim se koriste ljudi u svakodnevnoj komunikaciji, govoru i pismu. Današnja računala osposobljena su da razumiju programske jezike – softverski interpreteri razvijeni su da razumiju predefinirane naredbe i izvršavaju ih. Postoje i aplikacije koje razumiju matematičke izraze i formule te mogu rješavati matematičke zadatke. No, kad je u pitanju razumijevanje prirodnog jezika od strane računala, nailazimo na brojne poteškoće te su takvi sustavi još uvijek u razvoju.

Svrha obrade prirodnog jezika može biti višestruka. Najčešće se koristi u cilju ekstrakcije informacija iz teksta, strojnog prevođenja, analize osjećaja, sažimanje teksta i sl. Razlog zašto je obrada prirodnog jezika problematična leži u poteškoćama utvrđivanja značenja riječi i fraza koje se koriste u jeziku, a čije značenje najčešće ovisi o kontekstu. Različiti jezici koriste različite gramatičke i sintaktičke strukture, ljudi u svom govoru koriste različite dijalekte i riječi koje se ne nalaze u službenim rječnicima (žargon), te koriste sinonime (različite riječi za isto značenje), homonime (iste riječi s različitim značenjem koje su etimološki nepovezane) i poliseme (iste riječi s više različitih međusobno povezanim značenjima). Zbog svega toga izuzetno je teško računalu utvrditi smisao i pravo značenje teksta ili dijelova teksta.

U kontekstu ovog istraživanja obrada prirodnog jezika koristi se za obradu teksta na mrežnim stranicama. Prema SEO preporukama (kako je objašnjeno u poglavlju 2) tekst na mrežnim stranicama izuzetno je važan u SEO procesu. Same ključne riječi nalaze se naravno unutar teksta. Tekst se osim na tijelu stranice koristi i u „meta opisu“ stranice, a fraze se mogu nalaziti i na nekim drugim pozicijama unutar HTML dokumenta (npr. atributi „ALT“ i „TITLE“).

Obogaćivanje teksta ključnim riječima jedna je od zadaća ovog istraživanja. Također generiranje sažetaka za upotrebu u „meta opisu“ za one mrežne stranice koje nemaju „meta opis“ tehnika je koja će se upotrijebiti u ovom istraživanju. Postoje različiti softverski alati za

NLP, a jedan od najpopularnijih je NLTK[4] (Loper i Bird 2002) koji se koristi uz pomoć programskog jezika Python[5]. Ovi alati koriste se u ovom istraživanju.

Slijedi kratki pregled tehnika iz domene obrade prirodnog jezika koji se koriste u ovom radu.

4.1. Tokenizacija, normalizacija i segmentacija teksta

Početak svake obrade teksta uključuje tehnike pred-obrade, parsiranja i čišćenja teksta. Npr. ako se radi o HTML tekstu potrebno je izbaciti određene HTML oznake, *javascript* dijelove, CSS notacije i slično, ovisno o cilju obrade. Ako se tekst nalazi u XML-formatu onda nas možda interesiraju samo dijelovi teksta u određenim oznakama, pa će proces čišćenja uključivati procedure za izvlačenje tih dijelova teksta. U ovoj fazi pred-obrade često se koriste regularni izrazi (engl. *regular expressions*) za dohvaćanje određenih dijelova teksta i ekstrakciju podataka.

Opojavničenje (engl. *tokenization*) je proces razdvajanja teksta na njegove manje dijelove koji se nazivaju *pojavnice* (engl. *token*). Pojavnicom se najčešće smatra jedna riječ, ali može se sastojati i od dvije ili više riječi ako se radi o frazi ili nazivlju koji se u jeziku tako i koristi. Pojavnica je dakle skup znakova (slova) koji tvore jednu semantičku cjelinu. S druge strane, *tip* je skup svih pojavnica koji sadrže isti znakovni poredak. Tip je zapravo element rječnika, dok je pojavnica primjerak tipa u tekstu. Na primjer, promotrimo rečenicu:

„Mrežna stranica Hewlet Packarda je bolja od moje mrežne stranice.“

Ova rečenica ima 10 riječi, ali 9 pojavnica – „Hewlet Packard“ je naziv tvrtke i kao takav se smatra jednom pojavnicom. „Mrežna stranica“ također bi mogla biti jedna pojavnica, jer semantički označava jednu pojavu. U tekstu se nalaze i dvije riječi „mrežna“ i „mrežne“, te „stranica“ i „stranice“ koje imaju isti znakovni poredak (razlika je samo u jednini/množini), te se kao takve smatraju jednim tipom.

Različiti jezici imaju različite probleme opojavničavanja. U nekim jezicima se koriste apostrofi (npr. engleska riječ „aren't“, francuska „l'ensemble“), te se postavlja pitanje kako to opojavničavati. U većini slučajeva apostrofi ne razdvajaju riječ na dva tokena. U njemačkom jeziku postoji problem spajanja riječi: neki termini su sačinjeni od više riječi zajedno

[4] https://www.nltk.org/
[5] https://www.python.org/

spojenim (bez razmaka). U tom slučaju trebaju se najprije razdvojiti kako bi se izvukle pojavnice.

Da bismo mogli razdvojiti tekst na pojavnice ili rečenice potreban nam je znak koji će biti separator. Za rečenice je to točka ili neki drugi interpunkcijski znak koji označava kraj rečenice. No, pojava takvih interpunkcijskih znakova ne znači nužno kraj rečenice. Npr. točka se koristi kod kratica (npr. dr.sc.), ostali interpunkcijski znakovi mogu se koristiti u tekstu unutar zagrada i sl. Za razdvajanje po pojavnicama razmak također nije uvijek pouzdan separator (kao što je pokazano u prethodnim primjerima). Zbog toga su klasične procedure bazirane na regularnim izrazima nadograđene procesima nadgledanog strojnog učenja čime se pravila izvlače na osnovu prethodno ručno segmentiranog teksta određenog korpusa. Također procedure mogu uključivati označavanje vrste riječi (engl. *part-of-speech taging*) u tekstu fraza koje se često koriste u prirodnom jeziku.

U fazi pred-obrade teksta neizostavan proces je uklanjanje zaustavnih riječi (engl. *stop words*). To su riječi koje se često pojavljuju u dokumentima unutar korpusa, a njihovo uklanjanje ne utječe na ishode obrade teksta. Radi se najčešće o zamjenicama, članovima i veznicima (npr. u engleskom jeziku: „the“, „a“, „this“ itd.). Popis zaustavne riječi je najčešće ručno izrađen popis za određeni jezik, a može se i izgenrirati iz korpusa na temelju frekvencija pojave i pozicije u tekstu.

Proces normalizacije teksta uključuje svođenje riječi na osnovnu formu. Npr. „U.S.A.“ označava istu riječ kao i „USA“, množina riječi se svodi na jedninu, eliminiraju se gramatičke verzije riječi, mijenjaju se velika slova u mala i sl. Dio procesa normalizacije je i tzv. lematizacija i korjenovanje (engl. *stemming*).

4.1.1. Lematizacija

Lematizacija riječi je proces svođenja riječi na osnovnu formu tj. smanjivanje varijanti riječi. Ova forma naziva se još i forma riječnika – oblik riječi kako je naveden u riječniku. Lematizacija ovisi o kontekstu riječi tj. o upotrebi u jeziku. Npr. riječi „leti“, „letim“, „leteći“ su oblici riječi „letjeti“ u riječniku. Lematizacija dakle koristi morfološku analizu riječi, kontekst, te rječnike kako bi se riječ svela na njen osnovni oblik ili *lemu*.

4.1.2. Korjenovanje riječi

Korjenovanje (engl. *stemming*) je jednostavniji proces od lematizacije jer ne uzima u obzir kontekst, niti se povezuje s rječnikom. Radi se o uklanjanju prefiksa i/ili sufiksa u riječima koji su dodani iz gramatičkih razloga. Rezultat korijenovatelja može biti riječ koja nema značenje već predstavlja skup znakova (slova) koji se koriste u korijenu riječi. Najpoznatiji korijenovatelj za engleski jezik je Porterov korijenovatelj[6] (engl. *Porter stemmer*) osmišljen 1980. godine (Willett 2006). Drugi manje popularni korijenovatelji su Lovinsov (Lovins 1968) i Paice/Husk (Chris 1990).

4.2. Označavanje vrste riječi

Označvanje vrste riječi (engl. *part-of-speech tagging, POS tagging*) je proces kojim se svakoj riječi u tekstu dodjeljuje oznaka vrste tj. leksičke kategorije: imenica, glagol, pridjev, prilog, prijedlog, veznik, zamjenica, usklik, broj, čestica. Riječi mogu imati više vrsta, ovisno o kontekstu. Npr.:

„Danas je sunčan dan."

„Poklon je dan njemu."

Riječ „dan" je u prvoj rečenici imenica, a u drugoj glagol. Za uspješno POS označavanje bitan je dakle kontekst, odnosno susjedne riječi i vjerojatnost pojave riječi. Računalni sustavi za POS označavanje koriste prethodno ručno označene korpuse za određivanje vrste riječi u novom (neviđenom) tekstu. Najpoznatiji je Brownov korpus iz 1960. godine (Francis i Kucera 1964) koji predstavlja bogati skup tekstova na engleskom jeziku. Brownov korpus se i danas često koristi u istraživanjima u računalnoj lingvistici (Ng i Zelle 1997).

Današnji softveri za POS označavanje rezultate označavanja prikazuju na način da svakoj riječi pridodaju kraticu vrste riječi, npr.:

„John/?NP likes/VB the/DT blue/JJ house/NN at/IN the/DT end/NN of/IN the/DT street/NN"

Najčešće POS oznake prikazane su u Tablica 4.1.

[6] https://tartarus.org/martin/PorterStemmer/

Tablica 4.1 Kratice za najčešće korištene POS oznake

Kratica	**Značenje**
AT	Član (engl. article)
CC	Koordinacijska veza (engl. coordinating conjuction)
CD	Redni broj (engl. cardinal number)
DT	(engl. singular determiner)
IN	Prijedlog (engl. preposition)
JJ	Pridjev (engl. adjective)
NN	Imenica u jednini (engl. singular or mass noun)
NNS	Imenica u množini (engl. plural noun)
NP	Vlastita imenica ili ime (engl. proper noun or part of name phrase)
VB	Glagol (engl. verb)

Najpoznatiji je Stanfordov označavatelj.[7]

Česta procedura koja je slična POS označavanju je označavanje imenovanih entiteta (engl. *named entity recognition*, *NER*) koja predstavlja proces prepoznavanja nazivlja, imena osoba, geografskih lokacija i naziva organizacija u tekstu (Nadeau i Sekine 2007).

4.3. Matrica riječ-dokument i težinske funkcije

Nakon normalizacije tekst se za svrhu daljnje analize može predstaviti i matematički u vektorskom prostoru. Možemo konstruirati matricu riječ-dokument koja se sastoji od *i* redaka i *j* stupaca gdje w_i predstavlja riječ, a d_j dokument u kojem se ta riječ pojavljuje. Sadržaj matrice f_{ij} je frekvencija pojave riječi *i* u dokumentu *j*. Struktura matrice je prikazana na slici 4.1.

[7] https://nlp.stanford.edu/software/tagger.shtml

	d_1	d_2	...	d_j
w_1	f_{11}	f_{12}	...	f_{1j}
w_2	f_{21}	f_{22}	...	f_{2j}
...	...	...	...	...
w_i	f_{i1}	f_{i2}	...	f_{ij}

Slika 4.1 Matrica riječ-dokument

Značajnost određene riječi može se izračunati pomoću tzv. težinskih funkcija koje se baziraju na frekvenciji pojave riječi u dokumentu i u svim dokumentima ukupno (zbirci dokumenata). Riječima koje se često pojavljuju možemo smanjiti, dok riječima koje se rijetko pojavljuju možemo povećati značajnost. Najpoznatije funkcije za tu svrhu su TF, IDF i TF-IDF.

TF (engl. *term frequency*) pokazuje koliko se puta termin *t* pojavljuje u dokumentu *d*. Prema tome $tf(t,d)=f_{td}$ u najjednostavnijem obliku te funkcije. Postoje još sljedeći oblici:

- „Boolean“ oblik gdje je $tf(t,d)$=1 ako se termin *t* pojavljuje u dokumentu *d* (neovisno koliko puta), inače 0.
- Normalizirani oblik gdje je $tf(t,d)=f_{td} / n$, gdje je *n* broj riječi u dokumentu *d*
- Logaritamski oblik gdje je $tf(t,d)=1+log(f_{td})$ - koristi se za smanjivanje utjecaja čestih riječi
- Prošireni oblik gdje je

$$tf(t,d) = 0{,}5 + 0{,}5\,\frac{f_{td}}{max_{\{t\prime \in d\}} f_{t\prime d}} \tag{22}$$

koji predstavlja frekvenciju riječi podijeljenu s najvećom frekvencijom riječi u dokumentu (kako bi se smanjio utjecaj dužine dokumenata).

IDF (engl. *inverse document frequency*) predstavlja mjeru koja pokazuje značajnost riječi s obzirom na to da li je rijetka ili česta riječ na nivou zbirke dokumenata:

$$idf(t,D) = \log \frac{N}{1 + |\{d \in D : t \in d\}|} \tag{23}$$

gdje je N broj dokumenata u zbirci, a $|\{d \in D: t \in d\}|$ broj dokumenata gdje se pojavljuje riječ *t*.

TF-IDF se računa kao umnožak TF i IDF funkcije:

$$tf - idf(t, d, D) = tf(t, d) \cdot idf(t, D) \quad (24)$$

TF-IDF mjera će imati visoke vrijednosti za riječi (*t*) koje imaju visoku frekvenciju u dokumentu (*d*), a nisku frekvenciju u zbirci dokumenata (D). TF-IDF se često koristi u domeni pretraživanja informacija (engl. *information retrieval*) i u samim algoritmima tražilica za utvrđivanje pojmova koji mogu predstavljati temu mrežne stranice tj. ključnih riječi. Matematičko predstavljanje teksta u vektorskom prostoru korisno je i kod utvrđivanja sličnosti dokumenata gdje se koriste metrike za izračunavanje udaljenosti vektora.

4.4. Leksička baza WordNet

WordNet[8] je leksička baza za engleski jezik (Miller 1995). Radi se zapravo o javnom rječniku gdje su riječi organizirane u stablastu strukturu hiponima i hiperonima te grupirane u tzv. „synset“ grupe (grupe sinonima). WordNet je dizajniran za lak pristup pomoću računalnih programa i često se koristi u domeni obrade prirodnog jezika i računalnoj lingvistici.

Jedna riječ u WordNetu može pripadati jednoj ili više grupa sinonima. Riječi u jednoj grupi sinonima predstavljaju semantički jedan koncept, tj. radi se o sinonimima. Grupe sinonima (engl. *synset)* su povezane stablasto tvoreći hiperonime (općenitije koncepte) i hiponime (detaljnije koncepte). Na taj način možemo izračunati sličnost grupa sinonima – ovisno o njihovoj udaljenosti u stablastoj strukturi. Npr. grupa „reference_book“ (hrv. *knjiga uputa*) je hiperonim od „cookbook“ (hrv. *knjiga kuharica*), a „cookbook“ je samo jedan od mnogih hiponima od „reference_book“. S druge strane „reference_book“ je hiponim od „book“, „book“ je hiponim od „publication“ koji je hiponim od „work“ itd. sve dok ne dođemo do korijena stabla koji je u WordNetu grupa sinonima „entity“.

Budući da su riječi u WordNetu povezane relacijama „je“, „je dio od“, „sačinjen je od“, „je svojstvo od“ i sličnim relacijama, relativno je jednostavno utvrditi sličnost između riječi tj.

[8] https://wordnet.princeton.edu/

koncepata koje one predstavljaju. Što su koncepti bliži u hiperonimskom stablu to je sličnost između njih veća i obrnuto. Na taj način funkcionira *metrika puta* (engl. *path)*, koja traži najkraći put između dva koncepta u stablu tj. grafu:

$$Path(c1, c2) = path_length(c1, c2) \quad (25)$$

gdje su *c1* i *c2* koncepti (riječi), a *path_length* udaljenost u hiperonimskom stablu.

Leacock-Chodorow sličnost ili LCH (Leacock i Chodorow 1998) funkcionira na isti način kao i Path metrika s razlikom da koristi negativni logaritam:

$$LCH(c1, c2) = -log\ (path_length(c1, c2)) \quad (26)$$

Često se koristi i Wu-Palmerova sličnost (Wu i Palmer 1994) koja dodajte težinske faktore na udaljenosti između koncepata (veće udaljenosti dobivaju veći faktor i obrnuto).

4.5. Sažimanje teksta

Sažimanje teksta je proces generiranja sažetaka tj. kraćeg teksta koji opisuje duži tekst koji se sažima, tj. sadrži glavne koncepte iz teksta. Postoje dva pristupa sažimanju: ekstraktivni i apstraktivni pristup. Sažeci generirani metodom ekstrakcije sadrže rečenice ili dijelove rečenica iz teksta. Cilj je utvrditi značajne rečenice tj. one koje daju najviše informacija o temi teksta te se kao takve izvlače iz teksta i uključuju u sažetak. Abstrakcijski način je kompliciraniji način generiranja sažetaka koji se bazira na izvlačenju koncepata i tema iz teksta te generiranja sažetaka koji se ne sastoji nužno od rečenica iz teksta, tj. moguće je da sažetak sadrži riječi koje se ne spominju u tekstu.

Tehnike sažimanja teksta također možemo podijeliti i prema broju dokumenata koji se sažimaju: da li se sažima jedan dokument ili više njih. U kontekstu ovog istraživanja zanima nas sažimanje bazirana na jednom dokumentu gdje pod dokumentom razumijevamo mrežnu stranicu tj. tekst koji se nalazi na njoj.

Jedan od prvih algoritama sažimanja teksta su Luhnov (Luhn 1959) i Edmundsonov (Edmundson 1969) algoritam za generiranje sažetaka. Oba algoritma koriste slične ekstrakcijske metode i baziraju se na izračunu težinskih koeficijenata rečenica u tekstu na

temelju pozicija ključnih riječi. Ova jednostavna metoda obično daje bolje rezultate od apstraktivnih pristupa koji pate od semantičkih problema prirodnog jezika.

TextRank (Mihalcea i Tarau 2004) i LexRank (Erkan i Radev 2004) algoritmi baziraju se na graf reprezentaciji rečenica u dokumentu. U tom grafu rečenice su čvorovi, a bridovi su odnosi sličnosti između rečenica. Sličnost se izračunava pomoću modificirane TF-IDF metode što omogućava da kreiramo matricu sličnosti. Rečenice koje imaju više veza s visokim vrijednostima mjere sličnosti smatraju se više važnim te stoga i dobrim kandidatima za ulazak u sažetak. Ovakav način rangiranja rečenica sličan je PageRank algoritmu kojeg se koristi na Google-ovoj tražilici.

U literaturi se često koristi i tzv. osnovni algoritam (engl. *base line algorithm*) (Vanderwende, i dr. 2007) i KL-Sum algoritam (Haghighi i Vanderwende 2009). Vanderwendow osnovni algoritam ili „SumBasic“ algoritam počiva na ideji da se u sažetak trebaju izabrati rečenice koje sadrže riječi sa najvećom vjerojatnošću pojave u tekstu. Algoritam također kod odlučivanja uzima u obzir i rečenice koje je već stavio u sažetak na način da ažurira vjerojantosti riječima koje su dio sažetka i time povećava njihovu vjerojatnost da budu izabrane u sljedećim rečenicama. Time se postiže da manje frekventne riječi također mogu završiti u sažetku ako se nalaze u rečenicama zajedno sa visokofrekventnim riječima koje su prvotno izabrane.

KL-Sum algoritam također radi pomoću distribucija vjerojatnosti. Definira se kao pohlepna metoda koja dodaje rečenice u sažetak sve dok se KL divergencija spušta. KL divergencija (Kullback i Leibler 1951) je mjera za različitost dviju distribucija, a u ovom slučaju se radi o distribucijama vjerojatnosti u dokumentu i sažetku. Drugim riječima, rečenice će se dodavati u sažetak sve dok se distribucije vjerojatnosti riječi u dokumentu i sažetku ne poklope.

Latentno semantičko indeksiranje ili analiza (LSA) je također metoda (Deerwester, i dr. 1990; Gong i Liu 2001) koja se često koristi kod utvrđivanja teme tekstova. Ona polazi od stanovišta da riječi koje imaju slično značenje imaju tendenciju zajedničke pojave u tekstu. U prirodnom jeziku različite riječi mogu imati isto značenje, ili iste riječi različito značenje, ovisno o kontekstu. Kod sažimanja nas zapravo više zanima značenje iza riječi i usporedba značenja. LSA metoda konstruira riječ-dokument matricu i koristi SVD metodu (engl. *singular value decomposition*) za reduciranje veličine matrice bez značajnog gubitka informacija.

5. METODOLOGIJA ISTRAŽIVANJA

Istraživanje je podijeljeno u 6 faza. Prva faza je odabir uzorka mrežnih stranica nad kojima će se vršiti klasifikacija i daljnje istraživanje. Slijedi označavanje uzorka tj. ocjenjivanje stupnja prilagođenosti od strane tri različita SEO stručnjaka. Nakon toga slijedi priprema podataka za klasifikaciju i treniranje klasifikatora te evaluacija rezultata za potvrđivanje hipoteze H1. Slijedi izlučivanje začajnih faktora i izgradnja sustava za prilagodbu koji se i testira za potvrđivanje hipoteze H2.

5.1. Odabir uzorka

Skup podataka za ovo istraživanje formiran je slučajnim odabirom 600 mrežnih stranica iz direktorija DMOZ[9]. DMOZ se često koristi u istraživanjima iz domene pretraživanja informacija, najčešće kod validacije sustava za klasifikaciju mrežnih stranica (Aliakbary, i dr. 2009; Lee, Yeh i Chuang 2015; Marath, i dr. 2014), sažimanja teksta mrežnih stranica (Berger i Mittal 2000; Sun, i dr. 2005; Zhang, Zincir-Heywood i Milios 2004) ili ekstrakciju ključnih riječi (Mostafa 2013).

Ovo istraživanje koristit će samo mrežne stranice na engleskom jeziku budući da su takve najzastupljenije u DMOZ-u i budući da je leksička baza koja se koristi u kasnijim fazama najrazvijenija za engleski jezik. No, iste metode iz ovog istraživanja se mogu primijeniti i na stranice na nekim drugom jeziku za kojeg postoji leksička baza.

Direktorij DMOZ mrežne stranice svrstava u hijerarhijsku strukturu kategorija i potkategorija ovisno o temi stranice. Postoji 16 glavnih kategorija (engl.): „Arts“, „Business“, „Computers“, „Games“, „Health“, „Home“, „News“, „Recreation“, „Reference“, „Regional“, „Science“, „Shopping“, „Society“, „Sports“, „Kids & Teens Directory“ i „World“. Budući da će se naziv kategorije (i potkategorije) u ovom istraživanju koristiti kao ključna riječ, uzorak mrežnih stranica je uzet iz minimalno treće razine kao bi za svaku mrežnu stranicu imali najmanje tri ključne riječi, te kako bi izbjegli preopćenite ključne riječi koje bi mogle loše reprezentirati stranicu u kategoriji. Također, budući da uzorak za ovo istraživanje treba sadržavati samo stranice na engleskom jeziku, kategorije „Regional“ i „World“ nisu korištene

[9] http://ww.dmoz.org, https://dmoztools.net/, datum pristupa 10.3.2017.

za odabir uzorka. Kategorija „News“ je također izuzeta iz uzorkovanja budući da ona sadrži mrežne stranice koje se većinom bave vijestima u obliku web portala što otežava preciznije definiranje ključnih riječi. Uzorkovanje stranica vršeno je nad cijelim direktorijem DMOZ tj. nije određeno koliko će stranica biti uzeto iz koje kategorije, već je za svaku odabranu stranicu uzeta (zapisana) kategorija kojoj ona pripada. Ako je to bila kategorija „News“, „Regional“ ili „World“ ta stranica je preskočena tj. nije uzeta u uzorak. Konačni broj mrežnih stranica u uzorku po kategorijama prikazan je u tablici 5.1.

Tablica 5.1 Broj mrežnih stranica u uzorku po kategorijama DMOZ-a

Naziv kategorije	**Broj stranica**
Arts	63
Business	69
Computers	58
Games	35
Helath	59
Home	43
Recreation	37
Reference	31
Science	39
Shopping	59
Society	31
Sports	42
Kids & Teens	34
Ukupno:	**600**

Iz naziva kategorija (minimalno treće razine) formirane su ključne riječi za svaku mrežnu stranicu. U nekim slučajevima gdje se naziv kategorije sastoji od više riječi one su razdvojene, npr.:

Kids_and_Teens/Entertainment/Magazines_and_E-zines/

je pretvoreno u

Kids, Teens, Entertainment, Magazines, E-zines

Kategorije s nazivima kontinenata i država su skraćeni, npr.

Sports/Golf/Courses/North_America/United_States/Maryland/

je pretvoreno u

Sports, Golf, Courses, Maryland

Iz naziva kategorija izbačene su zaustavne riječi. Ako se stranica nalazila u više kategorija uzeta je samo prva iz treće ili više razine.

Rezultat ove faze istraživanja je skup podataka s 600 URL-ova i pripadajućih ključnih riječi[10].

5.2. Označavanje uzorka

Mrežne stranice iz uzorka potrebno je označiti tj. svrstati u tri kategorije: „neprilagođeno“, „djelomično prilagođeno“ i „prilagođeno“ s obzirom na stupanj prilagođenosti utvrđenim ključnim riječima iz naziva kategorije. Označavanje je potrebno radi treniranja klasifikatora – oznaka zapravo predstavlja klasni atribut. Navedeno označavanje vršila su tri neovisna SEO stručnjaka. Kod označavanja SEO stručnjaci su koristili vlastito iskustvo i subjektivnu procjenu da li je određena stranica prilagođena određenoj ključnoj riječi prema SEO pravilima. Informacije o stručnjacima i uputama koje su dobili prije ocjenjivanja date su u Prilogu 2, dok su ocjene vidljive u Prilogu 1.

Rezultat ove faze je skup podataka koji sadrži 3 atributa: URL mrežne stranice, ključne riječi i stupanj SEO prilagodbe. Distribucija oznaka nakon označavanja prikazana je u tablici 5.2.

Tablica 5.2 Distribucija oznaka u uzorku

Oznaka	**Stručnjak 1**	**Stručnjak 2**	**Stručnjak 3**	**MOD**
Neprilagođeno	180	119	112	146
Djelomično prilagođeno	307	341	341	293
Prilagođeno	113	140	147	161
Ukupno:	**600**	**600**	**600**	

Konačna klasa u koju je svrstana pojedina stranica je najčešća klasa od tri ocjene stručnjaka (stupac „MOD“ u Tablica 5.2). Najčešća klasa je postojala u svim ocjenjenim podacima, tj. nije bilo slučajeva da je jedna stranica ocjenjena s tri različite ocjene. Tako je dobivena vrijednost zavisne varijabla u ovom skupu podataka.

[10] Skup podataka je dan u Prilogu 1 – sadrži i vrijednosti varijabli i klasnu oznaku što je objašnjeno u nastavku teksta (poglavlja 5.2. i 5.3.).

Za izračunavanje stupnja međusobnog slaganja stručnjaka koristi se *kappa statistika* koja se pojavljuje u nekoliko oblika. Kada uzorke ocjenjuju dva stručnjaka koristi se *Cohenova kappa* (Cohen 1960), a kada se radi o više od dva ocjenjivača onda se koristi *Fleissova kappa* (Fleiss 1971). Također tip podataka diktira koju vrstu Kappa statistike treba upotrijebiti. Budući da u ovom istraživanju koristimo nominalnim (kategorijskim), odnosno ordinalnim podacima, prigodno je upotrijebiti Fleissovu Kappu. Za ordinalne podatke koristi se težinska kapa statistika (engl. *weighted kappa*), no nju je moguće izračunati samo između dva ocjenjivača (parovi).

Kappa statistika poprima vrijednosti u rasponu od -1 do 1. Vrijednost 1 označava savršeno slaganje ocjenjivača, vrijednost 0 označava da je slaganje isto kao i slučajni odabir, a vrijednost manja od 0 da je slaganje gore od slučajnog (što se rijetko događa).

Fleissova kappa statistika s tri ocjenjivača izračunata je nad ovim skupom podataka i prikazana je u tablici 5.3.

Tablica 5.3 Izračunata Fleissova Kappa statistika za ocjene tri stručnjaka nad 600 primjera uzorka

	Klasa „neprilagođeno“	**Klasa „djelomično prilagođeno“**	**Klasa „prilagođeno“**	**Prosječna vrijednost Kappa statistike**
Kappa	0,492362	0,340215	0,546786	0,44499570
Varijanca	0,003441	0,006293	0,003423	
Standardna pogreška	0,058662	0,079326	0,058503	
z vrijednost	8,393179	4,288804	9,346260	
p vrijednost	0,000000	0,000018	0,000000	

Iz tablice 5.3 možemo vidjeti da se stručnjaci najviše slažu oko klase „prilagođeno“ (Kappa=0,546786), nešto slabije kod klase „neprilagođeno“ (Kappa=0,492362), dok srednja klasa „djelomično prilagođeno“ ima najviše neslaganja, što je i razumljivo zbog graničnih vrijednosti koje SEO stručnjaci moraju uzeti u obzir kod ocjenjivanja (neka stranica koja je na rubu, tj. između najbolje ili najlošije i srednje klase češće je svrstana u srednju klasu zbog osjećaja da će tako najmanje pogriješiti kod ocjenjivanja). Bez obzira, ukupna Kappa statistika od 0,44499570 je dobar rezultat i govori nam da su ocjene pouzdane. Landis i Koch

(Landis i Koch 1977) daju tablicu s vrijednostima i objašnjenjima vrijednosti Kappa statistike (tablica 5.4).

Tablica 5.4 Značenje vrijednosti Kappa statistike (Izvor: Landis i Koch, 1977)

Vrijednost	**Značenje**
< 0	Loše slaganje
0.00 – 0.20	Slabo slaganje
0.21 – 0.40	Pošteno slaganje
0.41 – 0.60	Umjereno slaganje
0.61 – 0.80	Izrazito slaganje
0.81 – 1.00	Skoro savršeno slaganje

Međutim tablica 5.4 ne predstavlja standard za interpretaciju Kappa vrijednosti jer su autori navedene granične vrijednosti predložili po vlastitom nahođenju, bez konkretne potpore u podacima i istraživanju, što je i najčešća kritika skale na toj tablici.

Budući da je klasa ordinalna, izračunata je i težinska Kappa statistika prikazana u tablici 5.5.

Tablica 5.5 Težinska Kappa statistika za ocjene tri stručnjaka nad 600 primjera iz uzorka

	Stručnjak 2	**Stručnjak 3**
Stručnjak 1	0,637	0,564
Stručnjak 2	-	0,662

Iz tablice 5.5 vidljivo je da su se u ocjenjivanju stručnjaci međusobno podjednako slagali. Neznatno lošije slaganje je između stručnjaka 1 i 3.

Za analizu slaganja ocjenjivača kada se radi o ordinalnim varijablama često se koristi i tzv. *W* statistika tj. *Kendallov koeficijent slaganja* (engl. *Kendall coefficient of concordance*) (Kendall 1938). Njegova vrijednost je u rasponu 0-1 gdje vrijednost bliža 1 označava veće slaganje. Izračunata W statistika nad podacima u ovom istraživanju iznosi 0,611.

5.3. Izbor nezavisnih varijabli

Na temelju prethodnih istraživanja (Abdullah 2017; Andersson i Lindgren 2017; Giomelakis i Veglis 2016; Gupta, i dr. 2016; Hussien 2014; Mavridis i Symeonidis 2015; Zhu i Wu 2011;

Sujata, i dr. 2016; Zhang i Dimitroff 2005) formiran je popis nezavisnih varijabli koje će se koristiti u treniranju klasifikatora prikazan u tablici 5.6. Varijable su grupirane s obzirom na ulogu u HTML dokumentu.

Tablica 5.6 Nezavisne varijable korištene u istraživanju

Grupa	**Oznaka varijable**	**Objašnjenje**
Zaglavlje stranice	Tlen	dužina naslova iz HTML oznake „title“ (u broju riječi)
	Tkw	frekvencija ključnih riječi u naslovu „title“
	Mlen	dužina meta opisa (HTML oznaka "meta description")
	Mkw	frekvencija ključnih riječi u meta opisu
Naslovi	h1	broj pojavljivanja oznake H1
	h1len	prosječna dužina sadržaja oznake H1
	h1kw	frekvencija ključnih riječi u H1
	h2	broj pojavljivanja oznake H2
	h2len	prosječna dužina sadržaja oznake H2
	h2kw	frekvencija ključnih riječi u H2
	h3	broj pojavljivanja oznake H3
	h3len	prosječna dužina sadržaja oznake H3
	h3kw	frekvencija ključnih riječi u H3
Slike	alt	broj pojavljivanja ALT atributa u oznaci IMG (samo ako ALT sadrži neku vrijednost)
	altkw	frekvencija ključnih riječi u ALT atributu
Poveznice	linkkw	frekvencija ključnih riječi u tekstu poveznica (engl. *anchor text*)
	linkout	broj eksternih poveznica (engl. *outbound*)
URL	urllen	dužina URL-a (broj znakova)
	urlkw	frekvencija ključnih riječi u URL-u
Tekst	txtlen	dužina teksta na mrežnoj stranici (iz tijela)
	txtkw	frekvencija ključnih riječi u tekstu mrežne stranice

Zavisna varijabla je klasa tj. ocjena eksperta gdje mrežna stranica pripada. Moguće vrijednosti zavisne varijable su: 1-„neprilagođeno“, 2-„djelomično prilagođeno“ i 3-„prilagođeno“. Vrijednosti nezavisnih varijabli izvučene su automatski pomoću skripte napisane u Pythonu specifično za ovu svrhu. Kod izvlačenja frekvencija ključnih riječi bitno je napomenuti kako

su prije toga ključne riječi i riječi iz teksta svedene na korijensku formu (engl. *stemming*) korištenjem Porterovog algoritma za svođenje riječi na korijenski oblik..

Skup podataka koji je kreiran i korišten dalje u ovom istraživanju dan je u Prilogu 1. Sastoji se od 600 uzoraka opisanih s 21 nezavisne varijable i jednom zavisnom varijablom.

U tablici 5.7 prikazana je matrica korelacije između nezavisnih varijabli.

Tablica 5.7 Matrica Pearsonove korelacije nezavisnih varijabli

	Tlen	Tkw	Mlen	Mkw	h1	h1len	h1kw	h2	h2len	h2kw	h3	h3len	h3kw	alt	altKw	linkKw	inkOu	urlLen	urlKw	txtLer	txtKw
Tlen	1,00	0,34	0,21	0,18	0,05	0,03	0,10	-0,05	-0,05	0,01	-0,04	-0,01	0,05	0,05	0,07	0,05	0,04	-0,08	-0,01	-0,01	0,05
Tkw		1,00	0,18	0,55	0,07	0,05	0,45	0,02	0,01	0,24	0,01	0,03	0,28	0,07	0,36	0,42	0,05	0,07	0,39	-0,01	0,48
Mlen			1,00	0,50	-0,05	-0,04	0,07	0,01	-0,02	0,03	0,03	0,03	0,01	0,03	0,07	0,05	0,02	0,02	0,12	-0,04	0,02
Mkw				1,00	0,03	0,04	0,31	0,00	0,01	0,15	0,03	0,03	0,18	0,08	0,35	0,34	0,03	-0,02	0,24	-0,04	0,36
h1					1,00	0,67	0,35	0,03	0,09	0,03	0,04	0,00	0,02	-0,01	0,03	0,10	0,10	-0,02	-0,01	0,07	-0,01
h1len						1,00	0,34	0,06	0,11	0,05	0,08	0,04	0,00	-0,02	0,01	0,08	0,08	-0,03	-0,02	0,07	-0,03
h1kw							1,00	0,05	0,06	0,26	0,06	0,01	0,12	0,01	0,20	0,34	0,03	0,13	0,29	0,05	0,27
h2								1,00	0,78	0,44	0,37	0,24	0,07	0,30	0,08	0,11	0,27	-0,02	-0,02	0,34	-0,05
h2len									1,00	0,45	0,22	0,16	0,05	0,20	0,09	0,11	0,26	-0,04	-0,02	0,28	-0,03
h2kw										1,00	0,14	0,06	0,24	0,07	0,25	0,33	0,12	0,01	0,18	0,13	0,20
h3											1,00	0,63	0,22	0,26	-0,01	0,06	0,23	0,06	-0,01	0,21	-0,04
h3len												1,00	0,32	0,22	0,01	0,03	0,17	0,05	-0,01	0,25	-0,05
h3kw													1,00	0,06	0,23	0,32	0,12	0,03	0,13	0,06	0,23
alt														1,00	0,31	0,11	0,15	-0,03	0,03	0,18	0,09
altKw															1,00	0,45	0,08	0,01	0,27	0,01	0,40
linkKw																1,00	0,15	0,15	0,37	0,08	0,61
inkOut																	1,00	0,05	0,02	0,31	-0,02
urlLen																		1,00	0,42	0,12	0,07
urlKw																			1,00	0,03	0,29
txtLen																				1,00	0,00
txtKw																					1,00

U tablici 5.7 označene su korelacije koje prelaze vrijednost 0,50. Takve vrijednosti korelacija kod tih varijabli su i očekivane budući da se radi o povezanim varijablama, npr. broj ključnih riječi u Meta opisu (Mkw) ne može biti veći od nula ako je dužina Meta opisa nula (Mlen). Osim spomenutih nekoliko slučajeva, možemo primijetiti da je korelacija između većine varijabli mala ili umjerena.

Analiza glavnih komponenti (PCA) pokazala je kako bi tek 18 komponenti pokrilo 95% varijance podataka, a prvih 7 komponenti ima svojstvene vrijednosti veće od 1,0 i one bi objasnile 65% varijance. Analiza je prikazana u tablici 5.8.

Tablica 5.8 Analiza glavnih komponenti

Komponenta	Svojstvena vrijednost	Proporcija varijance	Kumulativ
1	3,92049	0,18669	0,18669
2	2,80444	0,13354	0,32023
3	1,85072	0,08813	0,40836
4	1,50494	0,07166	0,48003
5	1,45465	0,06927	0,5493
6	1,23531	0,05882	0,60812
7	1,06568	0,05075	0,65887
8	0,96933	0,04616	0,70503
9	0,84676	0,04032	0,74535
10	0,73178	0,03485	0,7802
11	0,6565	0,03126	0,81146
12	0,60421	0,02877	0,84023
13	0,52248	0,02488	0,86511
14	0,50634	0,02411	0,88922
15	0,4682	0,0223	0,91152
16	0,40323	0,0192	0,93072
17	0,3408	0,01623	0,94695
18	0,32746	0,01559	0,96254
19	0,31756	0,01512	0,97766
20	0,2811	0,01339	0,99105
21	0,18803	0,00895	1

Budući da je PCA analiza pokazala da nam je potrebno više glavnih komponenti za objašnjenje varijance (18 za 95% od ukupno 21), zaključujemo kako ne postoji dovoljno jak razlog za upotrebu glavnih komponenti za redukciju dimenzionalnosti.

5.4. Ekstrakcija faktora sadržaja

Analizom skupa podataka možemo saznati koje su varijable više značajne, a koje manje u predviđanju klase. Time možemo saznati koji faktori tj. varijable su bitne za prilagodbu sadržaja mrežne stranice. Najjednostavniji način je rangiranje varijabli s obzirom na

korelaciju s klasnom varijablom. U ovom slučaju zavisna varijabla je transformirana u numeričku (1="neprilagođeno", 2="djelomično prilagođeno", 3="prilagođeno"). U Tablica 5.9 dan je popis varijabli i vrijednost korelacije i drugih testova preporučenih za rangiranje varijabli kod klasifikacije (Ruiz Sánchez, i dr. 2005).

Upotrebom stabla odlučivanja moguće je i također dobiti uvid u prediktorska svojstva pojedine varijable. Mjere značajnosti koje se koriste u stablima odlučivanja mogu se koristiti za rangiranje varijabli, i to ne samo onih koje se pojavljuju u stablu, već svih, jer neka varijabla na nekom čvoru možda nije bila prvi izbor za taj čvor, nego drugi, što ne znači da je nevažna. Mjera važnosti varijable u stablima odlučivanja je informacijski dobitak koji se računa kao mjera redukcije entropije uzrokovanu razdvajanjem primjera na osnovu određene varijable[11]. Rangiranje varijabli prema informacijskom dobitku prikazano je u tablici 5.9, a grafički prikaz stabla odlučivanja na temelju kojeg se mogu vidjeti varijable bliže korijenu dan je u Prilogu 3.

Tablica 5.9 Važnost varijabli po različitim testovima

Varijabla	Korelacija	Informacijski dobitak	Hi kvadrat	Relief	Slučajna šuma
Tlen	0,257135	0,05544402	0,3264777	2,26E+04	**21,810107**
Tkw	**0,426339**	**0,12293193**	**0,4786585**	**6,65E+03**	**42,189959**
Mlen	**0,388327**	**0,20411246**	**0,5681143**	3,15E+04	**51,131839**
Mkw	**0,575413**	**0,23275999**	**0,6215979**	1,70E+04	**42,067928**
h1	0,022324	0,03913948	0,2757038	**6,64E+01**	10,321804
h1len	0,037721	0,03686936	0,2678983	-1,22E+02	11,734166
h1kw	**0,300638**	**0,06749659**	**0,3611576**	**7,46E+03**	15,475284
h2	0,135253	0,00000000	0,0000000	-8,69E+03	15,867193
h2len	0,099092	0,00000000	0,0000000	-7,56E+03	17,243045
h2kw	0,272581	0,05016715	0,3036099	1,96E+03	13,157693
h3	0,077233	0,00000000	0,0000000	4,78E+03	7,044129
h3len	0,057554	0,00000000	0,0000000	-4,67E+03	7,638001
h3kw	0,182965	0,04004664	0,2710277	-5,10E+03	4,511746
alt	0,123755	0,00000000	0,0000000	-3,28E+03	8,791717
altKw	0,244176	0,04367333	0,2905980	-2,45E+02	5,516950
linkKw	**0,277557**	0,05469602	0,3335990	2,80E+03	12,373683

[11] poglavlje 3.1.1

linkOut	0,082039	0,00000000	0,0000000	7,93E+01	13,193339
urlLen	0,013055	0,00000000	0,0000000	**6,23E+03**	1,648894
urlKw	0,210249	0,05061797	0,3141724	-7,25E+03	11,379830
txtLen	0,002467	0,03307188	0,2336873	1,23E+03	6,708839
txtKw	0,268072	**0,09057668**	**0,4262453**	**9,74E+03**	**31,804311**

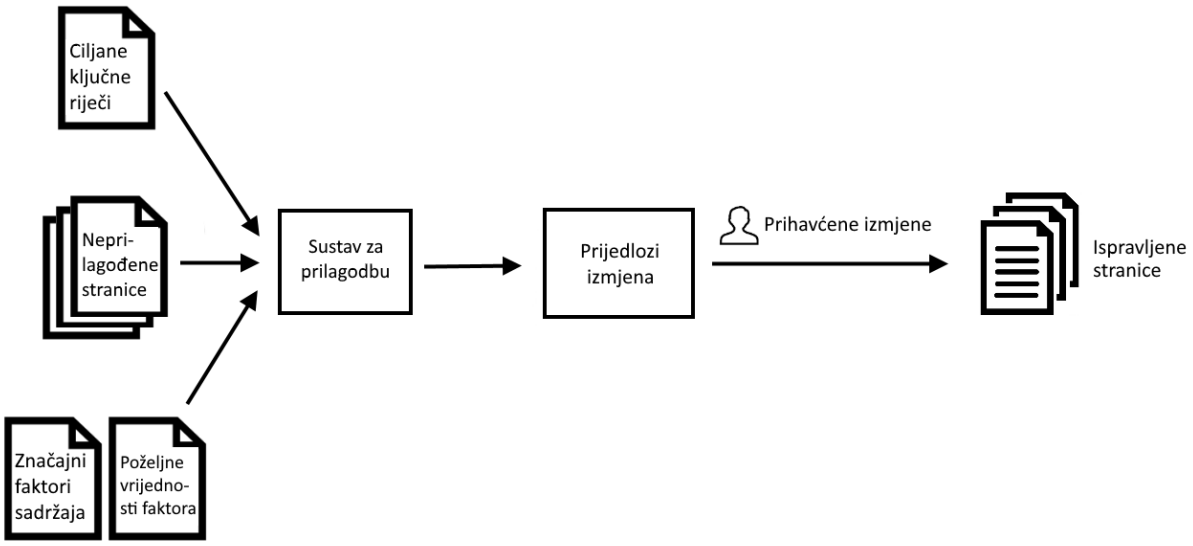
Varijable koje imaju najviše vrijednosti u većini testova iz tablice 5.9: Tkw (frekvencija ključnih riječi u oznaci „title“), Mlen (dužina oznake „meta description“), Mkw (frekvencija ključnih riječi u oznaci „meta description“), h1Kw (frekvencija ključnih riječi u oznaci „h1“) i txtKw (frekvencija ključnih riječi u tekstu stranice). Te varijable možemo smatrati značajnim faktorima kod prilagodbe sadržaja mrežnih stranica za internetske tražilice. One se spominju kao značajne i u većini prethodnih istraživanja (Zhang i Dimitroff, 2005; Zhu i Wu 2011; Su, i dr. 2014; Giomelakis i Veglis 2016; Buddenbrock 2016; Khan i Mahmood 2018).

5.5. Razvoj sustava za prilagodbu

Stranice koje klasifikator svrsta u grupu „neprilagođeno“ korigirat će se korištenjem algoritam za ispravak koji je kreiran za tu svrhu. Cilj algoritma je da korigirana mrežna stranica nakon ponovne klasifikacije bude klasificirana u grupu „djelomično prilagođeno“ ili „prilagođeno“, odnosno da se poveća stupanj prilagođenosti stranice. Evaluacija algoritma će se dakle vršiti pomoću modela klasifikatora izgrađenog u prethodnim fazama. Algoritam koristi leksičku bazu WordNet u kojoj su riječi organizirane u stablastu strukturu hipo i hiperonima i grupama sinonima. Navedena struktura se koristi za izračunavanje sličnosti između riječi iz teksta i ključnih riječi na temelju udaljenosti grupa sinonima (tzv. *synset grupa*) u hipo-hiper stablu. Riječi koje prelaze određeni prag sličnosti zamjenjuju se s ključnom riječi. Za ispravak „meta opisa“ tamo gdje on nedostaje koriste se metode sažimanja teksta, točnije sažimanje prema upitu (engl. *query based summarization*). Važno je naglasiti da je predloženi algoritam ispravka poluautomatski tj. algoritam predlaže promjenu za koju je potrebna suglasnost čovjeka kako se ne bi narušio smisao teksta. WordNet se često koristi u istraživanjima vezanim za kategorizaciju teksta, detekciju teme teksta te rješavanju problema smisla riječi i teksta (Cai, i dr. 2018; Wei i Chang, Measuring Word Semantic Relatedness Using WordNet-Based Approach. 2015; Wei, Lu, i dr. 2015).

Cilj ove faze je razviti sustav za poboljšavanje stupnja prilagođenosti onih mrežnih stranica koje su od strane eksperata označene kao „neprilagođene“ (najlošija klasa). Cilj sustava je što više stranica poboljšati na način da prijeđu u bolju klasu. Sustav će različitim algoritmima izvršiti promjene na stranici nakon čega će se ponovno izvući vrijednosti nezavisnih varijabli te testirati na modelu klasifikatora izgrađenom u prethodnim koracima.

Ulaz u sustav za prilagodbu su stranice koje su označene kao „neprilagođene“ (njihov HTML-kod), značajni faktori optimizacije sadržaja utvrđeni u prethodnim fazama (pomoću metoda strojnog učenja), te ciljane ključne riječi (Slika 5.1).



Slika 5.1 Shema sustava prilagodbe mrežnih stranica

Na izlazu iz sustava nalaze se prijedlozi za izmijenu mrežne stranice koji bi, ako bi bili implementirani, stranicu učinili boljom s obzirom na stupanj prilagodbe SEO pravilima na odabrane ključne riječi.

Osim samih značajnih faktora, u sustav prilagodbe trebaju ući i njihove poželjne vrijednosti kako bi sustav znao koje su ciljane vrijednosti. Za svaki od *n* značajnih faktora koji ulaze u sustav prilagodbe razvijeni su algoritmi i postupci pomoću kojih vršimo poboljšanje vrijednosti faktora. Zajedničko svojstvo za poboljšanje svih značajnih faktora je postupak obogaćivanja teksta ključnim riječima.

5.5.1. Obogaćivanje teksta ključnim riječima

Obogaćivanje teksta ključnim riječima možemo činiti na način da u tekst ubacujemo ključne riječi na određena mjesta bez da narušimo smisao teksta. Postoje tri situacije na koje možemo naići:

> SITUACIJA 1: Tekst ne sadrži niti jednu ključnu riječ, potrebno ih je ubaciti;
> SITUACIJA 2: Tekst sadrži ključne riječi, ali ne u dovoljnoj gustoći;
> SITUACIJA 3: Ne postoji tekst, potrebno ga je generirati na način da sadrži ključne riječi.

Za situacije 1 i 2 možemo koristiti dvije tehnike:

1) Zamjena sinonima, hiperonima ili sličnih riječi (upotrebom metrika sličnosti)
2) Ubacivanje ključnih riječi na mjesta gdje je to moguće utvrđivanjem n-grama (n=2) – ako se dvije riječi *a* i *b* pojavljuju često zajedno od kojih je riječ *b* ključna riječ, a izvorni tekst sadrži samo riječ *a*, sustav može dodati riječ *b*. Npr. ako je ključna riječ „hotel“, a n-gram je „hotel Sheraton“, ako se u tekstu hotel „Sheraton“ referencira samo sa „Sheraton“, može se dodati ključna riječ „hotel“ ispred kako bi poboljšali gustoću te ključne riječi u tekstu.

Za situaciju 3 je potrebno generirati tekst. U našem slučaju govorimo o generiranju sažetaka tj. kraćih verzija teksta iz većeg teksta gdje je veći tekst onaj koji se pojavljuje u tijelu mrežne stranice (tekst na stranici, čitljiv korisnicima). Ta tehnika se naziva sažimanje teksta koja može biti ekstraktivna ili apstraktivna. Ekstraktivno sažimanje utvrđuje značajne rečenice u tekstu i od njih generira sažetak – koristi dakle postojeće riječi i konstrukcije iz teksta. Abstrakcijska sažimanje generira sažetak koji se sastoji od riječi i rečenica koje se nužno ne moraju pojavljivati u izvornom tekstu. Ovo je teži način generiranja sažetaka pa je stoga ekstraktivni način i češći. Sažimanje bazirana na upitu (engl. *query based summarization*) se često koristi u području pretraživanja informacija.

Za generiranje sadržaja za oznaku „title“ (varijabla Tkw) potrebno je iz teksta tijela izlučiti 1-2 rečenice koje najbolje opisuju tekst i po mogućnosti da sadrže ključne riječi. Kod generiranja sadržaja za oznaku „meta description“ (varijabla Mkw) u obzir se može uzeti malo duža sažimanje (od nekoliko rečenica). Za generiranje naslova (oznake h1 i h2) biti će potrebno iz teksta izlučiti konstrukcije (rečenice ili dijelove rečenica) koje bi bile

najpogodnije za naslov. Nažalost, ako mrežna stranica nema tekst u tijelu, onda cijeli ovaj postupak nije moguć. To je ujedno i ograničenje ovog sustava.

5.5.2. Algoritam sažimanja teksta

Testiranje algoritama za sažimanje teksta za potrebe SEO prilagodbe „meta opisa“ vršeno je u prethodnom istraživanju (Matošević, 2018) nad uzorkom od 100 stranica iz DMOZ-a koji nisu imali „meta opis“. Korišteno je pet popularnih algoritama za sažimanje i poseban algoritam razvijen za tu svrhu kojeg navodimo u nastavku:

Algoritam 1: Predloženi algoritam sažimanja baziran na upitu

Ulaz: K=lista ciljanih ključnih riječi, T=tekst za sažimanje, N=broj rečenica, S=granična vrijednost sličnosti riječi

Izlaz: SUM=generirani sažetak

```
T=izbaci zaustavne riječi iz T
K=STEM(K)
Za svaku riječ u T
    tag=POS_TAG(riječ)
    ako je tag=imenica onda
        riječ=STEM(riječ)
        riječ_s=WordNet_Synset(riječ)
        Za svaku ključnu riječ u K
            sim=sličnost(riječ ili riječ_s, ključna riječ)
            ako je sim>S onda zamijeni riječ s ključnom riječi u T
Za svaku rečenicu u T
    Ako rečenica sadrži ključnu riječ iz K onda rang(rečenice)+=1
Vrati SUM=prvih N rečenica poredanih po rang(rečenice) u padajućem redosljedu
```

Predloženi Algoritam 1 je jednostavan – broji ključne riječi u rečenicama i rangira ih prema tom broju kako bi u sažetak odabrao prvih N rečenica po rangu. Algoritam je poboljšan upitom na leksičku bazu WordNet kako bi pronašao sinonime i identificirao potencijalne riječi za zamjenu i tako povećao rang rečenice i njenu priliku da uđe u prvih N kandidata za sažetak. Algoritam koristi i POS označavanje kako bi pretragu suzio samo na imenice (budući

da su one najčešće ključne riječi), te STEM funkciju koja služi za svođenje riječi na njen korijen (kako bi usporedbe riječi bile pouzdanije tj. kako bi eliminirali gramatičke sufikse). Za korjenovanje riječi upotrijebljen je popularni Porterov korijenovatelj. Funkcija sličnosti izračunava sličnost između riječi i bazira se na udaljenosti riječi u hiper-hipo stablu riječi u WordNetu. U ovom eksperimentu odabrana je granična sličnost S=0,8 i broj rečenica od N=5, kao ulazne vrijednosti u algoritam.

Algoritam 1 je testiran na način da je izvršen nad svih 100 stranica iz uzorka i uspoređen s ostalim poznatim algoritmima za sažimanje. Evaluacija je vršena od strane tri neovisna SEO stručnjaka koji su za zadatak imali odabrati najpovoljniji sažetak za meta opis (prema SEO preporukama, koristeći vlastito znanje i iskustvo) od predloženih 7 (Algoritam 1, LexRank, LSA, Luhn, KL, TextRank i Basic). Tablica 5.10 prikazuje primjer jednog zapisa (generiranog sažetka).

Tablica 5.10 Primjer jednog zapisa iz rezultata sažimanja teksta mrežne stranice

Ključne riječi	Recreation, Pets, Dogs
Naslov stranice (title):	Dog owner's guide: welcome to dog owner's guide
Algoritam 1:	You have a dog, want a dog, or just plain. Like dogs, you've come to the right place for all kinds of information. About living with and loving dogs. Along with dozens of articles about dogs, we also have a bookstore.
LexRank:	If an article fits in more than one topic it is listed. See our list of topics; choose of all topics and articles. Books for help choosing the right dog.
LSA:	Chock full of books for further reading and videos for family enjoyment and a mall for doggie purchases, than 300 pages of features, breed profiles, training tips, health information, how to teach that puppy good manners, how to choose a veterinarian, a boarding. Kennel, a groomer, or a trainer, what to expect at a dog show, and much, and be sure to check these special pages: gift books and movies for the whole family.
Luhn:	You'll find articles to help you choose a breed, a breeder and a puppy; the right dog, manners and training, dog owner's. See our list of topics; choose. See our list of related books, amazon's lists of best-selling dog books, or browse for anything in the entire amazon catalog.
KL:	Chock full of books for further reading and videos for family enjoyment and a mall for doggie purchases, than 300 pages of features, breed profiles, training tips, health information. Looking for information on one subject? Be sure to check these special pages: gift books and movies for the whole family. Books for new owners.
TextRank:	You have a dog, want a dog, or just plain. If an article fits in more than one topic it is listed. See a list of all articles listed by title. See our list of related books, amazon's lists of best-selling dog books, or browse for anything in the entire amazon catalog.

Basic:	Dog. See a list. All topics and articles. Books for new owners.

Iz tablice 5.10 možemo primijetiti kako neki algoritmi imaju problema s duljinom rečenice. Odabir prekratkih rečenica generira loše sažetke, naročito sa SEO gledišta (algoritam Basic). S druge pak strane, ako algoritam preferira duže rečenice dobijemo predugačak sažetak koji također nije pogodan za „meta opis“ sa SEO gledišta (npr. LSA algoritam). Rješenje bi bilo dodati ograničenje u dužini rečenica, ali kako ne bi izgubili relevantne dugačke rečenice potreban bi bio algoritam dijeljenja rečenica na dvije ili više. To može biti tema za buduća istraživanja u ovom području.

Rezultati evaluacije eksperata prikazani su u tablici 5.11. Oni su odabrali predloženi Algoritam 1 u 95 slučajeva (ukupno), što od mogućih 300 slučajeva čini nešto manje od jedne trećine. To je vrlo ohrabrujući i zadovoljavajući rezultat. Na drugom mjestu je LexRank s 68 slučajeva, a na trećem LSA s 46 slučaja. TextRank i Basic algoritmi pokazali su se kao najlošiji, a u 20 slučaja niti jedan sažetak nije bio odgovarajući kandidat za „meta opis“ s obzirom na dane ključne riječi.

Tablica 5.11 Ocjena eksperata o podobnosti sažetaka za „meta opis“ (prema Matošević, 2018)

Algoritam	Expert 1	Expert 2	Expert 3	Ukupno
Algoritam 1	24	39	32	95
LexRank	21	34	13	68
LSA	19	12	15	46
Luhn	11	10	9	30
KL	5	11	9	25
TextRank	10	7	5	22
Niti jedan	5	7	8	20
Basic	4	4	7	15

Mali broj uzorka s ocjenom „niti jedan“ govori nam da u većini slučajeva jedan od algoritama sažimanja može biti upotrijebljen za generiranje „meta opisa“, što znači da samom primjenom tih algoritama možemo poboljšati prilagođenost mrežne stranice SEO preporukama.

5.6. Analiza predloženih promjena

Stranice koje ispravi algoritam za ispravak analizirati će SEO stručnjaci na način da prihvate ili odbace određenu izmjenu. Na temelju toga izvršit će se statistička analiza korisnosti predloženih promjena, te odredit elementi algoritma koji su više, a koji manje uspješni u procesu ispravka.

Prihvaćanje ili odbijanje promjena od strane čovjeka (SEO stručnjaka) nužna je zbog kontrole smisla teksta. Ubacivanjem ključnih riječi i njihovom zamjenom pomoću predloženih algoritama i pravila može generirati nesmislene rečenice zbog čega nam je potrebna ljudska evaluacija.

Mrežne stranice iz skupa podataka koji su predmet ovog istraživanja, koje su bile označene najlošijom klasom („neprilagođene“) prolaze dakle kroz algoritam ispravka, te nakon toga prolaze kroz proces ljudske evaluacije tj. prihvaćanja ili odbijanja predloženih promjena. Promjene koje su odbijene ne ulaze u rezultat ispravka. Rezultat ispravka čine dakle izmijenjene stranice koje su odobrili SEO stručnjaci i kao takve čine ulazni skup u završnu analizu.

Završna analiza se sastoji od ponovne klasifikacije ispravljenih stranica korištenjem algoritama strojnog učenja da bi utvrdili njihovu novu klasu („prilagođeno“, „djelomično prilagođeno“ ili „neprilagođeno“). Kako bismo potvrdili hipotezu H2 potrebno je da više od polovica tih stranica prijeđe u bolju klasu nakon primjene algoritma ispravka nad njima.

6. REZULTATI ISTRAŽIVANJA

Ekstrakcijom vrijednosti značajnih varijabli formiran je skup podataka od 600 uzoraka i 21 nezavisnom varijablom, te 1 zavisnom varijablom koja predstavlja klasu u koju uzorak pripada (Prilog 1). Klasnu pripadnost su odredili SEO stručnjaci u procesu ocjenjivanja te smo time dobili skup podataka s određenom klasnom pripadnošću (odabrana je većinska klasa temeljem procjene tri SEO stručnjaka).

Navedeni skup podataka predstavlja ulaz u algoritme klasifikacije pomoću kojih se trenira klasifikator za označavanje mrežnih stranica. Evaluacija klasifikacije je vršena pomoću metode izdvajanja (engl. *holdout*) i pomoću unakrsne validacije. U metodi izdvajanja koristi se 2/3 uzorka za učenje, a 1/3 za validaciju. Za unakrsnu validaciju korištena je ugniježđena 10-struka unakrsna validacija (sa 10 iteracija u vanjskoj petlji i 10 iteracija u unutarnjoj)[12].

U nastavku predstavljamo rezultate klasifikacije i ostalih pokusa koji su provedeni u ovom istraživanju.

6.1. Rezultati klasifikacije mrežnih stranica

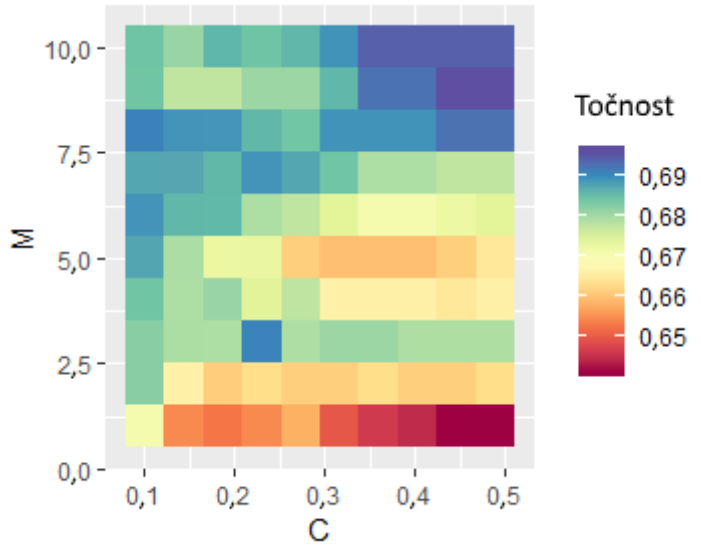
Za treniranje klasifikatora korišten je programski alat R i paket MLR[13]. Treniranje je vršeno nad svih 600 uzoraka s ugniježđenom unakrsnom validacijom pomoću pet najznačajnija klasifikatora: stabla odlučivanja, SVM, Naive Bayes, kNN i logistička regresija (Wu, i dr. 2008), te točnosti kao mjere evaluacije. Kod svih klasifikatora istražila se prikladna vrijednost hiperparametara za potrebe regularizacije pomoću mrežne metode (engl. g*rid search*). Podaci su normalizirani min-max metodom. Kod unakrsne validacije korištena je metoda stratifikacije.

6.1.1. Stabla odlučivanja

Za izgradnju stabla odlučivanja korišten je algoritam J48 koji predstavlja java implementaciju algoritma C4.5. Testirani su parametri pouzdanosti (engl. *confidence factor*, kontrolira

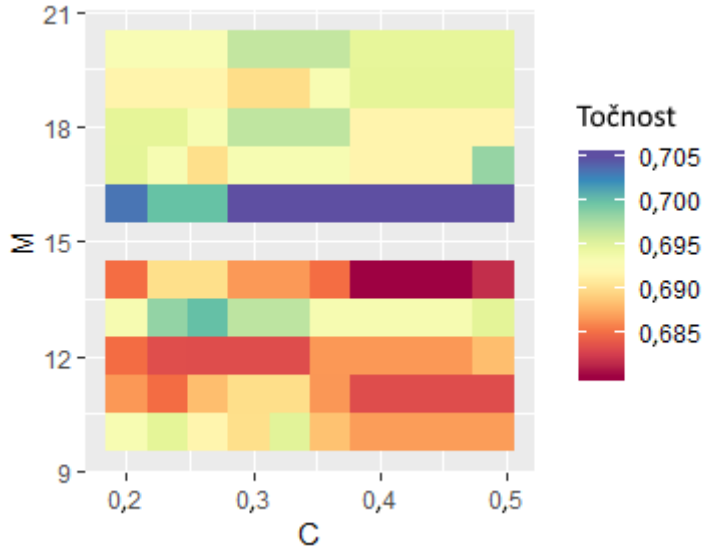
[12] proces ugniježđene unakrsne validacije objašnjen je na str. 33.
[13] https://github.com/mlr-org/mlr, datum pristupa 25.10.2018.

skraćivanje stabla) i minimalni broj uzorka po listu (oznaka. „minNumObj“). Parametri su ispitani u rasponu od 0,1 do 0,49 za pouzdanost (oznaka C) i od 1 do 10 za broj primjera po listu (M). Rezultati mrežne pretrage prikazani su na slikama 6.1 i 6.2.



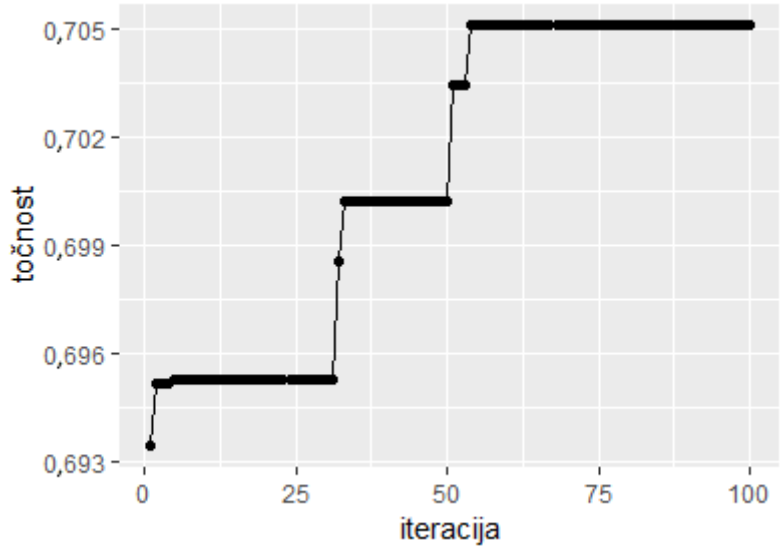
Slika 6.1 Rezultat optimizacije hiperparametra C i M kod algoritma stabla odlučivanja pomoću metode mrežne pretrage

Iz slike 6.1 zaključujemo da je područje u kojem možemo očekivati dobre rezultate kada je M>6 i C>0,2, stoga je to područje dodatno ispitano (u rasponu 6<M<20 i 0,2<C<0,5) te je rezultat prikazan na slici 6.2.



Slika 6.2 Rezultat detaljnije optimizacije parametra C i M kod algoritma stabla odlučivanja pomoću metode mrežne pretrage

Ovime smo utvrdili optimalne vrijednosti hiperparametara C=0,49 i M=16. Isti postupak je proveden u testiranju hiperparametara u narednim poglavljima: najprije se uzeo širi raspon vrijednosti da bi se utvrdila potencijalna područja, te nakon toga se odabrano područje detaljnije ispitalo. Na slici 6.3 prikazujemo odnos broja iteracija i točnosti tj. koliko je iteracija bilo potrebno da se dođe do optimalnih vrijednosti hiperparametera.



Slika 6.3 Iteracije i točnost kod mrežne pretrage hiperparametara stabla odlučivanja

Točnost algoritma J48 s pronađenim vrijednostima hiperparametara s obzirom na dva načina validacije prikazan je u tablici 6.1.

Tablica 6.1 Točnost algoritma J48 s različitim načinima validacije

Način validacije	**Točnost**
10-struka unakrsna	67,53%
66% treniranje, 33% test	65,67%

Matrica zabune za model s 10-strukom unakrsnom validacijom je:

```
        predviđeno
stvarno   1   2   3 -pogreška-
  1      92  51   3     54
  2      52 199  42     94
  3       1  46 114     47
 -pogr.- 53  97  45    195
```

Iz matrice zabune vidljivo je da je najviše grešaka (94) klasifikator učinio između klasa 1 (neprilagođeno) i 2 (djelomično prilagođeno), njih 52, te između klasa 2 (djelomično

prilagođeno) i 3 (prilagođeno), njih 42. Najmanje pogreška, samo 4, klasifikator je učinio između klasa 1 i 3 tj. „neprilagođeno“ i „prilagođeno“.

Rezultat je stablo veličine 133 grana i 97 listova (Prilog br. 3). Iz izgrađenog stabla vidljive su najznačajnije varijable (u vrhovima stabla): Mkw (ključne riječi u „meta opisu“), TKw (ključne riječi u oznaci „title“), h1kw (ključne riječi u oznaci „h1“) i Txtkw (ključne riječi u tekstu stranice); što je u skladu s ranijim rezultatima utvrđivanja značajnih faktora (Abdullah 2017; Andersson i Lindgren 2017; Gupta, i dr. 2016; Zhang i Dimitroff 2005) te u skladu s preporukama tražilica (Google Webmaster Guidelines n.d.; Bing Webmaster Guidelines n.d.).

6.1.2. Naivan Bayesov klasifikator

Budući da su nezavisne varijable u skupu podataka numeričke, a Naivan Bayesov klasifikator preferira nominalne vrijednosti, provedena je diskretizacija varijabli prije izvođenja klasifikatora. Rezultati su prikazani u tablici 6.2.

Tablica 6.2 Točnost naivnog Bayesovog klasifikatora

Način validacije	**Točnost**
10-struka unakrsna	54,69%
66% treniranje, 33% test	58,71%

Matrica zabune s 10-strukom unakrsnom validacijom je:

```
        predviđena
stvarna    1   2  3 -pogreška-
  1      119  23  4     27
  2      107 127 59    166
  3        9  70 82     79
 -pogr.- 116  93 63    272
```

Iz matrice zabune je vidljivo da i ovaj klasifikator kao i stabla odlučivanja najviše griješi između klasa 1 i 2 („neprilagođeno“ i „djelomično prilagođeno“), te 2 i 3 („djelomično prilagođeno“ i „prilagođeno“), gdje je napravio ukupno 166 pogrešnih klasifikacija. Broj grešaka između najgore i najbolje klase („neprilagođeno“ i „prilagođeno“) je 13 (9+4 iz matrice zabune), što je više od klasifikatora stabla odlučivanja. Znatno je veći i ukupan broj pogrešaka (272 naprema 195).

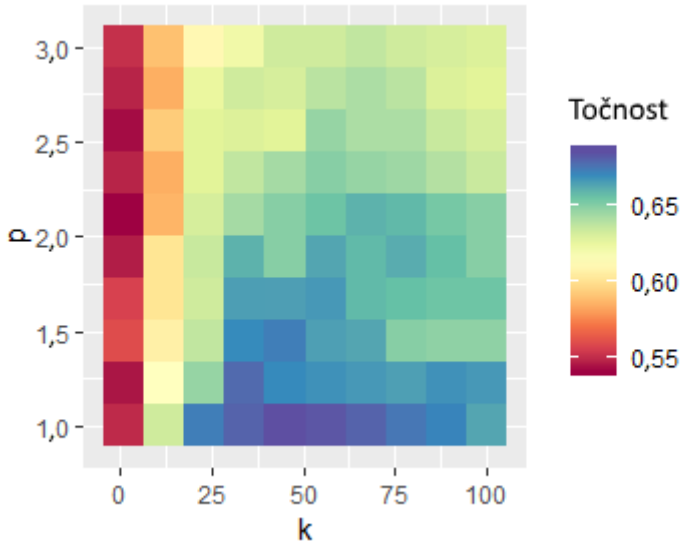
Naivan Bayesov klasifikator nema hiperparametara koje bi mogli optimizirati. Rezultati točnosti pokazuju da je to najlošiji klasifikator za ovaj skup podataka.

6.1.3. KNN klasifikator

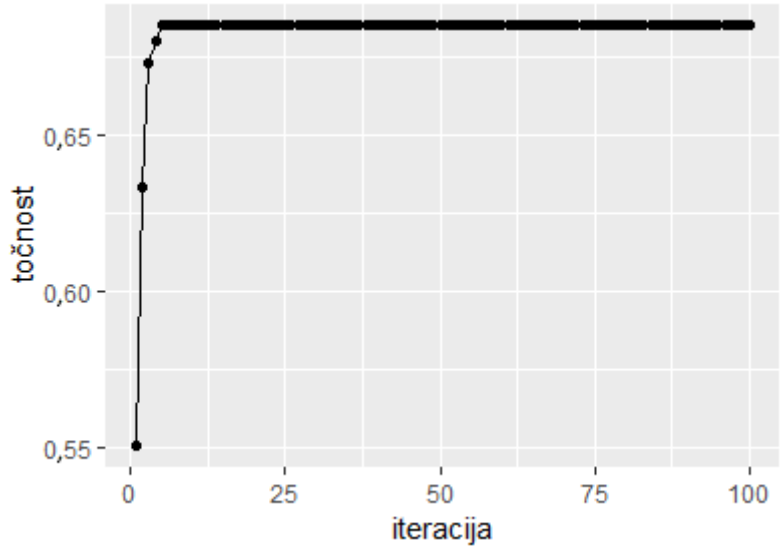
U potrazi za optimalnim rezultatima kNN klasifikatora testirani su hiperparametri *k* (broj susjeda) i mjera udaljenosti, odnosno parametar *p* Minkowskijeve udaljenosti (Teknomo n.d.). Varirajući parametar p iz formule Minkowskijeve udaljenosti (17) može se doći do Euklidske (kada je p=2) i Manhattan (kada je p=1).

$$Minkowski(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \tag{17}$$

Mrežna pretraga dala je rezultate prikazane na slici 6.4 i 6.5.



Slika 6.4 Rezultat optimizacije hiperparametra k i p kod KNN algoritma pomoću metode mrežne pretrage



Slika 6.5 Iteracije i točnost kod mrežne pretrage hiperparametara algoritma KNN

Optimalni hiperparametri su pronađeni vrlo brzo (nakon 5. iteracije), a to su k=45 i p=1, što znači upotrebu Manhattan udaljenosti. Rezultati validacije modela s tim hiperparametrima prikazani su u tablici 6.3.

Tablica 6.3 Točnost KNN klasifikatora s k=45 i p=1

Način validacije	**Točnost**
10-struka unakrsna	69,67%
66% treniranje, 33% test	65,17%

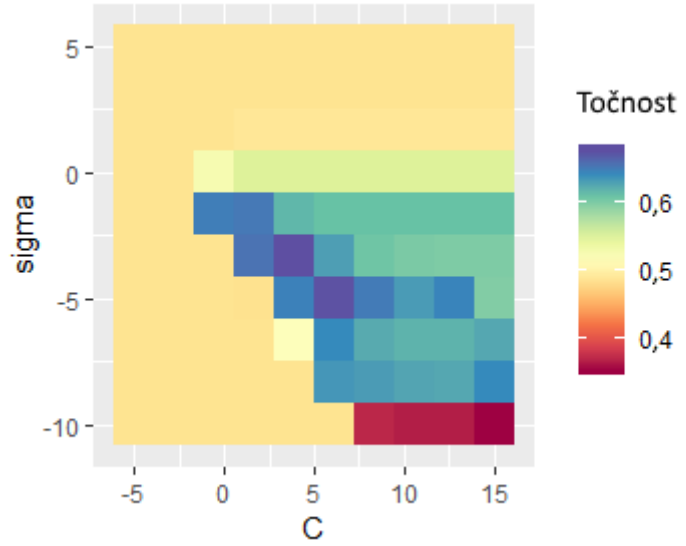
Matrica zabune 10-struke unakrsne validacije je:

```
        predviđena
stvarna    1   2  3 -pogreška-
  1      123  22  1     23
  2       65 197 31     96
  3        4  59 98     63
 -pogr.-  69  81 32    182
```

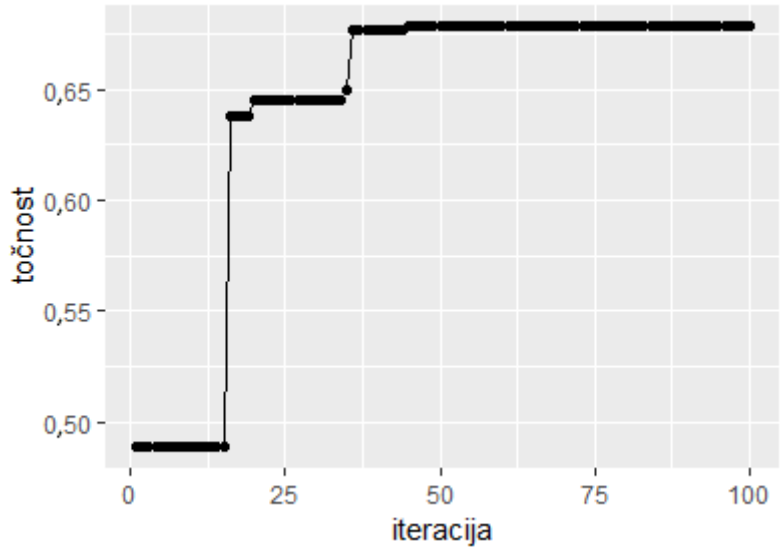
Iz matrice zabune vidljivo je da je kao i u prethodno testiranim klasifikatorima najviše pogrešaka između klasa 1 i 2 te 2 i 3 (ukupno 96). Pogrešaka između najlošije i najbolje klase (1 i 3) je ukupno 5, što predstavlja dobar rezultat – ne bi bilo poželjno da je ovaj broj velik jer bi to značilo da klasifikator „prilagođene“ stranice svrstava u „neprilagođene“ i obrnuto.

6.1.4. Metoda potpornih vektora

SVM algoritam uključuje parametar *C* koji označava kompleksnost modela. Pomoću programskog paketa MLR u alatu R istražene su optimalne vrijednosti parametra te se opetovanim postupkom utvrdilo da se one nalaze: *C* u rasponu od 10^{-5} do 10^{10} , te σ RBF jezgrene funkcije u rasponu 10^{-10} i 10^{5}. Metoda mrežne pretrage dala je rezultate prikazane na slikama 6.6 i 6.7.



Slika 6.6 Rezultat optimizacije hiperparametra c i σ (sigma) kod SVM algoritma pomoću metode mrežne pretrage



Slika 6.7 Iteracije i točnost kod mrežne pretrage hiperparametara algoritma SVM

Optimalni hiperparametri su *C*=7.74e+03 i σ=0.000464 što je dalo točnost prikazano u tablici 6.4.

Tablica 6.4 Točnost SVM klasifikatora s c=7.74e+03 i σ=0.000464

Način validacije	**Točnost**
10-struka unakrsna	66,18 %
66% treniranje, 33% test	62,68%

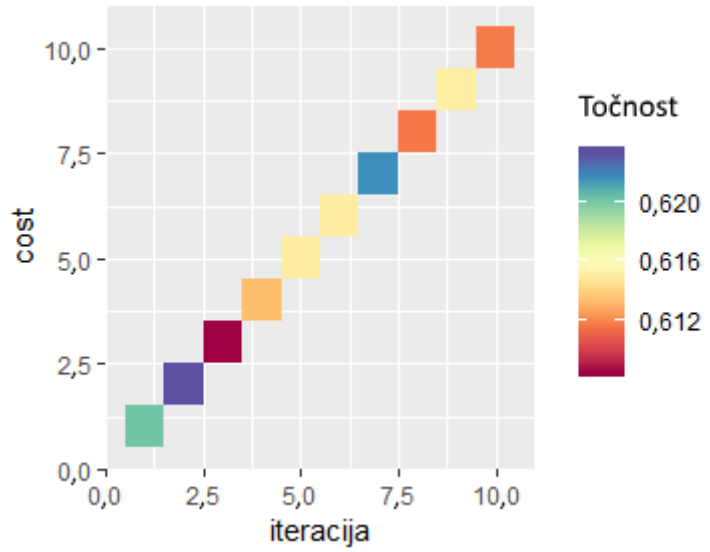
Matrica zabune 10-struke unakrsne validacije je:

```
        predviđena
stvarna   1   2   3 -pogreška-
  1      95  50   1     51
  2      51 200  42     93
  3       4  55 102     59
 -pog.-  55 105  43    203
```

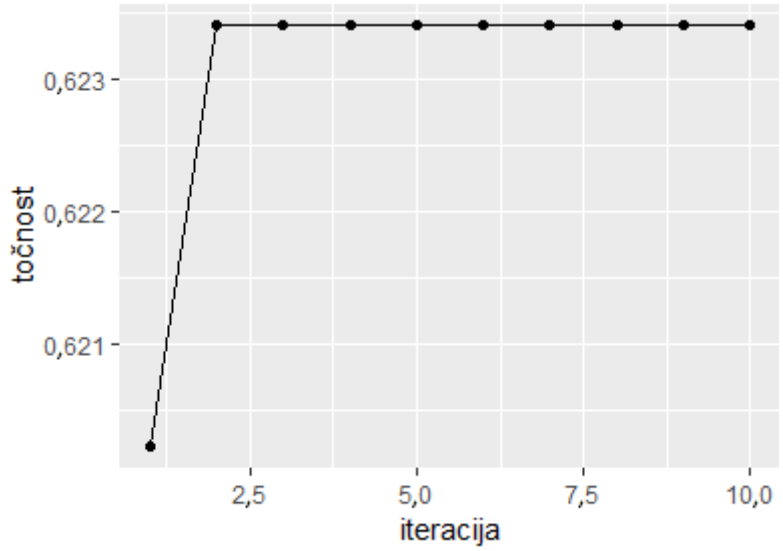
Iz matrice zabune vidimo slične karakteristike kao i kod prethodnih klasifikatora: najviše se griješi između klasa 1 i 2 te 2 i 3. Broj grešaka između klasa 1 i 3 je na niskoj razini od 5.

6.1.5. Logistička regresija

Kod logističke regresije pretreniranost se kontrolira s L2 „ridge“ parametrom („cost“) koji predstavlja penalizaciju pomoću težinskih faktora. Ispitan je raspon od 1 do 10, a metoda mrežne pretrage utvrdila je da je optimalna vrijednost 2 (slike 6.8 i 6.9).



Slika 6.8 Rezultat optimizacije hiperparametra „cost“ kod logističke regresije pomoću metode mrežne pretrage



Slika 6.9 Iteracije i točnost kod mrežne pretrage hiperparametara "cost" logističke regresije

Model s hiperparametrom „cost“=2 daje rezultate prikazane u tablici 6.5.

Tablica 6.5 Točnost klasifikacije pomoću logističke regresije s parametrom "cost"=2

Način validacije	**Točnost**
10-struka unakrsna	62,99%
66% treniranje, 33% test	62,19%

Matrica zabune 10-struke unakrsne validacije je:

```
        predviđena
stvarna   1   2  3 -pogreška-
  1      73  69  4     73
  2      28 219 46     74
  3       2  73 86     75
 -pogr.- 30 142 50    222
```

U ovoj matrici zabune možemo primjetiti podjednaki broj pogrešaka po klasama (73, 74 i 75). Ukupan broj pogrešaka je 222 što je nešto više od prethodno testiranih (osim Bayesovog klasifikatora). Ukupan broj pogrešaka između klase 1 i 3 („neprilagođeno“ i „prilagođeno“) je na niskoj razini, ukupno 6.

Rezultati signifikantnosti modela logističke regresije dani su u tablici 6.6.

Tablica 6.6 Signifikantnost varijabli kod modela logističke regresije

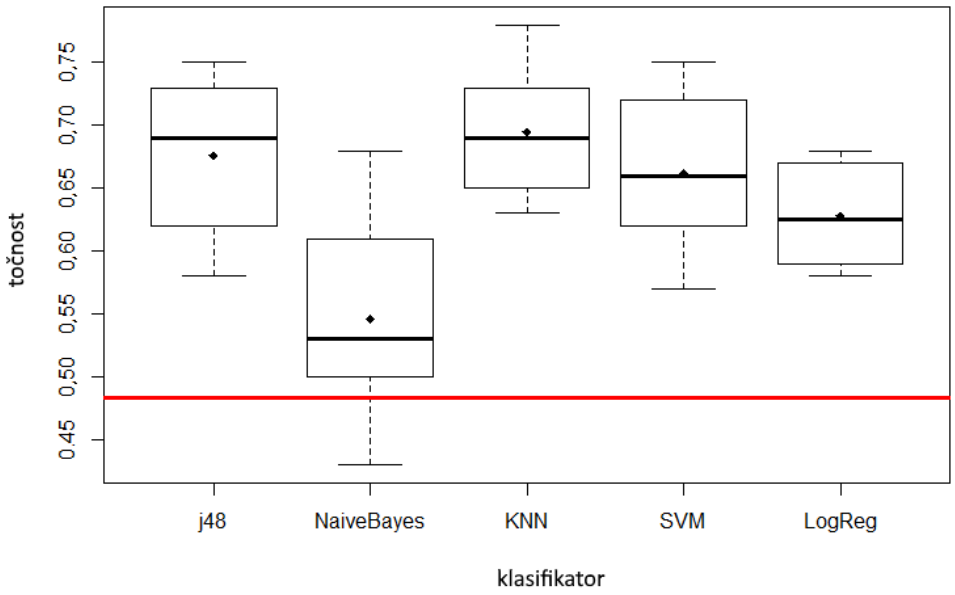
	Procjena koeficijenta	**Standardna pogreška**	**t vrijednost**	**Pr(>|t|)**
Slobodni koef.	1.411e+00	7.967e-02	17.707	< 2e-16 ***
Mlen	4.961e-03	1.348e-03	3.680	0.000255 ***
Tlen	1.915e-02	5.064e-03	3.782	0.000172 ***
Tkw	2.415e-02	3.051e-02	0.792	0.428875
Mkw	2.170e-01	2.528e-02	8.583	< 2e-16 ***
h1	-1.086e-02	1.274e-02	-0.852	0.394530
h1len	5.191e-04	2.095e-03	0.248	0.804351
h1kw	8.108e-02	3.194e-02	2.538	0.011400 *
h2	1.141e-02	5.682e-03	2.008	0.045115 *
h2len	-1.067e-03	1.061e-03	-1.005	0.315237
h2kw	1.014e-01	2.995e-02	3.384	0.000762 ***
h3	-7.203e-04	3.147e-03	-0.229	0.819040
h3len	6.141e-05	7.797e-04	0.079	0.937249
h3kw	3.088e-02	2.969e-02	1.040	0.298751
alt	2.278e-03	1.600e-03	1.423	0.155201
altKw	-1.273e-02	2.270e-02	-0.561	0.575264
linkKw	-3.793e-03	1.709e-02	-0.222	0.824379
linkOut	3.020e-04	3.463e-04	0.872	0.383580
urlLen	7.593e-04	1.805e-03	0.421	0.674149
txtLen	-2.346e-05	2.222e-05	-1.056	0.291547

txtKw	1.765e-02	1.570e-02	1.124	0.261385
Oznake značajnosti: 0 '***', 0,001 '**', 0,01 '*'				
Koeficijent determinacije: 0,4259 , Korigirani koeficijent determinacije: 0,4061				
F-statistika: 21,48 s 20 i 579 stupnjeva slobode, p-vrijednost: < 2,2e-16				

Iz tablice 6.6 vidljive su značajne varijable (označene zvjezdicama) te signifikantnost modela ($p<2,2e-16$). Koeficijent determinacije od 0,4259 upućuje na umjerenu reprezentativnost modela.

6.2. Usporedba rezultata klasifikacije

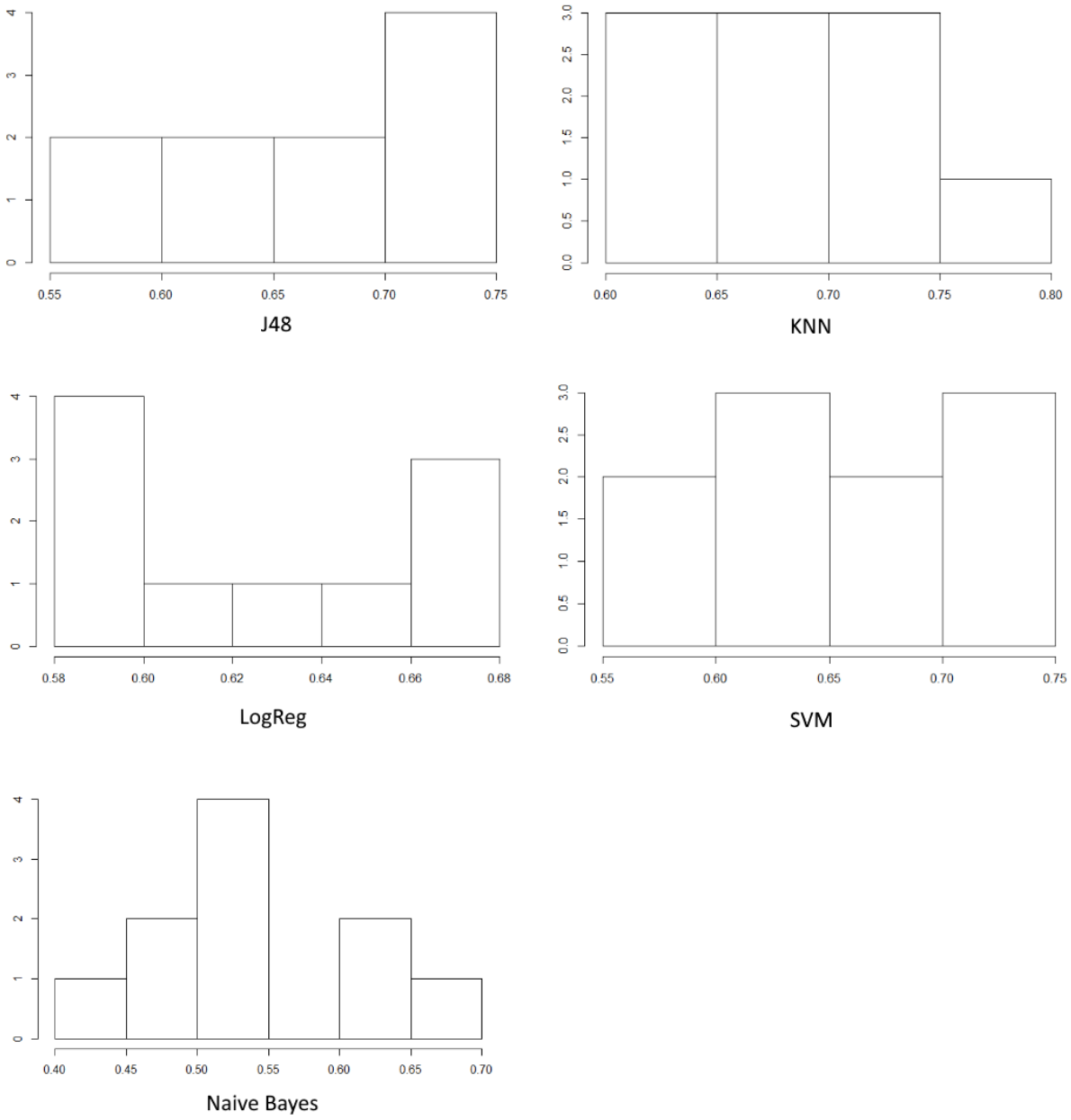
Za potvrđivanje hipoteze H1 uspoređeni su svi obrađeni klasifikatori tj. njihova točnost s postotkom najčešće klase (u daljnjem tekstu „bazna točnost“). Najčešća klasa je klasa 2, tj. „djelomično prilagođeno“ (293 uzoraka od ukupno 600), što čini 48,83% skupa. Kod 10-struke unakrsne validacije prosječna točnost (od 10 prolaza) svih testiranih klasifikatora je iznad te granice, što je prikazano i dijagramom pravokutnika na slici 6.10. Crvena linija na toj slici označava baznu točnost tj. točnost najčešće klase s kojom uspoređujemo klasifikatore u hipotezi H1. Središnja linija u pravokutnicima je medijan, a točka aritmetička sredina. Budući da je medijan (a i aritmetička sredina) svih testiranih klasifikatora iznad crvene linije tj. 0,4883 možemo potvrditi hipotezu H1 (McGill, Tukey i Larsen 1978).



Slika 6.10 Usporedba točnosti klasifikatora kod 10-struke unakrsne validacije s baznom točnošću tj. proporcijom najčešće klase (označeno crvenom linijom).

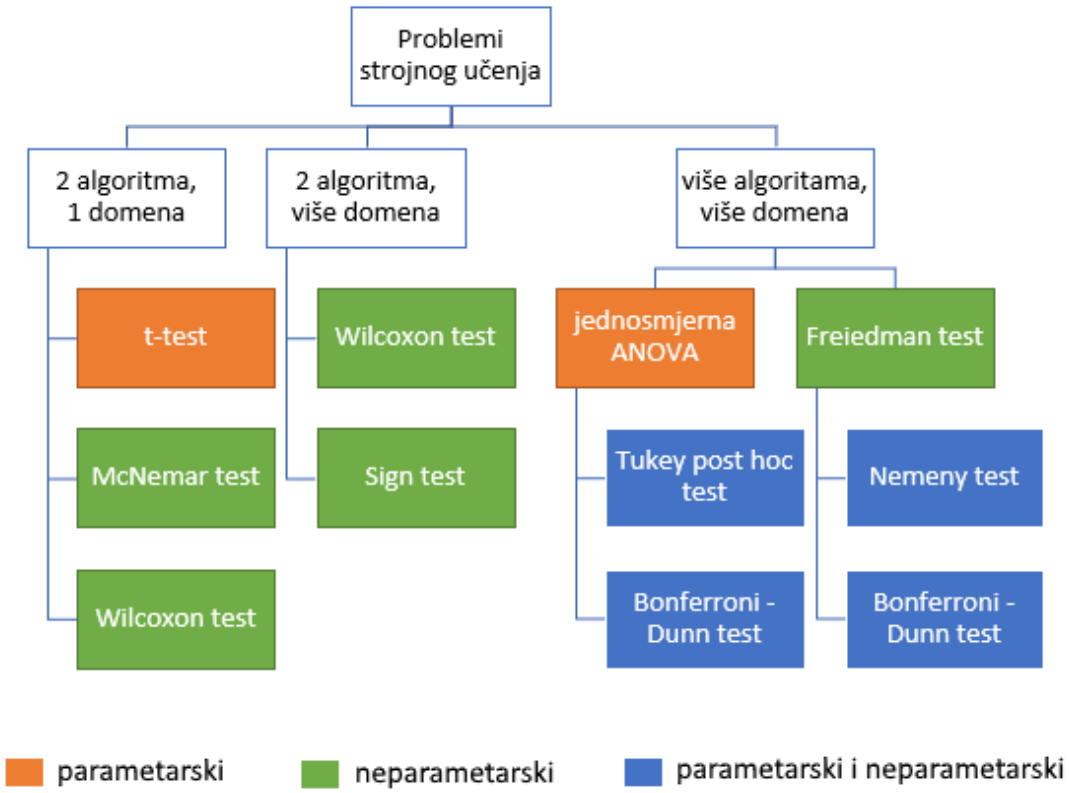
Kod uspoređivanja performansi više klasifikatora na jednom skupu podataka mogu se koristi parametarski i neparametarski testovi (Japkowicz i Shah 2011). Najčešći parametarski test je dvosmjerni t-test, no on zahtijeva određene preduvjete kao veličina uzorka, normalnost distribucije i homogenost varijanci. Zbog toga se kod uspoređivanja klasifikatora češće koriste neparametarski testovi McNemara i Wilcoxon test.

Slika 6.11 prikazuje distribuciju točnosti prema klasifikatorima. Iz slike možemo vidjeti da se normalnost distribucije može pretpostaviti samo kod naivnog Bayesovog klasifikatora. Rezultati ostalih klasifikatora (mjere točnosti) ne prikazuju karakteristike normalnosti, no treba se uzeti u obzir da se radi o vrlo malom uzorku (n=10), te se normalnost ne može sa sigurnošću utvrditi.



Slika 6.11 Histogrami distribucija točnosti po klasifikatorima

Na slici 6.12 prikazani su slučajevi kada se koriste koji testovi (Japkowicz i Shah 2011). Budući da u ovom slučaju uspoređujemo klasifikatore s baznom točnošću tj. najčećom klasom u domeni optimizacije mrežnih stranica za Internetske tražilice potrebno je uzeti u razmatranje testove kategorizirane pod „2 algoritma, 1 domena“ – dva algoritma zbog toga što ćemo uspoređivati u paru (svaki algoritam s baznom točnošću), jedna domena zbog toga što su svi klasifikatori trenirani nad istim podacima.



Slika 6.12 Podjela statističkih testova po problemima strojnog učenja i vrsti testa (Izvor: Japkowicz i Mohak, 2011)

McNemarov test je pogodan kada imamo zaseban skup podataka za validaciju (različit od skupa za treniranje) – nije pogodan za unakrsnu validaciju (Dietterich 1998). Stoga u ovom istraživanju možemo provesti McNemarov test na metodi izdvajanja (engl. *holdout*) tj. na 33% skupa podataka. Oba algoritma (modela) koja se testiraju moraju biti trenirani i testirani na istom skupu podataka. Test zahtijeva izradu kontingencijske tablice. Ako testiramo algoritme A i B koji generiraju modele M_A i M_B onda se kontingencijska tablica kreira ovako:

Broj primjera krivo klasificiranih s M_A i M_B	Broj primjera krivo klasificiranih s M_A, ali točno s M_B
Broj primjera krivo klasificiranih s M_B, ali točno s M_A	Broj primjera točno klasifciranih s M_A i M_B

Možemo upotrijebiti sljedeću notaciju:

n_{00}	n_{01}
n_{10}	n_{11}

gdje je $n_{00}+n_{01}+n_{10}+n_{11}$ ukupan broj uzorka u skupu za testiranje (validaciju) i gdje $n_{10}+n_{01}$ treba biti veće ili jednako 30.

Po null-hipotezi dva algoritma trebaju imati istu stopu pogreške, što znači $n_{10}=n_{01}$. McNemarov test se izračunava kao:

$$\frac{(|n_{01}-n_{10}|-1)^2}{n_{01}+n_{10}} \tag{27}$$

Kreirane su kontingencijske tablice za svaki par klasifikator-bazna točnost (ukupno 5) te je za svaki par izračunat McNemarov test. Sve kontingencijske tablice zadovoljavaju uvjet $n_{10}+n_{01}>=30$. Rezultati su prikazani u tablici 6.7.

Tablica 6.7 Rezultati McNemarovog testa usporedbe klasifikatora s baznom točnošću

Klasifikator	**Kontingencijska tablica**		**McNemara X^2**	**p-vrijednost**
J48	35	68	9,9417	0,001616
	35	66		
SVM	35	68	6,75	0,009375
	40	61		
LogReg	40	63	5,703	0,01694
	38	63		
KNN	44	59	5,6277	0,01768
	35	66		
NaiveBayes	36	67	0,28346	0,5944
	60	41		

McNemara vrijednost X^2 treba biti veća od 3,841 (hi-kvadrat distribucija s 1 stupnjem slobode i α=0,05) da bi s 95% pouzdanošću zaključili da je točnost klasifikatora značajno različita od bazne točnosti. Testovi su pokazali da je to slučaj za sve testirane klasifikatore osim za naivni Bayesov klasifikator koji je pokazao najlošije rezultate.

Wilcoxon signed-rank test je drugi neparametarski test koji nam može ukazati na sličnost tj. različitost distribucija dvaju uzorka iz iste populacije. Često se smatra alternativom parametarskom t-testu jer ne zahtijeva normalnost razdiobe. Izračunava se tako da se najprije izračunaju razlike između varijabli (promatranja), zatim se rangiraju od najmanjih do najvećih ignorirajući predznak, te se sumiraju pozitivni i negativni rangovi (W^+ i W^-). Odabere se minimalni $W=\min(W^-, W^+)$ što predstavlja Wilcoxon vrijednost.

Ako je broj promatranja (*n*) manji ili jednak 25 (Japkowicz i Shah 2011) onda se p-vrijednost traži u Wilcoxonovoj tablici kritičnih vrijednosti te se uspoređuje da li je vrijednost veća ili manja kako bi odlučili o odbacivanju ili neodbacivanju null hipoteze. Ako je *n* veći od 25 onda se može pretpostaviti normalnost distribucije te se izračunava z-statistika na sljedeći način:

$$z_{wilcox} = \frac{W - \mu_W}{\sigma_W} \tag{28}$$

Gdje je μ_w aritmetička sredina normalne aproksimacije distribucije W:

$$\mu_W = \frac{n(n+1)}{4} \tag{29}$$

a σ_W standardna devijacija normalne aproksimacije distribucije W:

$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{30}$$

Dobivena z_{wilcox} vrijednost se zatim traži u tablicama normalne distribucije kako bi se odlučilo o odbacivanju null hipoteze. Važno je napomenuti da se *n* umanjuje za broj gdje je razlika između parova nula.

Uzimajući rezultate 10-struke unakrsne validacije klasifikatora (n=10) izračunavamo W za svaki par klasifikatora (tj. par klasifikator-bazna točnost) te dobivamo p-vrijednost kako je prikazano u tablici 6.8.

Tablica 6.8 Rezultati Wilcox signed rank testa za 10-struku unakrsnu validaciju

	Bazna točnost	**J48**	**Knn**	**SVM**	**LogReg**	**N.Bayes**
1	0,48	0,59	0,73	0,63	0,63	0,49
2		0,70	0,73	0,62	0,59	0,53
3		0,68	0,67	0,57	0,68	0,68
4		0,62	0,68	0,64	0,65	0,53
5		0,75	0,78	0,69	0,62	0,50
6		0,71	0,75	0,75	0,67	0,43
7		0,65	0,63	0,58	0,58	0,54

8		0,58	0,65	0,68	0,68	0,63
9		0,75	0,63	0,74	0,60	0,61
10		0,73	0,70	0,72	0,58	0,52
Prosjek	0,48	0,68	0,69	0,66	0,63	0,55
W	-	55	55	55	55	52
p	-	0,001953	0,001953	0,001953	0,001953	0,009766

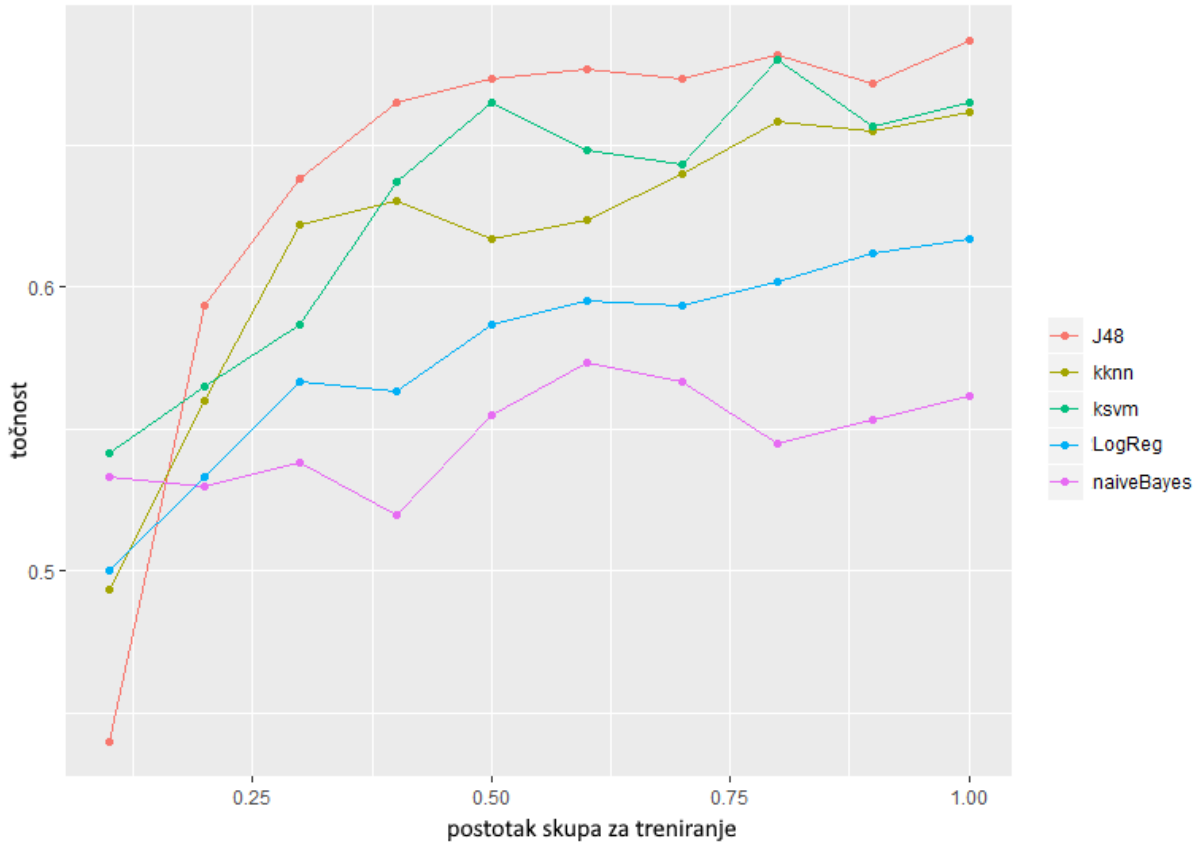
p-vrijednosti u tablici 6.8 pokazuju da su svi testirani klasifikatori značajno bolji od bazne točnosti. Nešto lošiji rezultat daje naivni Bayesov algoritam, ali i dalje s prihvatljivo niskom p-vrijednošću od 0,009766.

Za analizu razlika između točnosti klasifikatora koristimo Friedmanov test (Milton 1940). To je neparametarski test koji se smatra zamjenom za parametarsku ANOVU. On daje $X^2(4)$=14,812 i p-vrijednost od 0,05106 koja je veća od α=0,05 što znači da postoji statistički signifikantna razlika između nekih klasifikatora. Dodatnim post-hoc testovima utvrđujemo između kojih parova postoje te razlike. Nemenyi test (Nemenyi 1962) daje sljedeće rezultate:

```
           j48    NaiveBayes KNN    SVM
NaiveBayes 0.0377 -          -      -
KNN        0.9692 0.0048     -      -
SVM        0.9986 0.0808     0.8896 -
LogReg     0.8267 0.3925     0.4357 0.9371
```

Iz ovih rezultata možemo iščitati da značajne razlike postoje između NaiveBayes i j48 ($p<0,05$), te između NaiveBayes i KNN.

Za analizu podobnosti veličine uzorka za treniranje klasifikatora napravljena je krivulja učenja koja prikazuje kretanje odabrane metrike točnosti s obzirom na broj (tj. postotak) uzorka uključenih u skup za treniranje. Krivulja je prikazana na slici 6.13.



Slika 6.13 Krivulja učenja odabranih klasifikatora

Iz slike 6.13 je vidljivo da točnost počinje slabije rasti nakon upotrebe 50% podataka što pokazuje da je veličina uzorka korištena za klasifikaciju metodom izdvajanja podobna.

6.4. Karakteristike značajnih faktora sadržaja

Promatrajući vrijednosti 5 značajnih faktora za mrežne stranice koje su u kategoriji „prilagođeno“ možemo doći do pogodnih vrijednosti na koje sustav za prilagodbu stranica treba ciljati. Karakteristike tih vrijednosti možemo vidjeti u tablici 6.9 koja prikazuje frekvenciju pojave vrijednosti.

Tablica 6.9 Frekvencije vrijednosti značajnih faktora stranica (varijabli) u klasi „prilagođeno“

Vrijednost	Tkw	Mkw	Mlen	h1Kw	txtKw
0	10	7	0	77	3
1	69	48	0	43	27
2	54	59	0	31	43
3	21	27	0	5	34

4	3	11	2	2	30
5	1	6	2	1	14
6	2	1	2	1	5
7	0	1	2	1	1
8	1	1	3	0	1
9	0	0	4	0	3
10+	0	0	146	0	0

Svih 5 značajnih faktora osim „Mlen“ (dužina meta opisa) istog su tipa – predstavljaju frekvenciju pojave ključne riječi u tekstu, te prema tablici 6.9 možemo zaključiti da je najčešća pojava 1-3 ključnih riječi odnosno 2-4 ključnih riječi u tekstu tijela stranica (varijabla „txtKw“). Treba izuzeti vrijednost 0 iz promatranja jer je ona u kontradikciji sa značajnošću faktora (faktori ne bi bili značajni da je frekvencija pojave ključnih riječi 0) ,a pojavljuje se zbog utjecaja drugih faktora (koji su manje značajni). Varijabla „Mlen“ predstavlja dužinu teksta u broju riječi, a iz tablice 6.9 možemo iščitati da dobro prilagođene stranice imaju „meta opis“ duži od 10 riječi. S obzirom da frekvencija ključnih riječi ovisi i o dužini teksta, potrebno je sagledati i te varijable i promatrati gustoću ključnih riječi (frekvencija / dužina). Gustoća je prikazana u tablici 6.10.

Tablica 6.10 Gustoća ključnih riječi u značajnim faktorima

	Tkw	**Mkw**	**h1Kw**	**txtKw**
Prosjek	0,25641	0,10919	0,17514	0,00936
Standardna pogreška	0,01522	0,00915	0,02121	0,00078
Medijan	0,22222	0,08333	0,03922	0,00661
Mod	0,25	0,07692	0	0,01008
Standardna devijacija	0,19311	0,11605	0,26917	0,00986

Možemo zaključiti da je najveća gustoća ključnih riječi potrebna u naslovu stranice (varijabla „Tkw“) i „h1 naslovu“. Sustav za prilagodbu stranica (objašnjen u sljedećoj fazi) treba znati i kolika je pogodna dužina teksta na pozicijama značajnih faktorima. Analizirajući dužine teksta u „prilagođenim“ stranicama dolazimo do podataka prikazanim u tablici 6.11.

Tablica 6.11 Dužina teksta (broj riječi) na pozicijama značajnih faktora

	Tlen	Mlen	h1len	txtLen
Prosjek	7,96273	25,548447	6,645962	653,0683
Standardna pogreška	0,35682	1,1615643	1,027623	48,36581
Medijan	7	23	3	445
Mod	8	23	0	397
Standardna devijacija	4,52754	14,738599	13,03908	613,6933

Iz tablice 6.11 možemo izvući pogodne dužine za oznaku „title“ (7-8 riječi), „meta description“ (25 riječi) i „h1“ (6-7 riječi). Varijabla „txtLen“ (dužina teksta u tijelu stranice) ima visoku standardnu devijaciju te prosjek ovdje nije realan, ali promatrajući medijan možemo zaključiti da je oko 400 riječi za tekst pogodna vrijednost.

6.5. Sustav za prilagodbu

U prethodnoj fazi utvrđeno je 5 značajnih faktora: Tkw (ključna riječi u oznaci „title“), h1kw (ključna riječ u oznaci „h1“), Mkw (ključna riječ u oznaci „meta description“), Mlen (dužina teksta u oznaci „meta description“) i txtKw (ključna riječ u tekstu stranice). Sustav za prilagodbu ciljati će na tih 5 faktora, te osim njih, u sustav ulaze još ciljane ključne riječi i naravno HTML-kod neprilagođene mrežne stranice.

U sustav za prilagodbu ulaze mrežne stranice svrstane u klasu „neprilagođeno“. Takvih u ovom skupu podataka ima 146. Karakteristike značajnih varijabli tih stranica prikazan je u tablici 6.12.

Tablica 6.12 Frekvencije vrijednosti nekih varijabli stranica u klasi „neprilagođeno“

Vrijednost	Tlen	Tkw	Mkw	Mlen	h1	h1Kw	txtKw
0	2	102	139	123	86	131	62
1	17	28	5	0	43	9	26
2	24	14	1	0	3	4	28
3	32	2	1	2	2	2	15
4	16	0	0	0	2	0	7
5	12	0	0	2	0	0	2
6	8	0	0	2	1	0	0
7	8	0	0	0	3	0	2
8	6	0	0	1	1	0	4

9	8	0	0	0	0	0	0
10+	10	0	0	16	0	0	0

Iz tablice 6.12 vidimo da većina stranica koje su u klasi „neprilagođeno“ ima loše ili nema uopće vrijednosti na značajnim varijablama. Samo dvije stranice nemaju naslov stranice (varijabla „Tlen“), 17 njih ima naslov od samo jedne riječi, 24 njih samo dvije riječi, dok ostatak uglavnom ima naslov, ali s niskom frekvencijom ključnih riječi (varijabla „Tkw“). To su loše vrijednosti prema SEO preporukama. Što se tiče „meta opisa“ čak 123 stranice od 146 nema meta opis. Za ispravak te situacije dobro će poslužiti algoritam sažimanja teksta (Matošević, 2018). Velik je broj i stranica koje nemaju oznake „h1“ (86). U tu svrhu poslužit će algoritam utvrđivanja najznačajnije rečenice iz teksta kao kandidata za „h1“ naslov ukoliko naslov ne postoji, a ukoliko postoji, ali nije formatiran kao „h1“ u HTML-u, algoritmom se može pronaći takav naslov i preoblikovati u „h1“. Loša frekvencija ključnih riječi u bilo kojem segmentu tj. tekstnoj varijabli pokušati će se ispraviti pomoću WordNet-a tehnikom ubacivanja sinonima i hiperonima.

Za vršenje navedenih ispravaka nužno je da stranica ima tekst. Tekst tijela stranice glavna je ulazna varijabla u sustav ispravka. Ako stranica nema teksta u tijelu, ili je tekst prekratak, jedini način poboljšanja stranice je jednostavno ubacivanje, tj. dodavanje ključnih riječi – a to ima smisla samo u naslov stranice (oznaku „title“) gdje nije nužno da tekst bude narativnog karaktera.

Od 146 neprilagođenih stranica u promatranom skupu podataka, njih 11 nema tekst u tijelu ili je taj tekst kraći od 10 riječi.

Kompletan algoritam ispravka prikazan je u nastavku. Osim teksta mrežne stranice (T) i ciljanih ključni riječi (K), ulaznu varijablu predstavlja i lista ciljanih tj. pogodnih vrijednosti značajnih varijabli (C) koje su u tvrđene analizom stranica iz klase „prilagođeno“. Ulazna varijabla CV predstavlja varijablu koju prilagođavamo (npr. Tkw, Mkw, txtKw itd.), a varijabla S je granična sličnost između riječi koja će se primjenjivati kod uspoređivanja riječi pomoću WordNet-a. Izlazna vrijednost algoritma je ispravljena ulazna varijabla, tj. nova vrijednost ulazne varijable (u tekstnom obliku).

Algoritam ispravka je izgrađen dakle na temelju karakteristika „prilagođenih“ stranica i na temelju SEO preporuka koje izdaju internetske tražilice.

Algoritam 2: Predloženi algoritam ispravka

Ulaz: K=lista ciljanih ključnih riječi, T=tekst stranice, CV=ciljana varijabla, C=lista ciljanih vrijednosti, S=granična vrijednost sličnosti riječi

Izlaz: T2=ispravljeni tekst

Ovisno o CV:

Ako je Mkw

Ako je Mlen=0 onda T2=Sažimanje(K, T, S, C(dužina))

Ako je Mlen>0 onda T2=Obogati(K, Mtext, S, C(gustoća))

Ako je H1kw

Ako je H1=0 onda T2=GenerirajH1T(K, T, S, C(dužina, gustoća), „H1“)

Ako je H1>0 onda T2=Obogati(K, text, S, C(gustoća))

Ako je Tkw

Ako je Tlen=0 onda T2=GenerirajH1T(K, T, S, C(dužina), „Title“)

Ako je Tlen>0 onda T2=Obogati(K, text, S, C(gustoća))

Ako je txtKw

Ako je txtLen<10 onda Izlaz Inače T2=Obogati(K, T, S, C(dužina, gustoća))

Ako je Mlen

Ako je Mlen>C(max.dužina) onda T2=Skrati(K, text, S, C(dužina))

Ako je Mlen<C(min.dužina) onda T2=Sažimanje(K, T, S, C(dužina))

Vrati T2

Za sažimanje koristit će se Algoritam 1 iz poglavlja 5.5.2., a u nastavku su predstavljeni algoritmi „Obogati“ za obogaćivanje teksta ključnim riječima, „GenerirajH1“ za generiranje teksta za naslov „h1“, „GenerirajT“ za generiranje naslova stranice (oznaka „title“), te „Skrati“ za skraćivanje predugačkog teksta.

Algoritam 3 („Obogati“): Obogaćivanje teksta ključnim riječima

Ulaz: K=lista ciljanih ključnih riječi, T=tekst kojeg treba obogatiti, S=granična vrijednost sličnosti riječi, C=maksimalna gustoća ključnih riječi

Izlaz: O=obogaćeni tekst

T=T bez stop-riječi iz T

K=STEM(K)

```
Za svaku riječ u T
    tag=POS_TAG(riječ)
    ako je tag=imenica onda
        riječ=STEM(riječ)
        riječ_s=WordNet_Synset(riječ)
        za svaku ključnu riječ u K
            sim=sličnost(riječ ili riječ_s, ključna riječ)
            ako je sim>S onda zamijeni riječ s ključnom riječi u T
            ako je hiperonim(riječ)=hiperonim(K) onda
                zamijeni riječ s hiperonim(riječ) u T
            ako je gustoća(T)>=C onda izađi iz petlje
O=T
Vrati O
```

Algoritam 4 („GenerirajH1T“): Generiranje H1 naslova ili „Title“ naslova iz teksta

Ulaz: K=lista ciljanih ključnih riječi, T=tekst stranice, S=granična vrijednost sličnosti riječi, C=lista ciljanih vrijednosti (dužina i gustoća), Tip=H1 ili Title

Izlaz: H=predloženi tekst

```
H1prijedlog=pronađi rečenice u T koje su formatirane većim fontom
Ako je broj(H1prijedlog) = 0 onda
    H=pronađi najkraću najznačajniju rečenicu u T (K, T)
Inače ako je broj(H1prijedlog)=1 onda
    H=H1prijedlog
Inače ako je broj(H1prijedlog)>1 onda
    H=najkraća rečenica s najvećom gustoćom
Ako je gustoća(H)<C(gustoća) onda H=Obogati (K, H1, S, C(gustoća))
Ako je dužina(H)>C(max.dužina) onda H=Skrati(K, H1, S, C(dužina))
Ako je Tip="Title" i gustoća(H)<C(gustoća) onda H=H+K
Vrati H
```

Algoritam 5 („Skrati“): Skraćivanje teksta

Ulaz: K=lista ciljanih ključnih riječi, T=tekst stranice, S=granična vrijednost sličnosti riječi, C=maksimalna dužina

Izlaz: ST=skraćeni tekst

Sve dok je dužina(T)>C

ST=Izbaci rečenicu s najmanjom gustoćom ključnih riječi (K, S) iz T

Vrati ST

Navedeni algoritmi implementirani su u programskom jeziku Python korištenjem NLTK Toolkita (Prilog br. 4) te su sve stranice iz klase „neprilagođeno“ provedene kroz algoritam ispravka. Ispravljene stranice tj. nove vrijednosti varijabli ručno su pregledane od strane SEO eksperta te su izbačeni ispravci koji nemaju smisla (zbog jezične i logičke forme). Tablica 6.13 prikazuje broj predloženih i prihvaćenih zamjena riječi po pojedinoj varijabli.

Tablica 6.13 Broj predloženih (od strane algoritma) i prihvaćenih zamjena riječi (od strane stručnjaka) po pozicijama (varijablama)

	Title	**Meta**	**H1**	**Tekst**	**Ukupno**
Predloženo	23	15	8	25	71
Prihvaćeno	14	13	5	23	55

Od ukupno 71 zamjene njih 55 je prihvaćeno. U tablici 6.14 prikazan je broj poboljšanja po varijablama za ukupno 146 stranica koje se nalaze u najlošijoj klasi koje su ušle u sustav ispravka.

Tablica 6.14 Broj poboljšanja po varijablama nakon izlaza iz sustava ispravka

	Tlen	**Tkw**	**Mlen**	**Mkw**	**H1**	**H1len**	**H1kw**	**txtKw**
Broj poboljšanja	46	48	100	100	46	46	42	27

Iz tablice 6.14 je vidljivo da je najviše poboljšanja vrijednosti varijabli učinjeno na poziciji „meta opisa“, odnosno 100 od 146 stranica, a najmanji broj poboljšanja je u varijabli pojave ključnih riječi u tekstu stranice (varijbla „txtKw“) – samo na 27 stranica od ukupno 146. Poboljšanje varijable „H1“ sustav je pokušao (predložio) u 70 slučajeva, međutim prihvaćeno je 46. Radi se o poboljšanjima specifičnim za pojedinu varijablu. Kod varijabli dužine poboljšanja se odnose na povećanje dužine (naslova stranice, meta opisa, h1 naslova itd.). Kod varijabli povezanim s gustoćom ključnih riječi radi se o poboljšanjima te gustoće tj. dodavanjem ključnih riječi na određene pozicije.

Od ukupno 146 stranica iz klase „neprilagođeno“ njih 118 je imala bar jednu promjenu u vrijednostima varijabli, tj. njih 28 je izašlo iz sustava ispravka u istom stanju u kojem je i ušlo. To znači da tih 28 stranica zadržavaju svoju klasu „neprilagođeno“, a ostatak od 118 ulazi u model klasifikatora na ponovnu klasifikaciju. Taj skup podataka s ispravljenim vrijednostima dan je u prilogu br.5.

6.6. Ponovna klasifikacija ispravljenih stranica

Ispravljene stranice pomoću algoritama za ispravak iz prethodnog poglavlja učitane su u modele klasifikatora koji su razvijeni u prvom dijelu ovog istraživanja. Svrha ponovne klasifikacije je evaluacija sustava ispravka i odgovaranje na pitanja da li sustav ispravka može povećati stupanj prilagođenosti mrežnih stranica SEO preporukama. Rezultati ponovne klasifikacije prikazani su u tablici 6.15.

Tablica 6.15 Rezultati uspješnosti ponovne klasifikacije ispravljenih stranica

	J48	**Knn**	**SVM**	**LogReg**	**N.Bayes**
Klasa „neprilagođeno“	12	16	13	9	25
Klasa „djelomično prilagođeno“	62	72	53	51	55
Klasa „prilagođeno“	44	30	52	58	38
Postotak stranica koje su prešle u višu klasu – od ukupnog broja 146	72%	70%	72%	75%	64%
Postotak stranica koje su prešle u višu klasu – od broja stranica koje su ušle u ponovnu klasifikaciju - 118	89%	86%	89%	92%	79%

Iz tablice 6.15 vidljivo je da je hipoteza H2 potvrđena uporabom svih korištenih algoritama za klasifikaciju: svi klasifikatori su svrstali više od 50% „neprilagođenih“ stranica u višu klasu. Pri tome je najuspješniji bio klasifikator logističke regresije (samo 9 uzoraka je ostalo u klasi „neprilagođeno“), dok je najmanje uspjeha imao naivni Bayesov klasifikator (sa 25 uzoraka koje nisu prešli u višu klasu). Broju neuspješnih klasifikacija u višu klasu iz tablice 6.15 treba pridodati i 28 stranica koje nisu uopće ušle u ponovnu klasifikaciju jer im varijable nisu promijenjene u procesu ispravka (te instance su ostale u klasi „neprilagođeno“). Iznenađuje visok postotak stranica koje su prešle u najbolju klasu „prilagođeno“. Razlog tomu

možemo tražiti u izuzetno važnim faktorima „meta opisa“, „Mkw“ i „Mlen“. Proces sažimanja je vrlo dobro obavio posao kod generiranja „meta opisa“ koji je nedostajao kod većina stranica u klasi „neprilagođeno“. Budući da je to jedan od najvažnijih faktora (kako se pokazalo u prvom dijelu istraživanja), njegov kvalitetan „ispravak“ u sustavu ispravka pridonio je ovakvom rezultatu. Možemo pretpostaviti da je najmanji učinak u poboljšanju stranica imao broj ključnih riječi u tekstu (varijabla „txtKw“), budući da je nad tom varijablom sustav učinio najmanje promjena (tablica 6.14).

7. ZAKLJUČAK

Optimizacija mrežnih stranica za internetske tražilice proces je koji uključuje prilagodbu sadržaja stranice (HTML-koda i teksta), te rad na dobivanju što više dolaznih poveznica. U ovom radu istražena je upotreba strojnog učenja i obrade prirodnoga jezika u procesu prilagodbe sadržaja. Brojna prethodna istraživanja bave se utvrđivanjem značajnih faktora koji utječu na poziciju stranica na rezultatima pretrage. U ovom istraživanju koristi se strojno učenje, tj. modeli klasifikatora za utvrđivanje tih faktora. Koristeći znanje SEO stručnjaka izgrađeni su klasifikacijski modeli koji mogu s prosječnom točnošću između 54,69% i 69,67% (ovisno o klasifikatoru) uspješno predvidjeti stupanj prilagodbe određene stranice SEO preporukama koje objavljuju internetske tražilice. Klasifikatori su trenirani nad uzorkom od 600 stranica slučajno odabranih iz direktorija DMOZ i označenih od strane tri SEO stručnjaka u tri predefinirane kategorije: „prilagođeno“, „djelomično prilagođeno“ i „neprilagođeno“. Pri tome su korištene ključne riječi iz naziva kategorije u kojoj su stranice svrstane u direktoriju DMOZ. Formiran je popis od 21 varijable (na osnovu prethodnih istraživanja) te su njihove vrijedosti izvučene iz promatranih stranica. Testiranjem 5 odabranih klasifikatora (stabla odlučivanja, naivni Bayes, Logistička regresija, KNN metoda i metoda potpornih vektora) postignuta je točnost veća od bazne točnosti (48,83%) koja predstavlja većinsku klasu „djelomično prilagođeno“, čime je potvrđena hipoteza H1. Skup podataka koji je formiran u ovom istraživanju, a koji se sastoji od 600 primjera s vrjednostima 21 nezavise varijable i zavisne varijable klase u koju stranica pripada može se koristiti u daljnjim istraživanjima o utjecaju faktora (varijabli) ili upotrebi raznih klasifikacijskih algoritama u ovoj domeni.

Stranice koje su SEO stručnjaci svrstali u najbolju klasu „prilagođeno“ poslužile su za ekstraciju poželjnih vrijednosti promatranih varijabli. Istraživanjem je utvrđeno da modeli klasifikatora mogu poslužiti za ekstrakciju značajnih faktora (varijabli), te da je popis tako dobivenih varijabli u skladu s prethodnim istraživanjima. Radi se o sljedećim varijablama: Mkw (ključne riječi u meta opisu stranice), Tkw (ključne riječi u oznaci „title“), Mlen (dužina meta opisa), h1kw (ključne riječi u oznaci „h1“) i txtKw (ključne riječi u tekstu stranice).

Stranice koje su SEO stručnjaci svrstali u klasu „neprilagođeno“ korištene su za testiranje predloženog sustava prilagodbe. Osim navedenih stranica u sustav prilagodbe ušli su i utvrđeni značajni faktori i njihove poželjne vrijednosti. Predloženi algoritmi koji koriste

metode iz obrade prirodnoga jezika predložili su ispravke koji su dani na evaluaciju SEO stručnjaku. Samo prihvaćeni ispravci od strane SEO stručnjaka su i implementirani nad varijablama tj. stranicama koje su ušle u sustav prilagodbe. Sustav prilagodbe je dakle ispravio stranice, a koliko su navedeni ispravci bili dobri ili ne testirano je upotrebom klasifikatora izgrađenih u prvoj fazi ovog istraživanja. Rezultati su pokazali da su izgrađeni modeli klasifikatora u prosjeku 70,6% neprilagođenih stranica nakon prilagodbe svrstali u višu klasu („djelomično prilagođeno“ ili „prilagođeno“). Pri tome je nauspješniji bio klasifikator logističke regresije (75%), a najlošiji naivni Bayesov klasifikator (64%). Najveći doprinos ovako visokoj uspješnosti sustava prilagodbe možemo tražiti u metodama sažimanja teksta koje su korištene za generiranje meta opisa stranica, čime se utjecalo na varijable Mkw (ključne riječi u meta opisu) i Mlen (dužina meta opisa) koje imaju zanačajan utjecaj na klasu stupnja prilagodbe.

Znanstveni doprinos ovog istraživanja je višestruk. Kreiran je skup podataka koji se može koristiti u budućim istraživanjima vezanim za karakteristike mrežnih stranica i SEO. Korišteno je znanje stručnjaka i strojno učenje za izgradnju modela klasifikatora radi utvrđivanja stupnja prilagođenosti stranice SEO preporukama te utvrđivanju značajnih faktora sadržaja. Izgrađeni su algoritmi za sustav ispravka loše prilagođenih stranica uz korištenje tehnika iz domene obrade prirodnoga jezika koji mogu poslužiti kao polazna točka u izgradnji softvera za tu namjenu.

Metode korištene u ovom istraživanju nisu vezane za internetsku tražilicu ili jezik stranica. Iste metode mogu se primijeniti ako SEO stručnjaci za označavanje uzoraka primjene svoje znanje vezano za neku drugu tražilicu ili područje primjene (npr. otkrivanje zlonamjernih stranica). Također, jezik samih stranica nije ograničavajući faktor dok god postoji leksička baza za taj jezik (leksičkosemantička baza riječi sa skupovima sinonima i leksičkim vezama hipo-hiperonima).

U ovom istraživanju promatrani su samo faktori sadržaja. Vanjski faktori poveznica nisu uključeni u istraživanje. Iz toga i proizlaze određena ograničenja koja se prvenstveno odnose na postojanje teksta na stranici. Ako mrežna stranica nema teksta onda nije ni moguća sažimanje i obogaćivanje teksta ključnim riječima, što su glavni procesi u sustavu ispravka.

Daljnja istraživanja mogla bi uključivati procese iz domene generiranja prirodnog jezika na osnovu konteksta mrežne stranice, ali i uključivanje dodatnih varijabli, posebice onih vanjskih

(vezanih uz hiperveze). Također, povećanjem uzorka stranica mogle bi se istražiti karakteristike stranica prema kategorijama u koje pripadaju.

8. POPIS LITERATURE

Abdullah, Kegesa Danvas. »Search Engine optimization Techniques by Google's Top Ranking Factors: Website Ranking Signals.« (samostalna naklada) 2017.

Agrawal, Sukrati, Antriksha Somani, i Vishal Chhabra. »Discernment of Search Engine Spamming and Counter Measure for It.« *International Journal of Computer Applications* (Foundation of Computer Science) 147, br. 8 (2016).

Aliakbary, Sadegh, Hassan Abolhassani, Hossein Rahmani, i Behrooz Nobakht. »Web page classification using social tags.« *Computational Science and Engineering, 2009. CSE'09. International Conference on.* IEEE, 2009. 588-593.

Andersson, Viktor, i Daniel Lindgren. »Ranking factors to increase your positionon the search engine result page: Theoretical and practical examples.« 2017.

Ben-Hur, Asa, i Jason Weston. »A user's guide to support vector machines.« U *Data mining techniques for the life sciences*, autor Asa Ben-Hur i Jason Weston, 223-239. Springer, 2010.

Berger, Adam L, i Vibhu O Mittal. »OCELOT: a system for summarizing Web pages.« *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2000. 144-151.

Bergstra, James S, Rémi Bardenet, Yoshua Bengio, i Balázs Kégl. »Algorithms for hyper-parameter optimization.« *Advances in neural information processing systems.* 2011. 2546-2554.

Bing Webmaster Guidelines. n.d. https://www.bing.com/webmaster/help/webmaster-guidelines-30fba23a. (datum pristupa: 15.5.2018)

Breiman, Leo. »Bagging predictors.« *Machine learning* (Springer) 24, br. 2 (1996): 123-140.

Breiman, Leo. *Classification and regression trees.* Routledge, 2017.

Brin, Sergey, i Lawrence Page. »The anatomy of a large-scale hypertextual web search engine.« *Computer networks and ISDN systems* (Elsevier) 30, br. 1-7 (1998): 107-117.

Buddenbrock, Frank. »Search Engine Optimization: Getting to Google's First Page.« U *Google It*, autor Frank Buddenbrock, 195-204. Springer, 2016.

Cai, Yuanyuan, Qingchuan Zhang, Wei Lu, i Xiaoping Che. »A hybrid approach for measuring semantic similarity based on IC-weighted path distance in WordNet.« *Journal of Intelligent Information Systems* (Springer) 51, br. 1 (2018): 23-47.

Chris, D Paice. »Another stemmer.« *ACM SIGIR Forum.* 1990. 56-61.

Cohen, Jacob. »A coefficient of agreement for nominal scales.« *Educational and psychological measurement* (Sage Publications Sage CA: Thousand Oaks, CA) 20, br. 1 (1960): 37-46.

Cover, Thomas, i Peter Hart. »Nearest neighbor pattern classification.« *IEEE transactions on information theory* (IEEE) 13, br. 1 (1967): 21-27.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, i Richard Harshman. »Indexing by latent semantic analysis.« *Journal of the American society for information science* (Wiley Online Library) 41, br. 6 (1990): 391-407.

Dietterich, Thomas G. »Approximate statistical tests for comparing supervised classification learning algorithms.« *Neural computation* (MIT Press) 10, br. 7 (1998): 1895-1923.

Edmundson, Harold P. »New methods in automatic extracting.« *Journal of the ACM (JACM)* (ACM) 16, br. 2 (1969): 264-285.

Erkan, Günes, i Dragomir R Radev. »Lexrank: Graph-based lexical centrality as salience in text summarization.« *Journal of artificial intelligence research* 22 (2004): 457-479.

Fleiss, Joseph L. »Measuring nominal scale agreement among many raters.« *Psychological bulletin* (American Psychological Association) 76, br. 5 (1971): 378.

Francis, W Nelson, i Henry Kucera. »Brown corpus.« *Department of Linguistics, Brown University, Providence, Rhode Island* 1 (1964).

Freund, Yoav, i Robert E Schapire. »Experiments with a new boosting algorithm.« *Icml.* Citeseer, 1996. 148-156.

Gamberger, D., i T. Šmuc. »Poslužitelj za analizu podataka [http://dms.irb.hr].« Institut Rudjer Bošković, Laboratorij za informacijske sustave, 2001.

García-Pedrajas, Nicolás, Juan A Romero del Castillo, i Gonzalo Cerruela-García. »A Proposal for Local $ k $ Values for $ k $-Nearest Neighbor Rule.« *IEEE transactions on neural networks and learning systems* (IEEE) 28, br. 2 (2017): 470-475.

Gaudette, Lisa, i Nathalie Japkowicz. »Evaluation methods for ordinal classification.« *Canadian Conference on Artificial Intelligence.* Springer, 2009. 207-210.

Giomelakis, Dimitrios, i Andreas Veglis. »Investigating search engine optimization factors in media websites: the case of Greece.« *Digital Journalism* (Taylor & Francis) 4, br. 3 (2016): 379-400.

Gong, Yihong, i Xin Liu. »Generic text summarization using relevance measure and latent semantic analysis.« *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2001. 19-25.

Google Webmaster Guidelines. n.d. https://support.google.com/webmasters/answer/35769?hl=en (datum pristupa: 15.5.2018).

Gupta, Swati, Nitin Rakesh, Abha Thakral, i Dev Kumar Chaudhary. »Search engine optimization: Success factors.« *Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on.* IEEE, 2016. 17-21.

Haghighi, Aria, i Lucy Vanderwende. »Exploring content models for multi-document summarization.« *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, 2009. 362-370.

Hsu, Chih-Wei, i Chih-Jen Lin. »A comparison of methods for multiclass support vector machines.« *IEEE transactions on Neural Networks* (IEEE) 13, br. 2 (2002): 415-425.

Hussien, A S. »Factors Affect Search Engine Optimization.« *International Journal of ComputerScience and Network Security* 14, br. 9 (2014): 28-33.

James, Gareth, Daniela Witten, Trevor Hastie, i Robert Tibshirani. *An introduction to statistical learning.* Svez. 112. Springer, 2013.

Japkowicz, Nathalie, i Mohak Shah. *Evaluating learning algorithms: a classification perspective.* Cambridge University Press, 2011.

Kass, Gordon V. »An exploratory technique for investigating large quantities of categorical data.« *Applied statistics* (JSTOR), 1980: 119-127.

Kendall, M. G. »A New Measure of Rank Correlation.« *Biometrika* (Oxford University Press) 30, br. 1/2 (1938): 81-93.

Khan, M N A, i A Mahmood. »A distinctive approach to obtain higher page rank through search engine optimization.« *Sādhanā* (Springer) 43, br. 3 (2018): 43.

Kilgore, Ty. »What is an SEO specialist?« *SEO.com*, https://www.seo.com/blog/what-is-an-seo-specialist/ (datum pristupa 1.4.2019).

Kleinberg, Jon M, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, i Andrew S Tomkins. »The web as a graph: measurements, models, and methods.« *International Computing and Combinatorics Conference.* Springer, 1999. 1-17.

Korayem, Mohammed, Camilo Ortiz, Khalifeh AlJadda, i Trey Grainger. »Query sense disambiguation leveraging large scale user behavioral data.« *Big Data (Big Data), 2015 IEEE International Conference on.* IEEE, 2015. 1230-1237.

Kullback, S., i R. A. Leibler. »On Information and Sufficiency.« *The Annals of Mathematical Statistics* 22, br. 1 (1951): 79-86.

Landis, J Richard, i Gary G Koch. »The measurement of observer agreement for categorical data.« *biometrics* (JSTOR), 1977: 159-174.

Larose, Daniel T., i Chantal D. Larose. *Data mining and predictive analytics.* New Jersey: John Wiley & Sons, Inc., 2015.

Leacock, Claudia, i Martin Chodorow. »Combining local context and WordNet similarity for word sense identification.« *WordNet: An electronic lexical database* 49, br. 2 (1998): 265-283.

Lee, Ji-Hyun, Wei-Chang Yeh, i Mei-Chi Chuang. »Web page classification based on a simplified swarm optimization.« *Applied Mathematics and Computation* (Elsevier) 270 (2015): 13-24.

Liaw, Andy, i Matthew Wiener. »Classification and regression by randomForest.« *R news* 2, br. 3 (2002): 18-22.

Loh, Wei-Yin, i Yu-Shan Shih. »Split selection methods for classification trees.« *Statistica sinica* (JSTOR), 1997: 815-840.

Loper, Edward, i Steven Bird. »NLTK: The natural language toolkit.« *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1.* Association for Computational Linguistics, 2002. 63-70.

Lovins, Julie Beth. »Development of a stemming algorithm.« *Mech. Translat. & Comp. Linguistics* 11, br. 1-2 (1968): 22-31.

Luh, Cheng-Jye, Sheng-An Yang, i Ting-Li Dean Huang. »Estimating Google's search engine ranking function from a search engine optimization perspective.« *Online Information Review* (Emerald Group Publishing Limited) 40, br. 2 (2016): 239-255.

Luhn, Hans Peter. »The Automatic Creation of Literature Abstracts.« *IBM Journal of Research and Development,* 1959: 159-165.

Mantovani, Rafael G, Tomáš Horváth, Ricardo Cerri, Joaquin Vanschoren, i André CPLF de Carvalho. »Hyper-parameter tuning of a decision tree induction algorithm.« *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on.* IEEE, 2016. 37-42.

Marath, Sathi T, Michael Shepherd, Evangelos Milios, i Jack Duffy. »Large-scale web page classification.« *System Sciences (HICSS), 2014 47th Hawaii International Conference on.* IEEE, 2014. 1813-1822.

Matošević, Goran. »Measuring the utilization of on-page search engine optimization in selected domain.« *Journal of Information and Organizational Sciences* (Fakultet organizacije i informatike Sveučilišta u Zagrebu) 39, br. 2 (2015): 199-207.

Matošević, Goran. »Text Summarization Techniques for Meta Description Generation in Process of Search Engine Optimization.« *Computer Science On-line Conference.* Springer, 2018. 165-173.

Mavridis, Themistoklis, i Andreas L Symeonidis. »Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms.« *Engineering Applications of Artificial Intelligence* (Elsevier) 41 (2015): 75-91.

McGill, Robert, John W Tukey, i Wayne A Larsen. »Variations of box plots.« *The American Statistician* (Taylor & Francis Group) 32, br. 1 (1978): 12-16.

Metaxas, P Takis, i Yada Pruksachatkun. »Manipulation of search engine results during the 2016 US congressional elections.« 2017.

Mihalcea, Rada, i Paul Tarau. »Textrank: Bringing order into text.« *Proceedings of the 2004 conference on empirical methods in natural language processing.* 2004.

Miller, George A. »WordNet: a lexical database for English.« *Communications of the ACM* (ACM) 38, br. 11 (1995): 39-41.

Milton, Friedman. »A comparison of alternative tests of significance for the problem of m rankings.« *The Annals of Mathematical Statistics*, 1940: 86-92.

Moreno, Lourdes, i Paloma Martinez. »Overlapping factors in search engine optimization and web accessibility.« *Online Information Review* (Emerald Group Publishing Limited) 37, br. 4 (2013): 564-580.

Mostafa, Lamiaa. »Webpage Keyword Extraction Using Term Frequency.« *International Journal of Computer Theory and Engineering* (IACSIT Press) 5, br. 1 (2013): 174.

Nadeau, David, i Satoshi Sekine. »A survey of named entity recognition and classification.« *Lingvisticae Investigationes* (John Benjamins publishing company) 30, br. 1 (2007): 3-26.

Nemenyi, Peter. »Distribution-free multiple comparisons.« *Biometrics* 18, br. 2 (1962): 263.

Ng, Hwee Tou, i John Zelle. »Corpus-based approaches to semantic interpretation in NLP.« *AI magazine* 18, br. 4 (1997): 45.

Quinlan, J Ross. *C4. 5: programs for machine learning.* Elsevier, 2014.

Quinlan, J Ross. »Induction of decision trees.« *Machine learning* (Springer) 1, br. 1 (1986): 81-106.

Roul, Rajendra Kumar, Shubham Rohan Asthana, i Gaurav Kumar. »Spam web page detection using combined content and link features.« *International Journal of Data Mining, Modelling and Management* (Inderscience Publishers (IEL)) 8, br. 3 (2016): 209-222.

Ruiz Sánchez, Roberto, Jesús Salvador Aguilar Ruiz, José Cristóbal Riquelme Santos, i Norberto Díaz Díaz. »Analysis of Feature Rankings for Classification.« *Advances in Intelligent Data Analysis VI, Lecture Notes in Computer Science, Volume 3646, pp 362-372 (2005)*, 2005.

Rutz, Oliver J, i Randolph E Bucklin. »Paid search advertising.« U *Advanced Database Marketing*, autor Oliver J Rutz i Randolph E Bucklin, 251-268. Routledge, 2016.

Sagot, Sylvain, Egon Ostrosi, i Alain-Jérôme Fougères. »A multi-agent approach for building a fuzzy decision support system to assist the SEO process.« *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on.* IEEE, 2016. 4001-4006.

Su, Ao-Jan, Y Charlie Hu, Aleksandar Kuzmanovic, i Cheng-Kok Koh. »How to improve your search engine ranking: Myths and reality.« *ACM Transactions on the Web (TWEB)* (ACM) 8, br. 2 (2014): 8.

Sujata, Joshi, Shrivastava Noopur, Nair Neethi, Prakash Jubin, i Pandey Udit. »On-Page Search Engine Optimization: Study of Factors Affecting Online Purchase Decisions of Consumers.« *Indian Journal of Science and Technology* 9, br. 15 (2016).

Sun, Jian-Tao, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, i Zheng Chen. »Web-page summarization using clickthrough data.« *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2005. 194-201.

Teknomo, Kardi. *Minkowski distance.* n.d. https://people.revoledu.com/kardi/tutorial/Similarity/MinkowskiDistance.html (datum pristupa 10. 10. 2018).

Vanderwende, Lucy, Hisami Suzuki, Chris Brockett, i Ani Nenkova. »Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion.« *Information Processing & Management* (Elsevier) 43, br. 6 (2007): 1606-1618.

Wei, Tingting, i Huiyou Chang. »Measuring Word Semantic Relatedness Using WordNet-Based Approach.« *JCP* 10, br. 4 (2015): 252-259.

Wei, Tingting, Yonghe Lu, Huiyou Chang, Qiang Zhou, i Xianyu Bao. »A semantic approach for text clustering using WordNet and lexical chains.« *Expert Systems with Applications* (Elsevier) 42, br. 4 (2015): 2264-2275.

White, Ryen W, Matthew Richardson, i Wen-tau Yih. »Questions vs. queries in informational search tasks.« *Proceedings of the 24th International Conference on World Wide Web.* ACM, 2015. 135-136.

Willett, Peter. »The Porter stemming algorithm: then and now.« *Program* (Emerald Group Publishing Limited) 40, br. 3 (2006): 219-223.

Witten, Ian H, Eibe Frank, Mark A Hall, i Christopher J Pal. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

Wu, Xindong, i dr. »Top 10 algorithms in data mining.« *Knowledge and information systems* (Springer) 14, br. 1 (2008): 1-37.

Wu, Zhibiao, i Martha Palmer. »Verbs semantics and lexical selection.« *Proceedings of the 32nd annual meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 1994. 133-138.

Zhang, Jin, i Alexandra Dimitroff. »The impact of metadata implementation on webpage visibility in search engine results (Part II).« *Information processing & management* (Elsevier) 41, br. 3 (2005): 691-715.

Zhang, Jin, i Alexandra Dimitroff. »The impact of webpage content characteristics on webpage visibility in search engine results (Part I).« *Information Processing & Management* (Elsevier) 41, br. 3 (2005): 665-690.

Zhang, Sonya, i Neal Cabage. »Search engine optimization: Comparison of link building and social sharing.« *Journal of Computer Information Systems* (Taylor & Francis) 57, br. 2 (2017): 148-159.

Zhang, Yongzheng, Nur Zincir-Heywood, i Evangelos Milios. »Term-based clustering and summarization of web page collections.« *Conference of the Canadian Society for Computational Studies of Intelligence.* Springer, 2004. 60-74.

Zhu, Cen, i Guixing Wu. »Research and analysis of search engine optimization factors based on reverse engineeing.« *Multimedia Information Networking and Security (MINES), 2011 Third International Conference on.* IEEE, 2011. 225-228.

9. PRILOZI

Prilog br. 1: Skup podataka

ID;Ključne riječi;
Tlen;Tkw;Mlen;Mkw;h1;h1len;h1kw;h2;h2len;h2kw;h3;h3len;h3kw;alt;altKw;linkKw;linkOut;urlLen
;urlKw;txtLen;txtKw;stručnjak 1;stručnjak 2; stručnjak 3;klasa

1;Arts, Animation, Artists;3;0;52;1;0;0;0;0;0;0;0;0;0;18;0;0;8;26;0;90;1;2;2;2;2

2;Arts, Animation, Awards;2;1;0;0;2;4;1;1;5;1;2;8;1;0;0;1;20;24;1;55;2;2;2;2;2

3;Arts, Animation, Cartoons;3;0;6;1;0;0;0;1;6;0;0;0;0;28;0;5;4;27;0;5836;5;2;2;2;2

4;Arts, Animation, Cartoons;14;6;18;6;11;59;7;1;2;0;0;0;0;5;2;11;16;29;2;450;9;3;3;2;3

5;Arts, Animation, Contests;8;1;0;0;1;3;0;2;3;0;0;0;0;7;0;3;12;29;0;181;4;2;2;2;2

6;Arts, Animation, Festivals;6;1;0;0;0;0;0;47;409;2;0;0;0;49;0;4;117;20;0;685;4;2;2;2;2

7;Arts, Animation, Movies;5;1;26;3;1;4;1;5;19;2;0;0;0;19;5;6;7;31;0;314;5;3;3;2;3

8;Arts, Animation, Organizations;4;1;0;0;1;4;1;10;66;2;4;4;0;9;1;2;253;16;0;958;3;2;2;2;2

9;Arts, Animation, Stop-Motion;31;1;0;0;0;0;0;1;8;1;0;0;0;3;0;1;6;44;1;617;2;2;2;2;2

10;Arts, Animation, Stop-Motion;6;1;23;1;0;0;0;0;0;0;0;0;0;0;0;0;27;0;59;1;2;3;3;3

11;Arts, Animation, Stop-Motion;9;1;58;2;0;0;0;2;2;0;0;0;0;4;1;1;7;27;1;119;2;3;3;2;3

12;Arts, Animation, Training;6;0;0;0;8;40;0;0;0;0;0;0;0;1;0;0;34;28;0;63;0;1;1;1;1

13;Arts, Animation, Web;3;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;37;0;73;2;2;2;2;2

14;Arts, Animation, Web;10;0;102;0;0;0;0;0;0;0;0;0;7;0;0;1;28;0;16;0;2;2;2;2

15;Arts, Animation, Writers;8;0;8;2;0;0;0;0;0;0;0;0;13;0;0;4;31;0;179;2;2;2;2;2

16;Arts, Animation, Writers;7;1;27;1;0;0;0;0;0;0;0;0;0;0;0;0;32;0;95;1;2;3;3;3

17;Arts, Architecture, Building, Types;2;1;21;1;0;0;0;0;0;0;0;0;4;0;2;7;36;1;172;5;2;3;3;3

18;Arts, Architecture, History;1;0;22;3;0;0;0;0;0;0;0;0;17;4;2;5;23;0;93;3;2;2;2;2

19;Arts, Architecture, History, Materials;6;0;0;0;1;17;0;1;15;0;1;8;0;1;0;0;3;30;0;400;2;2;2;2;2

20;Arts, Architecture, History, Organizations;3;0;5;0;1;1;0;2;15;0;14;33;0;5;0;3;16;27;0;227;3;1;1;1;1

21;Arts, Architecture, History, Vernacular;6;2;0;0;1;10;1;0;0;0;1;3;0;0;0;0;0;64;1;650;4;2;2;2;2

22;Arts, Architecture, Landscape;9;0;26;0;0;0;0;3;7;0;0;0;0;11;0;0;17;32;0;152;2;2;2;2;2

23;Arts, Architecture, Landscape;7;2;0;0;4;8;2;2;5;0;6;13;0;5;0;1;35;28;0;239;3;1;1;2;1

24;Arts, Architecture, Preservation;3;1;0;0;0;0;0;0;0;0;0;0;0;0;0;1;32;31;1;178;2;1;1;1;1

25;Arts, Awards, Golden, Globe, Awards;10;5;34;5;1;3;3;2;4;2;1;3;2;37;3;3;117;41;2;561;3;3;3;2;3

26;Arts, Bodyart, Piercing, Studios;8;2;23;1;1;0;0;14;46;1;5;10;0;40;3;5;228;29;0;1022;2;3;3;2;3

27;Arts, Bodyart, Schools, Instruction;5;1;20;2;0;0;0;0;0;0;0;0;0;0;0;0;0;32;0;0;0;2;2;2;2

28;Arts, Bodyart, Studios;3;0;12;0;0;0;0;0;0;0;0;0;0;0;0;0;0;35;0;0;0;1;1;2;1

29;Arts, Bodyart, Tattoo, Removal;5;2;35;3;1;4;2;0;0;0;7;18;0;15;1;1;100;87;2;430;5;2;3;3;3

30;Arts, Classical, Studies, Journals;3;1;0;0;1;0;0;13;19;2;0;0;0;10;1;3;17;23;0;463;2;2;2;2;2

31;Arts, Comics, Comic, Strips, Panels;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;2;1;33;0;78;2;1;1;1;1

32;Arts, Comics, Comic, Strips, Panels;3;1;33;3;0;0;0;0;0;0;0;0;0;0;0;0;13;29;0;204;4;2;2;2;2

33;Arts, Comics, Conventions;3;0;0;0;0;0;0;0;0;0;0;0;0;0;17;3;1;18;39;0;303;3;1;1;1;1

34;Arts, Comics, Creators;3;0;0;0;0;0;0;1;5;0;0;0;0;11;0;0;12;25;0;860;1;1;1;2;1

35;Arts, Comics, Fan, Pages;2;0;0;0;0;0;0;0;0;0;2;91;2;6;0;3;30;38;0;455;0;1;1;1;1

36;Arts, Comics, Fan, Pages;2;1;24;1;0;0;0;0;0;0;0;0;0;0;6;1;2;23;50;1;451;4;2;2;2;2

37;Arts, Comics, Manga, Fandom;2;0;33;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;32;0;156;0;1;2;1;1

38;Arts, Comics, Organizations;3;1;16;2;0;0;0;5;15;0;7;39;0;0;0;0;54;23;0;438;2;2;2;3;2

39;Arts, Comics, Organizations;17;1;34;1;1;9;1;0;0;0;0;0;0;3;0;1;0;49;0;183;1;2;2;2;2

40;Arts, Comics, Organizations;5;1;0;0;0;0;0;21;205;2;0;0;0;15;3;2;215;18;0;2115;2;2;2;2;2

41;Arts, Comics, Publishers;3;1;0;0;14;98;1;0;0;0;0;0;0;34;0;2;33;26;0;182;2;2;2;2;2

42;Arts, Costumes, Exhibitions;11;0;0;0;1;7;0;0;0;0;4;18;0;1;0;0;68;28;1;377;0;2;2;2;2

43;Arts, Costumes, Exhibitions;4;0;0;0;0;0;0;0;0;0;0;0;0;3;0;4;7;34;0;313;5;2;1;1;1

44;Arts, Crafts, Candles;4;1;26;2;1;3;1;3;33;2;9;26;1;0;0;3;154;57;1;961;5;3;3;3;3

45;Arts, Crafts, Candles;4;1;0;0;0;0;0;1;7;1;0;0;0;15;1;2;18;33;0;757;2;2;1;1;1

46;Arts, Crafts, Mosaics, Glass;2;1;0;0;1;4;1;0;0;0;1;5;0;1;1;2;1;30;1;182;4;2;2;2;2

47;Arts, Crafts, Textiles, Dyeing;2;1;0;0;1;3;1;0;0;0;5;14;1;2;0;0;0;50;1;1263;3;2;1;2;2

48;Arts, Crafts, Woodcraft, Carving;2;1;0;0;0;0;0;17;73;2;4;10;0;36;4;5;117;35;1;754;5;2;2;2;2

49;Arts, Digital, Magazines, E-zines;8;1;14;1;1;1;0;0;0;0;0;0;0;0;0;3;47;23;1;164;4;2;2;1;2

50;Arts, Education, Educators;3;0;0;0;0;0;0;0;0;0;0;0;0;0;0;5;29;0;631;0;1;1;1;1

51;Arts, Education, Educators;2;0;0;0;0;0;0;1;1;0;0;0;0;4;0;1;19;25;0;122;8;2;1;1;1

52;Arts, Education, Schools, Academies;4;2;18;3;1;0;0;1;5;0;4;20;0;26;4;3;6;23;0;424;5;3;2;3;3

53;Arts, Education, Schools, Academies;11;1;48;5;0;0;0;4;34;1;0;0;0;15;2;3;23;35;1;325;6;3;2;3;3

54;Arts, Education, Schools, Academies;8;1;24;0;2;11;0;1;4;1;1;3;0;0;0;0;51;21;0;242;3;2;2;2;2

55;Arts, Humanities, Literature, Art;3;0;0;0;1;0;0;0;0;0;0;0;0;5;0;0;0;30;0;3;0;1;1;1;1

56;Arts, Humanities, Scholarship, Technology;1;0;0;0;1;1;0;1;2;0;0;0;0;9;0;2;4;37;0;109;3;1;1;1;1

57;Arts, Illustration, Airbrush;11;1;24;1;2;3;1;7;26;3;0;0;0;2;0;4;61;26;1;397;4;3;3;3;3

58;Arts, Illustration, Airbrush;6;1;0;0;0;0;0;0;0;0;0;0;0;0;0;2;2;34;0;96;3;1;1;1;1

59;Arts, Music, Concerts, Events, Folk;4;1;26;2;0;0;0;0;0;0;0;0;0;0;0;1;0;32;1;50;2;2;3;3;3

60;Arts, Performing, Arts, Acting, Actors, Actresses, Galvin, Anna;5;1;0;0;0;0;0;0;0;0;0;0;0;9;1;2;7;45;2;131;8;2;1;1;1

61;Arts, Photography, Contests;7;1;0;0;5;13;1;0;0;0;0;0;0;3;1;2;46;33;0;442;2;2;2;1;2

62;Arts, Photography, Education;8;1;21;1;2;11;1;6;34;2;0;0;0;14;1;4;10;22;0;905;3;2;3;3;3

63;Arts, Photography, Techniques, Styles, Polaroid, Processes;12;1;17;2;1;6;1;1;2;0;0;0;0;10;3;5;38;48;2;397;4;3;2;3;3

64;Business, Accounting, Associations;1;0;20;3;0;0;0;1;1;0;4;4;0;3;0;1;38;20;0;400;4;2;2;2;2

65;Business, Accounting, Associations;28;1;43;1;0;0;0;0;0;0;1;7;0;16;1;2;14;21;0;640;3;3;2;3;3

66;Business, Accounting, Associations;5;2;7;2;0;0;0;0;0;0;3;6;0;11;0;2;26;19;0;335;1;2;3;3;3

67;Business, Automotive, Car, Rentals;5;2;53;4;1;2;0;1;4;2;4;14;3;17;2;3;18;24;0;556;4;3;3;3;3

68;Business, Automotive, Car, Rentals;8;3;27;2;1;7;2;2;4;0;1;11;2;2;2;2;3;28;1;255;2;2;3;3;3

69;Business, Automotive, Car, Rentals;11;1;21;1;0;0;0;1;5;1;7;22;1;26;1;0;18;36;1;368;2;2;3;3;3

70;Business, Automotive, Import, Export;2;0;22;0;1;2;0;0;0;0;4;6;0;25;0;1;51;28;0;140;2;2;2;2;2

71;Business, Automotive, Import, Export;4;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;1;28;0;185;1;2;1;1;1

72;Business, Automotive, Import, Export;13;1;18;3;1;2;0;0;0;0;0;0;0;7;0;0;15;31;1;466;1;2;2;2;2

73;Business, Automotive, Marketing,
Advertising;3;1;0;0;1;6;1;4;12;0;5;23;0;1;0;2;19;34;1;321;3;1;2;2;2

74;Business, Automotive, Marketing,
Advertising;4;2;17;3;3;9;2;3;14;0;0;0;0;4;2;0;34;25;0;235;2;2;3;3;3

75;Business, Automotive, Motorcycles;8;0;15;0;1;2;0;0;0;0;19;62;0;7;0;0;68;34;0;699;0;1;1;2;1

76;Business, Automotive, Snow, Clearing;1;0;19;1;0;0;0;0;0;0;0;0;0;22;0;0;0;26;0;535;1;1;2;2;2

77;Business, Automotive, Snow, Clearing;8;1;23;1;0;0;0;1;5;0;4;11;1;51;2;1;110;24;0;659;0;3;3;3;3

78;Business, Automotive, Snow, Clearing;8;2;24;2;0;0;0;10;43;0;1;0;0;42;1;0;29;23;1;1333;1;3;3;2;3

79;Business, Automotive, Software;15;0;0;0;0;0;0;3;7;0;7;24;0;52;0;0;32;26;1;405;0;1;1;1;1

80;Business, Automotive, Software;10;0;0;0;0;11;56;0;1;8;0;7;12;0;2;0;0;735;28;0;7160;0;1;2;1;1

81;Business, Automotive, Software;9;2;21;2;5;32;1;5;11;1;1;1;0;7;0;0;99;22;0;789;3;3;3;2;3

82;Business, Automotive, Software;31;1;31;2;1;8;1;1;6;0;0;0;0;3;2;0;2;26;0;231;2;2;3;3;3

83;Business, Biotechnology, Pharmaceuticals;7;0;37;0;0;0;0;0;0;0;0;0;0;1;0;0;19;24;0;2354;0;1;2;1;1

84;Business, Biotechnology, Pharmaceuticals;10;0;0;0;0;0;0;7;51;0;3;22;0;3;1;0;54;17;0;489;0;1;1;2;1

85;Business, Biotechnology, Pharmaceuticals;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;26;0;0;0;1;1;1;1

86;Business, Biotechnology,
Pharmaceuticals;16;1;0;0;24;48;1;0;0;0;0;0;0;8;1;1;58;19;0;1031;1;1;1;2;1

87;Business, Chemicals, Associations;6;0;0;0;1;6;0;1;2;0;11;3;0;9;0;2;166;49;0;732;2;2;1;1;1

88;Business, Chemicals, Associations;5;2;25;2;0;0;0;1;4;2;2;11;2;0;0;0;0;25;0;203;3;3;2;2;2

89;Business, Construction, Maintenance, Consulting, General,
Consulting;8;0;23;0;0;0;0;0;0;0;0;0;0;0;0;0;34;0;11;0;1;1;2;1

90;Business, Construction, Maintenance, Consulting, General,
Consulting;2;0;7;0;1;0;0;56;107;0;87;223;0;243;0;0;158;33;0;5852;1;2;2;2;2

91;Business, Construction, Maintenance, Consulting, General,
Consulting;7;3;23;5;5;11;0;11;20;0;9;29;2;9;0;4;72;31;2;851;4;3;2;3;3

92;Business, Construction, Maintenance, Consulting, Project,
Management;4;0;0;0;1;6;0;3;17;0;0;0;0;1;0;0;5;35;1;66;0;1;1;2;1

93;Business, Construction, Maintenance, Consulting, Project,
Management;4;1;0;0;0;0;0;0;0;0;0;0;0;12;1;0;0;29;1;156;6;2;1;2;2

94;Business, Construction, Maintenance, Consulting, Project,
Management;7;1;63;1;0;0;0;7;16;3;1;6;1;8;1;1;3;22;0;242;4;3;3;2;3

95;Business, Construction, Maintenance, Consulting,
Safety;5;1;0;0;0;0;0;2;16;0;0;0;0;39;4;1;12;26;0;460;4;2;2;2;2

96;Business, Construction, Maintenance, Consulting,
Safety;8;2;0;0;0;0;0;0;0;0;0;0;0;1;2;1;16;29;0;177;3;2;2;2;2

97;Business, Construction, Maintenance,
Environmental;4;0;18;0;0;0;0;3;7;0;7;22;0;17;0;0;2;34;0;235;0;2;1;1;1

98;Business, Construction, Maintenance, Facilities,
Management;3;0;74;2;0;0;0;0;0;0;0;0;0;11;0;0;5;26;0;156;1;2;2;2;2

99;Business, Construction, Maintenance, Facilities,
Management;7;0;24;0;0;0;0;5;12;0;4;8;0;38;0;0;50;22;0;550;1;2;2;2;2

100;Business, Construction, Maintenance, Facilities,
Management;5;2;0;0;1;3;2;1;5;0;0;0;0;0;0;3;22;31;0;100;2;2;2;2;2

101;Business, Construction, Maintenance, Facilities, Management;2;0;43;0;1;6;0;7;16;2;0;0;0;2;0;0;18;25;0;594;0;2;2;2;2

102;Business, Construction, Maintenance, Historic, Preservation;4;0;0;0;1;3;0;2;2;0;1;7;0;3;1;1;27;35;0;209;3;1;2;2;2

103;Business, Construction, Maintenance, Landscaping, Materials, Supplies;3;0;11;1;9;39;1;2;297;2;2;3;0;1;0;0;105;55;0;782;0;2;2;2;2

104;Business, Construction, Maintenance, Restoration;22;1;0;0;0;0;0;0;0;0;0;0;0;44;3;0;0;29;0;1079;1;1;2;2;2

105;Business, Construction, Maintenance, Restoration;5;1;0;0;1;4;1;0;0;0;4;7;0;0;0;2;52;21;0;495;3;1;1;2;1

106;Business, Construction, Maintenance, Restoration;5;1;18;1;0;0;0;1;13;0;4;17;1;0;0;0;18;32;0;289;1;2;2;2;2

107;Business, Construction, Maintenance, Specifications;7;1;19;1;0;0;0;1;5;0;1;2;0;4;1;1;17;32;1;157;1;2;2;2;2

108;Business, Construction, Maintenance, Specifications;9;0;22;0;0;0;0;0;0;0;4;19;2;1;0;1;41;20;0;348;4;2;2;3;2

109;Business, Construction, Maintenance, Tools, Equipment;5;0;0;0;0;0;0;0;0;0;0;0;0;7;1;0;1;25;0;94;0;1;1;2;1

110;Business, Construction, Maintenance, Tools, Equipment;16;2;41;2;1;4;0;0;0;0;4;6;0;4;0;3;4;31;0;296;3;2;2;2;2

111;Business, Construction, Maintenance, Tools, Equipment;7;1;25;1;0;0;0;8;54;1;9;99;1;9;0;0;50;43;0;802;1;3;3;3;3

112;Business, Education, Training, Manual, Skills;2;0;0;0;0;0;0;2;2;0;0;0;0;0;0;0;0;0;45;0;139;1;1;1;1;1

113;Business, Education, Training, Manual, Skills;8;1;11;1;0;0;0;9;43;1;0;0;0;2;0;1;1;26;0;528;3;2;2;2;2

114;Business, Education, Training, Manual, Skills;5;1;60;2;0;0;0;1;5;0;0;0;0;13;1;1;4;28;0;344;1;3;3;2;3

115;Business, Education, Training, Manual, Skills;16;1;0;0;7;36;0;4;18;0;0;0;0;3;0;2;8;29;0;420;1;2;2;2;2

116;Business, Electronics, Electrical, Engineering, Product, Development;4;0;23;1;0;0;0;0;0;0;0;0;0;8;1;1;39;31;0;189;1;2;2;2;2

117;Business, Electronics, Electrical, Engineering, Product, Development;10;3;13;3;0;0;0;0;0;0;0;0;0;5;2;0;0;30;0;108;11;2;2;2;2

118;Business, Electronics, Electrical, Engineering, Product, Development;11;1;9;1;0;0;0;5;15;1;1;18;0;5;2;3;45;33;1;419;5;3;2;3;3

119;Business, Electronics, Electrical, Heating, Cooling;2;0;39;0;1;1;0;6;9;0;1;1;0;0;0;0;16;22;0;231;5;2;2;2;2

120;Business, Electronics, Electrical, Heating, Cooling;9;3;17;3;0;0;0;0;0;0;0;0;0;16;2;2;1;32;1;366;4;3;2;3;3

121;Business, Electronics, Electrical, Surge, Protectors;4;2;0;0;1;13;2;0;0;0;0;0;0;2;1;0;2;25;1;53;5;2;2;2;2

122;Business, Financial, Services, Banking, Services;12;0;41;0;1;1;0;18;52;0;0;0;0;4;0;1;16;26;0;992;0;2;2;2;2

123;Business, Financial, Services, Banking, Services;2;0;0;0;1;3;0;0;0;0;0;0;0;10;2;2;12;25;0;301;6;2;2;2;2

124;Business, Financial, Services, Loans;3;0;24;0;0;0;0;1;10;0;0;0;0;7;0;2;30;30;0;143;2;1;2;2;2

125;Business, Financial, Services, Loans;7;1;22;2;1;5;1;13;52;2;14;42;2;26;2;3;37;29;1;743;2;3;3;3;3

126;Business, Financial, Services, Loans;10;2;0;0;0;0;0;0;0;0;3;8;3;9;0;4;5;34;0;594;5;2;2;2;2

127;Business, Healthcare, Nursing;6;0;22;0;1;7;0;7;18;0;6;11;0;10;0;2;65;35;0;653;2;3;2;2;2

128;Business, Healthcare, Products, Services,
Nutrition;4;0;0;0;1;0;0;6;14;0;9;55;0;1;1;1;64;28;0;401;0;1;2;1;1

129;Business, Healthcare, Products, Services,
Nutrition;5;0;18;0;1;5;0;0;0;0;0;0;0;1;0;1;76;26;0;259;0;2;2;2;2

130;Business, Healthcare, Products, Services,
Nutrition;11;1;0;0;0;0;0;7;21;1;3;11;0;19;1;0;52;24;0;338;0;1;2;2;2

131;Business, Small, Business, Start, Up,
Entrepreneurship;10;3;0;0;0;0;0;0;0;0;0;0;0;4;3;8;83;38;1;417;6;2;2;2;2

132;Business, Small, Business, Start-Up,
Entrepreneurship;9;2;17;0;1;5;0;1;6;0;0;0;0;22;2;2;210;85;3;459;2;3;3;3;3

133;Computers, Artificial, Intelligence, Neural,
Networks;5;1;0;0;0;0;0;1;5;1;0;0;0;0;0;4;5;49;0;217;3;1;2;2;2

134;Computers, Computer, Science, Database,
Theory;2;0;0;0;1;3;0;0;0;0;0;0;0;2;0;0;10;38;0;277;0;1;1;1;1

135;Computers, Computer, Science, Database,
Theory;3;0;0;0;1;8;0;0;0;0;0;0;0;0;0;0;14;21;0;281;1;1;1;2;1

136;Computers, Computer, Science,
Organizations;3;0;17;3;0;0;0;2;4;0;10;35;0;10;0;0;11;35;0;403;4;2;2;2;2

137;Computers, Education, Certification,
Cisco;4;1;12;2;1;2;0;1;2;0;7;23;2;3;1;4;30;51;1;203;4;3;2;3;3

138;Computers, Education, Commercial,
Services;10;1;14;1;1;0;0;3;9;0;0;0;0;6;1;1;280;40;1;927;1;2;2;2;2

139;Computers, Education, Courses;3;0;0;0;1;4;0;0;0;0;0;0;0;2;0;1;2;30;0;315;2;1;1;2;1

140;Computers, Emulators, Amiga;3;1;0;0;4;19;1;1;0;0;5;6;0;2;1;2;29;23;0;972;3;1;2;2;2

141;Computers, Emulators, Thomson;2;0;0;0;1;2;0;1;4;0;1;1;0;1;0;1;4;42;0;190;5;2;2;2;2

142;Computers, Graphics, Fonts;5;1;31;1;2;5;1;1;0;0;0;0;0;6;1;2;33;28;1;253;2;3;3;2;3

143;Computers, Graphics, Textures;11;1;0;0;0;0;0;0;0;0;0;0;0;17;1;2;1;30;1;237;2;2;2;1;2

144;Computers, Hacking, Cryptography;1;0;0;0;1;1;0;6;19;0;0;0;0;1;0;0;9;44;1;692;0;1;1;2;1

145;Computers, Hacking, Cryptography;1;0;0;0;1;3;0;1;1;0;2;11;0;0;0;2;0;42;0;1513;1;1;1;2;1

146;Computers, Home, Automation, Products,
Manufacturers;10;2;25;2;1;51;2;1;8;2;13;0;0;18;3;2;13;32;0;127;2;3;3;2;3

147;Computers, Home, Automation, Products,
Manufacturers;6;1;20;2;1;0;0;9;34;1;0;0;0;10;0;2;38;29;0;205;2;3;3;3;3

148;Computers, Home, Automation, Products,
Manufacturers;8;2;21;2;0;0;0;3;15;0;4;8;1;4;2;4;18;23;0;407;5;3;3;3;3

149;Computers, Internet, Abuse;2;0;0;0;3;7;0;1;2;0;0;0;0;0;0;0;0;28;0;601;1;1;1;2;1

150;Computers, Internet, Cybercafes, Australia;21;1;0;0;0;0;0;0;0;0;0;0;0;2;1;0;1;27;0;0;0;1;2;2;2

151;Computers, Internet, Domain, Names,
Registrars;2;0;8;0;0;0;0;3;7;0;10;29;0;24;0;1;41;21;0;239;2;2;2;2;2

152;Computers, Internet, Domain, Names,
Registrars;11;1;24;4;1;5;0;2;10;0;2;4;0;9;0;3;15;23;0;212;3;2;3;3;3

153;Computers, Internet, Telephony;7;0;32;0;0;0;0;2;6;0;3;5;0;9;0;0;37;36;0;1131;0;3;3;3;3

154;Computers, Mobile, Computing, Wireless,
Data;8;0;26;1;1;10;0;0;0;0;7;21;0;23;0;0;19;29;0;261;0;2;2;3;2

155;Computers, Mobile, Computing, Wireless,
Data;12;0;14;0;0;0;0;1;3;0;0;0;0;1;0;0;0;22;0;147;0;2;2;2;2

156;Computers, Open, Source, Software;1;0;0;0;7;22;1;0;0;0;0;0;0;2;3;0;8;26;0;494;1;1;1;1;1

157;Computers, Organizations, Committees;9;1;0;0;0;0;0;7;23;0;0;0;0;2;0;4;22;22;0;300;1;2;1;2;2

158;Computers, Organizations, Non-Profit;1;0;0;0;7;12;0;13;35;1;1;1;0;5;0;0;15;23;0;285;0;1;1;2;1

159;Computers, Organizations, Non-Profit;9;0;0;0;0;0;0;0;0;0;0;0;0;7;0;1;8;48;1;225;3;1;1;1;1

160;Computers, Organizations, Non-Profit;9;0;0;0;1;6;0;7;36;1;23;78;3;4;0;2;40;25;0;1614;2;1;1;2;1

161;Computers, Parallel, Computing, Programming,
Languages;10;3;8;3;1;12;2;4;9;1;2;5;0;1;0;3;2;55;0;497;9;3;2;3;3

162;Computers, Parallel, Computing, Projects;8;3;6;3;0;0;0;3;4;0;0;0;0;6;0;2;120;29;0;363;0;1;1;2;1

163;Computers, Programming, Compilers;2;0;5;2;0;0;0;0;0;0;0;0;0;2;0;2;14;26;1;849;2;2;2;2;2

164;Computers, Programming, Contests;4;2;0;0;0;0;0;0;0;0;0;0;0;46;0;0;3;19;0;322;8;2;1;1;1

165;Computers, Programming, Education;17;1;0;0;1;46;1;15;56;0;6;184;1;2;0;0;18;23;0;611;2;2;2;2;2

166;Computers, Programming, Languages, Delphi,
Tools;10;1;47;1;1;9;0;0;0;0;0;0;0;0;0;1;42;0;526;3;3;3;3;3

167;Computers, Programming, Languages,
Haskell;2;1;8;2;0;0;0;5;10;0;6;9;0;1;0;1;79;25;1;1343;3;2;2;2;2

168;Computers, Programming, Languages, Ruby;2;1;0;0;0;0;0;4;4;0;0;0;0;2;0;0;2;24;0;367;3;2;2;2;2

169;Computers, Programming, Methodologies;5;0;34;0;1;2;0;1;5;0;0;0;0;1;0;1;2;25;0;331;2;2;2;2;2

170;Computers, Robotics, Robots;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;2;28;0;336;3;1;1;1;1

171;Computers, Robotics, Robots;9;2;2;2;0;0;0;0;0;0;0;0;0;5;2;2;0;29;2;260;4;2;2;2;2

172;Computers, Security, Authentication,
Kerberos;3;0;53;2;2;18;2;10;27;0;0;0;0;4;0;0;19;25;1;685;2;3;3;3;3

173;Computers, Security, Authentication,
RADIUS;9;1;5;1;1;1;0;13;22;0;2;2;0;1;0;0;5;46;1;1086;2;2;3;3;3

174;Computers, Security, Biometrics;2;0;23;1;0;0;0;9;21;0;3;4;0;8;0;1;124;24;0;299;2;2;2;2;2

175;Computers, Security, Consultants,
Training;2;0;0;0;1;2;0;2;6;1;7;15;2;11;0;4;78;27;1;799;2;2;2;2;2

176;Computers, Security, Firewalls;5;1;0;0;2;13;2;24;47;1;0;0;0;5;1;2;80;75;1;3865;3;2;3;2;2

177;Computers, Security, Malicious, Software, Trojan,
Horses;5;2;0;0;0;0;0;1;10;0;0;0;0;8;1;6;23;33;1;1571;7;2;2;2;2

178;Computers, Security, Malicious, Software,
Worms;6;0;19;1;1;4;0;10;51;1;40;158;3;47;3;2;22;45;0;2093;2;3;2;2;2

179;Computers, Security, Products, Tools,
Cryptography;8;1;22;1;1;11;1;7;32;0;28;221;3;2;0;4;239;25;0;1026;2;3;3;3;3

180;Computers, Software, Business, Drawing;12;0;33;0;0;0;0;0;0;0;0;0;0;19;2;1;5;24;0;440;2;3;2;2;2

181;Computers, Software, Business,
Drawing;8;1;19;1;1;7;0;10;53;0;8;35;0;24;1;0;65;29;0;1398;1;3;3;3;3

182;Computers, Supercomputing, Companies;13;2;28;1;1;7;1;0;0;0;0;0;0;1;0;1;3;20;0;82;1;3;3;3;3

183;Computers, Supercomputing,
Conferences;3;0;34;1;4;13;0;1;10;0;3;8;0;8;0;0;13;24;0;338;2;2;2;3;2

184;Computers, Supercomputing, Cray;7;2;24;2;1;0;0;4;13;0;6;22;0;17;0;3;52;21;1;699;2;2;3;3;3

185;Computers, Usenet, Newsgroup,
Hosting;13;0;26;0;1;19;1;0;0;0;0;0;0;16;1;1;13;43;0;713;2;2;2;2;2

186;Computers, Usenet, Software;4;0;0;0;0;0;0;13;56;0;4;5;0;22;0;0;143;30;0;3002;0;1;1;2;1

187;Computers, Usenet, Web, Based;6;0;0;0;7;14;0;2;13;0;0;0;0;1;0;1;7;28;0;199;1;2;2;2;2

188;Computers, Virtual, Reality, Hardware;3;0;0;0;0;0;0;8;20;0;4;6;0;26;0;0;134;24;0;216;0;1;1;2;1

189;Computers, Virtual, Reality, Hardware;2;0;0;0;1;4;0;2;6;0;4;57;0;0;0;0;25;30;0;289;0;1;1;2;1

190;Computers, Virtual, Reality, Hardware;9;0;17;2;0;0;0;1;5;0;4;5;0;4;0;0;1;20;0;242;2;2;2;2;2

191;Games, Board, Games, Fantasy, Small,
World;16;2;36;5;4;16;5;0;0;0;0;0;0;7;0;7;30;44;2;231;6;3;2;3;3

192;Games, Board, Games, Resources;8;3;40;3;1;4;3;2;13;3;5;31;3;1;3;5;6;29;3;345;7;3;3;3;3

193;Games, Board, Games, Science, Fiction, Car,
Wars;5;5;187;5;0;0;0;0;0;0;0;0;0;0;0;0;0;54;8;0;0;2;2;2;2

194;Games, Board, Games, War, Politics;4;0;0;0;0;0;0;0;0;0;0;0;0;12;0;2;2;30;1;244;4;1;1;1;1

195;Games, Card, Games, Trick, Capturing;5;4;30;3;1;5;4;0;0;0;0;0;0;0;0;3;6;28;0;249;7;2;2;2;2

196;Games, Card, Games, Trick, Capturing,
Bridge;1;0;0;0;2;12;3;10;32;1;3;13;3;1;0;0;4;27;1;347;5;1;1;1;1

197;Games, Dice, Dragon, Dice;2;0;13;4;0;0;0;0;0;0;0;0;0;4;5;3;27;24;0;84;2;1;2;2;2

198;Games, Dice, Yahtzee;11;0;13;2;1;0;0;1;1;0;0;0;0;4;1;0;9;27;1;781;3;3;3;2;3

199;Games, Gambling, Poker;10;2;25;2;2;26;2;28;105;1;6;20;1;6;1;2;7;27;1;792;2;2;2;2;2

200;Games, Gambling, Poker;9;1;16;1;1;3;0;3;10;1;4;23;2;0;0;3;17;35;1;620;1;2;3;3;3

201;Games, Gambling, Poker;8;1;25;1;1;1;0;1;2;0;1;1;1;0;0;1;33;36;1;979;1;2;2;2;2

202;Games, Game, Studies, Conferences;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;2;33;1;229;2;2;1;1;1

203;Games, Game, Studies, Education;5;0;37;3;1;3;0;10;19;0;28;159;0;10;2;2;61;35;0;3785;4;2;2;2;2

204;Games, Hand, Games, Thumb, Wrestling;5;0;0;0;4;2;0;7;15;0;0;0;0;0;0;2;5;48;0;499;7;2;1;1;1

205;Games, Hand-Eye, Coordination, Skittles;2;1;0;0;0;0;0;2;4;0;0;0;0;0;0;2;1;30;1;400;3;2;2;2;2

206;Games, Online, Roleplaying, Horror;7;0;21;0;0;0;0;0;0;0;0;0;0;6;0;0;10;43;0;219;0;1;1;2;1

207;Games, Paper, Pencil, Tic-Tac-Toe;4;0;24;0;1;5;0;2;16;0;0;0;0;2;0;3;18;55;0;1134;2;3;3;2;3

208;Games, Party, Games, Drinking, Games;2;0;22;4;1;3;3;0;0;0;0;0;0;6;4;4;6;30;4;142;6;3;3;2;3

209;Games, Play, Groups, Canada;2;0;0;0;1;1;0;2;3;0;0;0;0;3;0;1;4;33;0;185;4;2;2;2;2

210;Games, Puzzles, Brain, Teasers, Riddles;6;0;0;0;1;1;0;0;0;0;0;0;0;1;0;0;2;31;1;355;3;1;1;2;1

211;Games, Puzzles, Brain, Teasers, Sudoku;3;1;0;0;1;2;1;0;0;0;2;8;1;1;0;2;3;37;2;268;4;2;2;2;2

212;Games, Puzzles, Jigsaw, Manufacturers;8;3;24;3;0;0;0;5;0;0;3;5;0;5;4;2;352;25;1;262;2;2;3;3;3

213;Games, Puzzles, Mechanical, Wooden;11;2;0;0;2;0;0;1;0;0;2;8;1;10;2;5;5;24;0;254;4;1;2;2;2

214;Games, Roleplaying, Software;11;1;0;0;0;0;0;0;0;0;7;60;3;28;0;0;3;24;0;627;5;2;2;2;2

215;Games, Tile, Games, Dominoes;3;1;14;1;1;2;1;0;0;0;0;0;0;1;0;3;78;47;1;321;7;2;2;2;2

216;Games, Tile, Games, Quinto;6;2;42;4;1;4;0;1;3;0;0;0;0;3;2;0;4;24;3;48;2;2;3;3;3

217;Games, Trading, Card, Games, Pokmon;3;0;0;0;1;3;0;0;0;0;0;0;0;5;0;4;18;45;0;228;6;1;2;2;2

218;"Games, Video, Games, Action, Space, Combat,
Massive, Multiplayer, Online, EVE, Online";3;0;9;1;0;0;0;6;25;3;0;0;0;0;0;3;14;27;1;880;6;2;2;2;2

219;"Games, Video, Games, Action, Space, Combat,
Star, Trek, Games, Invasion, Cheats,
Hints";10;5;18;3;0;0;0;0;0;0;0;0;0;4;1;11;221;42;1;1011;13;2;2;2;2

220;Games, Video, Games, Educational;6;1;19;1;1;8;1;0;0;0;0;0;0;14;0;1;1;44;0;1188;3;2;2;2;2

221;"Games, Video, Games, Fighting, Final,
Fight, Series, Final, Fight, Streetwise";5;5;24;2;1;3;3;1;3;3;1;3;0;6;4;9;40;63;8;1040;8;2;3;3;3

222;Games, Video, Games, Simulation, Flight;3;2;0;0;1;3;2;0;0;0;0;0;0;13;1;0;1;39;0;497;3;2;1;1;1

223;Games, Video, Games, Strategy, Browser, Based;8;2;25;2;1;7;2;0;0;0;4;7;0;6;0;2;2;27;0;271;2;3;2;3;3

224;Games, Video, Games, Word, Games;10;1;0;0;1;3;0;5;15;2;0;0;0;0;0;2;7;24;1;352;2;1;1;1;1

225;Games, Yard, Deck, Table, Games, Table, Soccer, Foosball;4;0;11;0;0;0;0;0;0;0;1;4;0;0;0;0;1;27;1;134;0;2;2;2;2

226;Health, Addictions, Internet;8;2;22;2;1;63;2;18;78;2;18;87;2;18;2;3;92;30;1;1066;2;3;3;3;3

227;Health, Aging, Life, Expectancy;5;2;0;0;2;5;2;0;0;0;0;0;0;104;1;2;31;49;2;830;4;1;2;2;2

228;Health, Alternative, Acupuncture, Chinese, Medicine;32;4;39;4;0;0;0;0;0;0;0;0;0;0;4;38;28;1;546;4;2;2;2;2

229;Health, Alternative, Acupuncture, Chinese, Medicine;3;1;24;0;0;0;0;13;77;3;0;0;0;0;0;3;100;29;1;2806;4;2;2;2;2

230;Health, Alternative, Aromatherapy;7;1;0;0;1;7;1;0;0;0;0;0;0;0;0;0;0;29;0;13;0;1;1;1;1

231;Health, Alternative, Aromatherapy;5;1;18;1;1;17;1;12;45;1;7;40;1;44;1;1;129;25;0;801;1;3;3;3;3

232;Health, Alternative, Ayurveda;9;1;29;1;0;0;0;0;0;0;0;0;0;0;0;0;0;18;0;3;0;1;2;1;1

233;Health, Alternative, Chiropractic;2;0;0;0;1;0;0;7;16;0;2;5;0;9;0;0;67;23;0;275;2;2;2;2;2

234;Health, Alternative, Herbs;3;0;0;0;0;0;0;0;0;0;0;0;0;1;0;2;1;31;1;427;2;1;1;2;1

235;Health, Alternative, Herbs;6;0;27;0;0;0;0;2;8;0;4;11;0;8;0;0;290;34;0;1325;2;2;2;3;2

236;Health, Alternative, Herbs;9;1;6;0;0;0;0;0;0;0;0;0;0;1;0;0;2;32;0;177;2;1;1;3;1

237;Health, Alternative, Homeopathy;1;0;117;1;2;1;0;16;47;2;2;4;0;4;0;2;43;40;0;2132;2;2;2;2;2

238;Health, Alternative, Homeopathy;3;1;4;1;6;20;1;21;47;1;0;0;0;6;0;2;50;34;0;2064;1;1;2;2;2

239;Health, Alternative, Massage, Therapy, Bodywork;4;0;21;0;1;1;0;2;24;0;9;18;0;1;0;0;49;24;0;431;0;2;2;2;2

240;Health, Alternative, Massage, Therapy, Bodywork;7;1;23;1;1;1;0;1;5;0;0;0;0;7;1;3;72;27;0;110;2;3;3;3;3

241;Health, Alternative, Ozone, Therapy;10;1;0;0;0;0;0;0;0;0;0;0;0;7;0;1;0;30;1;360;1;1;2;2;2

242;Health, Alternative, Ozone, Therapy;9;0;0;0;0;0;0;0;0;0;0;0;0;0;0;3;68;35;0;100;3;2;2;2;2

243;Health, Alternative, Reiki;6;0;21;0;0;0;0;0;0;0;0;0;0;25;2;0;4;21;1;357;2;2;3;3;3

244;Health, Alternative, Reiki;3;0;0;0;3;8;0;22;106;0;0;0;0;12;0;0;4;31;1;4543;0;1;2;1;1

245;Health, Alternative, Reiki;3;0;0;0;0;0;0;0;0;0;0;0;0;0;0;11;41;1;695;0;1;1;1;1

246;Health, Alternative, Tibetan, Medicine;2;0;16;2;0;0;0;0;0;0;0;0;0;2;0;0;1;32;2;485;3;2;2;2;2

247;Health, Alternative, Tibetan, Medicine;3;2;0;0;0;0;0;0;0;0;0;0;0;2;1;3;49;49;2;938;2;1;1;1;1

248;Health, Beauty, Hair;3;1;0;0;0;0;0;0;0;0;0;0;0;8;1;2;1;45;1;187;3;2;1;1;1

249;Health, Beauty, Salons, Spas;3;1;3;1;0;0;0;0;0;0;0;0;0;6;2;1;1;34;1;172;2;2;1;1;1

250;Health, Beauty, Schools;14;1;24;1;4;38;1;2;46;1;0;0;0;6;0;1;64;30;0;384;1;3;2;2;2

251;Health, Beauty, Schools;10;1;16;1;1;3;0;7;48;3;0;0;0;5;2;3;27;31;1;615;3;3;3;3;3

252;Health, Beauty, Skin, Care;7;0;29;2;1;6;0;0;0;0;7;18;0;16;1;5;104;72;4;381;4;3;2;3;3

253;Health, Beauty, Skin, Care, Acne;12;1;0;0;2;10;0;11;35;3;6;20;2;6;3;3;92;51;2;1295;3;2;2;2;2

254;Health, Child, Health, Fitness;9;0;9;0;1;6;0;3;8;0;8;30;2;17;0;2;14;38;0;945;2;1;2;2;2

255;Health, Child, Health, Nutrition;7;2;0;0;1;2;0;11;16;2;25;72;3;12;1;3;19;63;2;404;3;2;2;2;2

256;Health, Child, Health, Nutrition;4;2;0;0;0;0;0;0;0;0;0;0;0;12;1;2;1;23;0;187;7;1;1;1;1

257;Health, Child, Health, Special, Needs;11;0;0;0;0;0;0;0;0;0;1;2;0;3;0;1;25;24;0;451;0;1;2;1;1

258;Health, Conditions, Diseases, Allergies;4;1;22;2;1;3;1;0;0;0;7;18;0;15;4;6;99;89;4;362;5;3;2;3;3

259;Health, Conditions, Diseases, Blood, Disorders, Anemia;5;0;13;0;1;3;0;13;38;1;10;33;0;16;0;3;27;78;1;1234;4;3;3;3;3

260;Health, Conditions, Diseases, Cancer;4;1;0;0;0;0;0;10;38;2;0;0;0;3;1;2;2;31;1;456;2;2;2;2;2

261;Health, Conditions, Diseases, Cancer;6;1;0;0;1;2;0;1;2;0;1;6;0;3;0;2;4;24;1;959;2;1;2;2;2

262;Health, Conditions, Diseases, Cardiovascular, Disorders;8;1;20;1;0;0;0;20;63;2;4;34;1;7;1;3;52;22;0;359;3;3;3;2;3

263;Health, Conditions, Diseases, Eye, Disorders, Blindness;5;1;0;0;0;0;0;5;36;0;1;7;0;2;1;0;77;45;1;474;1;1;1;1;1

264;Health, Conditions, Diseases, Sleep, Disorders;8;2;0;0;2;10;0;5;27;2;3;14;3;8;1;5;88;87;3;1040;3;2;2;2;2

265;Health, Fitness, Aerobics;7;2;27;2;0;0;0;1;4;1;3;13;0;2;1;0;18;24;0;156;3;2;3;3;3

266;Health, Fitness, Gyms;5;2;29;2;3;14;1;8;22;1;7;16;1;25;1;4;69;25;1;1828;4;2;3;3;3

267;Health, Fitness, Pilates, Method;7;1;12;1;1;2;1;8;40;2;1;5;1;2;2;2;60;25;1;1047;1;2;3;3;3

268;Health, Mental, Health, Policy, Advocacy;27;4;66;4;1;4;0;0;0;0;0;0;0;15;4;1;5;25;0;703;4;2;3;3;3

269;Health, Mental, Health, Stress;5;1;0;0;1;3;1;11;26;1;15;46;1;13;1;3;60;40;0;1087;2;2;2;2;2

270;Health, Mental, Health, Stress;10;1;9;1;1;8;1;0;0;0;3;8;1;4;0;1;19;43;1;827;3;2;3;3;3

271;Health, Nursing, Students;10;1;11;1;1;10;1;5;43;4;2;5;1;0;0;5;58;36;2;1977;3;3;2;3;3

272;Health, Pharmacy, Nuclear, Pharmacy;9;3;0;0;0;0;0;0;0;0;0;0;0;0;0;0;7;13;44;1;113;8;1;1;1;1

273;Health, Pharmacy, Schools, of, Pharmacy;3;2;0;0;1;5;2;2;10;2;0;0;0;14;5;5;69;25;1;496;7;2;2;2;2

274;Health, Public, Health, Safety, Emergency, Services;6;0;0;0;0;0;0;4;12;0;0;0;0;3;0;2;22;21;0;5309;4;1;2;1;1

275;Health, Public, Health, Safety, First, Aid;8;2;0;0;1;0;0;6;21;2;3;8;2;4;0;2;10;31;2;335;2;1;2;2;2

276;Health, Senior, Health, Assisted, Living;4;2;19;2;0;0;0;7;54;3;4;19;2;0;0;3;57;26;0;262;3;2;2;2;2

277;Health, Senior, Health, Drugs;3;1;0;0;1;0;0;0;0;0;2;3;0;9;0;7;47;31;1;153;4;1;1;1;1

278;Health, Senses, Smell, Taste;12;2;26;2;0;0;0;1;3;1;3;7;0;13;2;3;7;30;2;382;5;2;3;3;3

279;Health, Teen, Health, Drugs, Alcohol;5;1;17;0;5;22;2;1;9;0;5;47;3;5;1;3;111;22;1;784;3;2;2;2;2

280;Health, Teen, Health, Teen, Pregnancy;9;3;26;3;3;12;3;10;34;1;0;0;0;29;3;3;20;35;3;689;3;3;3;2;3

281;Health, Weight, Loss, Clinics, Services;4;2;0;0;0;0;0;1;4;1;0;0;0;3;2;3;1;39;2;157;2;1;2;2;2

282;Health, Women's, Health, Organizations;4;0;0;0;0;0;0;11;51;0;10;55;2;16;0;2;240;26;0;881;2;2;2;2;2

283;Health, Women's, Health, Organizations;38;2;86;2;0;0;0;0;0;0;0;0;0;4;2;2;28;21;0;353;4;2;2;2;2

284;Health, Women's, Health, Smoking;6;2;0;0;0;0;0;9;21;0;3;21;0;5;0;0;41;53;0;870;2;1;2;2;2

285;Home, Apartment, Living, Roommates;11;1;0;0;1;5;1;4;22;2;0;0;0;0;0;3;0;30;1;268;3;2;2;3;2

286;Home, Apartment, Living, Roommates;10;1;30;1;1;5;1;10;35;1;5;18;1;1;1;2;0;29;0;945;2;2;2;2;2

287;Home, Consumer, Information, Complaints;2;1;0;0;0;0;0;4;13;2;6;16;3;1;0;3;18;44;2;269;4;2;2;2;2

288;Home, Cooking, Breakfast;3;1;10;2;1;3;1;13;38;0;7;25;0;25;1;1;135;35;1;1437;4;3;2;3;3

289;Home, Cooking, Breakfast;4;1;18;1;1;2;0;0;0;0;5;8;0;7;0;1;6;51;1;236;0;3;2;2;2

290;Home, Cooking, Chocolate;3;1;23;1;1;1;1;14;42;1;8;35;2;15;2;1;11;55;1;1018;2;2;3;2;2

291;Home, Cooking, Desserts;7;1;0;0;1;2;0;9;55;0;3;7;1;9;0;3;199;32;0;599;2;1;2;2;2

292;Home, Cooking, Desserts;4;1;0;0;0;0;0;25;55;0;0;0;0;11;2;3;79;27;1;3297;1;2;2;2;2

293;Home, Cooking, Desserts;10;1;59;2;1;2;1;0;0;0;0;0;0;25;0;1;127;53;1;1237;2;3;2;2;2

294;Home, Cooking, Desserts;7;1;18;1;1;2;1;21;136;2;0;0;0;34;1;3;83;30;1;244;2;3;3;3;3

295;Home, Cooking, Eggs;5;0;6;0;0;0;0;0;0;0;0;0;0;1;0;1;1;70;0;158;5;2;2;3;2

296;Home, Cooking, Fish, Seafood, Bluefish;4;0;28;0;1;4;0;2;7;0;2;8;0;1;0;4;99;55;0;644;3;2;2;3;2

297;Home, Cooking, For, Children;14;0;28;0;0;0;0;0;0;0;0;0;0;14;1;1;4;49;0;281;0;2;2;2;2

298;Home, Cooking, For, Children;7;0;0;0;0;0;0;1;4;0;1;3;0;18;1;1;5;56;0;147;1;1;1;1;1

299;Home, Cooking, For, Children;5;1;0;0;1;6;1;1;5;0;0;0;0;1;0;5;47;42;0;1273;3;2;2;2;2

300;Home, Family, Grandparents;6;0;8;0;0;0;0;0;0;0;0;0;0;26;1;1;1;47;1;450;1;2;3;3;3

301;Home, Family, Grandparents;2;0;13;0;3;15;0;0;0;0;3;12;1;11;1;2;4;28;0;1969;3;2;2;2;2

302;Home, Family, Grandparents;8;2;0;0;0;0;0;0;0;0;0;0;0;0;1;10;23;0;250;2;2;2;2;2

303;Home, Family, Pregnancy;3;0;0;0;1;3;0;4;13;0;1;1;0;5;0;0;7;24;0;547;0;1;2;1;1

304;Home, Family, Pregnancy;2;1;7;1;1;2;1;0;0;0;0;0;0;4;1;1;373;31;0;2954;0;1;2;2;2

305;Home, Family, Pregnancy;6;1;13;1;1;6;0;1;10;0;5;28;1;3;0;2;7;34;0;311;1;2;3;3;3

306;Home, Family, Pregnancy;6;1;3;1;1;2;1;10;51;1;8;15;1;28;1;2;252;26;1;793;1;2;2;2;2

307;Home, Gardening, Bonsai, Suiseki;2;0;32;2;1;22;0;0;0;0;0;0;0;1;0;2;1;30;1;70;1;1;2;2;2

308;Home, Gardening, Landscaping;6;1;24;1;1;1;0;2;27;1;4;73;3;2;1;4;27;34;1;2196;3;3;2;3;3

309;Home, Gardening, Landscaping;6;2;0;0;4;9;3;0;0;0;0;0;0;1;0;6;3;34;1;645;4;2;2;2;2

310;Home, Gardening, Organic;4;0;0;0;1;3;1;10;37;3;8;34;2;2;1;5;11;33;2;531;4;2;2;2;2

311;Home, Gardening, Organic;9;3;34;2;2;3;2;2;12;3;7;16;2;8;3;4;246;54;2;1175;2;2;3;3;3

312;Home, Gardening, Plants, Herbs;12;3;27;2;5;19;3;3;7;0;1;2;0;2;0;3;16;83;3;1166;5;2;2;2;2

313;Home, Home, Improvement,
Bathrooms;5;1;20;1;1;4;1;1;6;1;10;70;1;10;1;3;107;65;1;979;3;3;2;3;3

314;Home, Home, Improvement, Energy,
Efficiency;3;3;27;4;9;32;3;20;117;4;9;34;2;29;4;8;35;23;2;892;4;2;3;3;3

315;Home, Home, Improvement, Floors;13;1;0;0;1;6;3;0;0;0;0;0;0;0;0;0;0;25;1;167;5;2;2;2;2

316;Home, Home, Improvement, Furniture;6;1;14;1;0;0;0;0;0;0;0;0;0;1;0;1;3;34;1;236;3;2;3;3;3

317;Home, Home, Improvement,
Storage;23;3;176;3;1;2;1;0;0;0;27;144;3;0;0;6;110;39;1;794;5;2;2;2;2

318;Home, Homemaking, Christian;8;3;6;1;1;2;2;15;156;0;0;0;0;34;3;3;132;37;2;1798;3;2;2;2;2

319;Home, Homemaking, Cleaning, Stains;7;1;103;3;1;4;1;10;18;0;3;5;0;2;1;3;19;49;0;244;3;3;2;3;3

320;Home, Homemaking, Frugality;9;1;0;0;0;0;0;1;3;0;0;0;0;4;1;2;5;29;1;959;2;1;2;2;2

321;Home, Homeowners, Treehouses;9;0;0;0;1;4;0;2;7;0;0;0;0;4;1;1;6;34;1;424;0;1;2;1;1

322;Home, Moving, Relocating, Moving;9;4;7;2;1;2;2;6;38;4;1;7;2;6;4;4;19;30;0;116;6;2;3;3;3

323;Home, Moving, Relocating, Moving;3;0;9;2;0;0;0;0;0;0;0;0;0;18;0;3;2;29;2;109;7;2;2;2;2

324;Home, Moving, Relocating, Moving;3;2;8;2;0;0;0;10;42;5;5;7;0;16;2;9;131;33;0;3103;7;2;2;2;2

325;Home, Moving, Relocating, Publications;3;0;0;0;3;22;1;4;13;2;0;0;0;12;4;5;2;26;0;381;5;1;2;2;2

326;Home, Personal, Finance, Investing;12;3;24;2;0;0;0;0;0;0;0;0;0;1;0;0;5;30;1;397;4;3;3;2;3

327;Home, Rural, Living, Homesteading;9;2;18;3;2;11;2;3;21;2;1;2;0;4;2;7;62;29;1;2517;5;3;3;3;3

328;Kids, Teens, Computers, Education;7;3;0;0;1;5;2;0;0;0;0;0;0;2;0;1;10;36;1;130;4;2;2;2;2

329;Kids, Teens, Entertainment, Bands,
Artists;5;0;3;0;1;2;0;3;11;0;6;19;0;1;0;0;24;21;0;324;0;1;2;1;1

330;Kids, Teens, Entertainment, Comics,
Superman;15;1;0;0;0;0;0;0;0;0;5;38;1;21;1;3;333;34;1;913;3;2;2;2;2

331;Kids, Teens, Entertainment, Magazines, E-
zines;2;1;5;0;1;0;0;1;4;0;26;154;0;0;0;0;7;21;0;785;2;1;1;2;1

332;Kids, Teens, Entertainment, Movies;8;2;16;3;1;6;2;22;170;5;1;3;0;52;2;4;98;46;3;1345;2;2;2;2;2

333;Kids, Teens, Entertainment, Museums, Arts;5;2;0;0;0;0;0;0;0;0;0;0;0;8;0;5;68;33;2;345;3;2;2;1;2

334;Kids, Teens, Games, Puzzles, Mazes;2;2;20;5;1;2;2;0;0;0;0;0;0;8;2;7;40;37;1;162;7;2;3;2;2

335;Kids, Teens, Health, Dental;2;0;20;1;1;2;0;3;7;0;6;21;0;2;0;3;9;48;3;1448;1;1;2;2;2

336;Kids, Teens, Health, Dental;5;2;24;2;0;0;0;3;9;0;6;33;1;9;2;4;22;35;2;288;5;3;2;3;3

337;Kids, Teens, Health, Nutrition;3;0;10;0;1;3;0;3;7;0;5;13;0;2;0;5;9;45;3;662;1;2;2;2;2

338;Kids, Teens, Pre-School, Drawing,
Coloring;9;2;23;3;1;5;2;1;17;3;0;0;0;0;0;1;3;29;1;240;5;2;3;3;3

339;Kids, Teens, Pre-School, Food;6;1;11;1;0;0;0;0;0;0;0;0;0;16;1;0;9;37;0;698;1;1;2;2;2

340;Kids, Teens, Pre-School, Games;2;0;9;1;0;0;0;1;1;0;0;0;0;5;0;0;0;34;0;450;2;2;2;2;2

341;Kids, Teens, Pre-School, Online, Stories;9;0;20;1;3;7;0;2;3;1;5;10;0;14;0;0;82;29;0;428;1;3;2;2;2

342;Kids, Teens, School, Time, Homework,
Help;1;0;0;0;1;8;0;5;23;0;1;6;0;15;0;0;7;28;0;148;0;1;1;2;1

343;Kids, Teens, School, Time, Homework,
Help;4;1;3;1;1;0;0;1;1;0;0;0;0;1;0;2;12;29;0;248;2;1;2;2;2

344;Kids, Teens, School, Time, Learning,
Games;14;0;0;0;1;4;0;1;8;1;0;0;0;1;0;1;25;26;0;363;3;2;1;2;2

345;Kids, Teens, School, Time, Social, Studies,
Geography;6;0;0;0;1;3;0;4;24;1;2;10;1;10;0;2;64;40;1;764;0;1;2;2;2

346;Kids, Teens, School, Time, Social, Studies,
History;8;2;23;1;2;11;2;5;48;2;14;41;1;1;0;3;102;29;1;708;1;3;3;3;3

347;Kids, Teens, Sports, Hobbies, Drawing,
Coloring;6;2;27;2;0;0;0;0;0;0;0;0;0;17;1;3;122;30;1;807;3;2;3;3;3

348;Kids, Teens, Sports, Hobbies, Sports;9;2;31;2;0;0;0;0;0;0;0;0;0;27;2;0;4;40;2;173;3;2;2;2;2

349;Kids, Teens, Sports, Hobbies, Summer, Camps,
Computer;6;3;0;0;1;3;1;0;0;0;0;0;0;3;1;8;113;89;8;854;6;2;2;2;2

350;"Kids, Teens, Sports, Hobbies, Summer, Camps,

Horseback, Riding";13;3;25;4;2;6;3;2;18;2;0;0;0;4;2;5;11;27;0;335;6;3;3;3;3

351;Kids, Teens, Sports, Hobbies, Summer, Camps,
Space;5;1;25;2;3;7;1;1;6;0;0;0;0;19;0;2;13;21;0;348;2;3;3;3;3

352;Kids, Teens, Teen, Life, Advice;6;0;35;1;1;2;0;2;3;0;4;7;0;5;0;0;73;23;0;407;0;2;3;2;2

353;Kids, Teens, Teen, Life, Fashion;2;1;20;1;1;2;1;3;36;1;0;0;0;2;0;1;31;34;1;442;1;3;3;3;3

354;Kids, Teens, Teen, Life, Magazines, E-
zines;7;0;34;0;1;0;0;0;0;0;10;54;1;49;1;1;96;16;0;350;0;2;2;2;2

355;Kids, Teens, Teen, Life, Proms;12;1;0;0;1;6;0;2;20;1;0;0;0;13;2;1;31;47;1;2336;1;2;2;2;2

356;Kids, Teens, Your, Family, Adoption;2;1;47;1;1;1;1;7;26;0;0;0;0;1;0;0;8;107;1;957;1;3;3;3;3

357;Kids, Teens, Your, Family, Divorce;20;1;0;0;0;0;0;1;8;0;1;11;0;4;1;1;17;93;2;593;1;2;2;2;2

358;Kids, Teens, Your, Family, Divorce;4;1;14;2;1;4;1;3;7;0;4;15;2;2;1;3;9;43;3;909;3;2;3;3;3

359;Kids, Teens, Your, Family, Grandparents;3;1;19;2;1;3;1;3;7;0;4;16;1;2;1;5;9;48;3;819;4;2;3;3;3

360;Kids, Teens, Your, Family, Pets;3;0;0;0;0;0;0;0;0;0;0;0;0;12;0;1;10;51;2;1221;2;1;2;1;1

361;Kids, Teens, Your, Family, Pets;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;42;1;0;0;1;1;1;1

362;Recreation, Aviation, Aircraft, Restoration;6;2;17;2;0;0;0;0;0;0;0;0;0;1;0;2;13;30;0;193;2;2;3;3;3

363;Recreation, Aviation, Military, Guides;10;2;21;2;1;6;2;4;8;2;7;16;0;13;2;2;17;24;0;827;2;2;3;3;3

364;Recreation, Aviation, Model, Aviation;4;0;0;0;1;10;0;0;0;0;0;0;0;0;0;0;0;2;27;3;10;0;1;1;2;1

365;Recreation, Aviation, Model, Aviation;3;0;0;0;0;0;0;1;4;0;6;8;0;4;0;2;2;18;0;362;3;1;2;1;1

366;Recreation, Boating, Sailing;19;2;27;3;1;0;0;4;17;1;3;10;1;6;2;2;25;20;0;338;3;2;3;3;3

367;Recreation, Boating, Sailing;6;1;20;1;1;4;1;0;0;0;0;0;0;2;0;2;0;34;2;81;2;2;3;3;3

368;Recreation, Camps, Academic;2;0;0;0;0;0;0;0;0;0;2;4;0;0;0;0;18;33;0;846;1;1;2;1;1

369;Recreation, Camps, Academic;5;1;0;0;1;4;1;0;0;0;2;4;0;1;0;0;101;37;0;660;1;1;1;2;1

370;Recreation, Camps, Academic;3;1;0;0;2;18;0;0;0;0;4;8;0;6;1;1;60;23;0;413;2;2;1;2;2

371;Recreation, Camps, Alumni;3;0;0;0;0;0;0;0;0;0;0;0;0;17;2;1;11;27;1;429;1;1;2;1;1

372;Recreation, Camps, Alumni;7;2;14;2;0;0;0;0;0;0;0;0;0;0;0;2;1;40;2;135;3;2;2;2;2

373;Recreation, Camps, Equestrian;5;0;9;1;0;0;0;0;0;0;0;0;0;12;0;1;2;28;0;372;3;2;2;2;2

374;Recreation, Camps, Equestrian;3;0;16;1;3;12;1;1;42;1;8;98;2;10;1;1;15;30;1;292;2;3;2;2;2

375;Recreation, Camps, Space;4;1;0;0;1;4;1;1;1;0;6;14;1;11;2;2;13;33;1;277;3;2;2;2;2

376;Recreation, Camps, Space;8;1;24;1;2;7;1;0;0;0;0;0;0;5;3;2;26;26;1;152;2;2;2;3;2

377;Recreation, Camps, Space;7;2;0;0;0;0;0;1;9;1;0;0;0;2;0;0;3;30;2;232;2;2;2;2;2

378;Recreation, Camps, Sports;3;2;0;0;2;21;0;1;3;0;1;1;0;20;0;3;23;31;2;8879;7;1;2;2;2

379;Recreation, Camps, Sports;13;2;29;3;1;3;1;1;2;0;5;32;2;1;0;5;56;28;1;386;4;2;3;3;3

380;Recreation, Camps, Travel, Adventure;5;3;28;4;1;5;3;3;33;3;1;6;0;22;3;2;75;33;1;388;3;3;3;3;3

381;Recreation, Camps, Travel, Adventure;7;1;17;2;9;38;2;1;28;0;3;12;2;1;0;5;132;28;1;472;4;3;3;3;3

382;Recreation, Camps, Travel, Adventure;9;2;23;2;1;5;1;2;11;1;16;43;1;36;0;0;106;39;1;467;0;3;3;3;3

383;Recreation, Camps, Weight, Loss;12;4;21;4;1;7;1;12;55;3;7;30;2;0;0;3;9;29;1;757;5;3;3;2;3

384;Recreation, Collecting, Stamps;5;1;0;0;1;5;1;1;1;0;0;0;0;0;0;1;213;33;1;1119;2;1;2;1;1

385;Recreation, Collecting, Stamps;8;1;14;1;0;0;0;2;4;1;0;0;0;18;1;1;3;29;1;93;2;2;3;3;3

386;Recreation, Drugs, Cannabis;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0;2;49;30;0;126;0;1;1;1;1

387;Recreation, Drugs, Cannabis;2;0;0;0;0;0;0;14;74;0;29;236;3;13;3;1;266;22;0;773;2;2;2;1;2

388;Recreation, Guns, Education;8;0;25;0;1;4;0;10;98;0;9;103;1;6;0;1;52;22;0;815;2;2;2;3;2

389;Recreation, Guns, Education;6;0;14;1;0;0;0;0;0;0;0;0;0;0;0;0;0;30;0;39;0;1;1;3;1

390;Recreation, Guns, Education;4;0;0;0;1;3;0;0;0;0;9;37;0;5;0;0;7;40;0;405;1;1;1;2;1

391;Recreation, Kites, Organizations;3;0;0;0;0;0;0;0;0;0;0;0;0;2;0;0;2;38;0;10;0;1;1;1;1

392;Recreation, Kites, Organizations;6;1;0;0;1;1;0;0;0;0;0;0;0;1;1;2;18;23;0;240;3;2;2;2;2

393;Recreation, Pets, Dogs;8;1;0;0;0;0;0;0;0;0;0;0;0;0;0;1;0;31;1;348;3;1;2;1;1

394;Recreation, Pets, Dogs;3;0;0;0;1;1;0;6;15;1;2;7;1;9;1;5;109;32;0;745;5;2;2;2;2

395;Recreation, Scouting, Campsites;2;0;0;0;1;2;0;1;8;0;0;0;0;3;0;0;7;62;0;249;1;1;1;2;1

396;Recreation, Scouting, Philanthropy;1;0;8;1;0;0;0;1;1;0;0;0;0;8;1;1;8;36;0;113;1;1;1;2;1

397;Recreation, Tobacco, Cigars;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;10;32;1;101;0;1;2;1;1

398;Recreation, Tobacco, Cigars;6;0;41;2;0;0;0;8;32;2;7;32;1;39;2;2;124;29;0;784;2;2;2;3;2

399;Reference, Almanacs, Arts;20;1;14;0;1;6;1;2;7;1;2;11;1;2;0;1;67;27;1;2387;2;2;2;2;2

400;Reference, Almanacs, History;5;1;0;0;1;11;0;1;10;0;33;372;1;10;0;1;2;54;1;7172;1;1;1;2;1

401;Reference, Almanacs, Sports;9;1;25;1;0;0;0;0;0;0;1;8;1;31;1;1;27;33;1;1268;3;2;2;3;2

402;Reference, Dictionaries, Acronyms;9;2;9;2;1;2;1;0;0;0;0;0;0;0;0;1;1;26;1;240;2;2;2;1;2

403;Reference, Dictionaries, Etymology;1;1;0;0;0;0;0;1;9;0;1;4;0;2;0;1;8;41;0;250;1;1;2;2;2

404;Reference, Dictionaries, Pronouncing;4;2;0;0;1;4;2;0;0;0;4;27;1;0;0;1;10;45;0;419;2;1;1;1;1

405;Reference, Dictionaries, Rhyming;6;2;0;0;0;0;0;0;0;0;0;0;0;39;2;2;0;47;0;231;4;2;1;1;1

406;Reference, Education, How, Study;8;2;21;1;1;0;0;0;0;0;2;3;0;38;2;1;20;30;1;455;2;2;3;3;3

407;"Reference, Encyclopedias, Subject,
Encyclopedias, Paranormal";2;2;0;0;0;0;0;0;0;0;0;0;8;3;1;7;40;3;167;1;1;1;1;1

408;"Reference, Encyclopedias, Subject,
Encyclopedias, Plants";9;2;28;3;0;0;0;0;0;0;0;0;0;3;0;1;1;27;1;179;3;2;3;3;3

409;Reference, Flags, Etiquette;3;2;0;0;1;3;2;7;17;3;4;12;1;5;2;3;5;41;2;1123;4;2;2;2;2

410;Reference, Flags, Nautical;9;1;15;2;2;12;2;0;0;0;0;0;0;2;0;2;3;38;1;319;3;3;3;2;3

411;Reference, Flags, Vexillology;2;0;0;0;2;2;0;0;0;0;3;15;1;3;1;1;66;17;0;535;2;1;2;1;1

412;Reference, Knots, Boating;4;2;9;2;1;2;2;3;4;1;0;0;0;4;0;4;26;47;3;537;4;3;3;3;3

413;Reference, Knots, Fishing;5;2;19;2;1;5;2;8;19;1;3;24;3;1;0;4;37;36;2;285;4;3;3;2;3

414;Reference, Knowledge, Management, Information,
Architecture;7;2;0;0;3;11;2;9;23;2;15;109;2;8;0;2;166;92;2;3320;3;2;2;2;2

415;Reference, Knowledge, Management, Knowledge, Discovery,
Software;10;4;15;1;0;0;0;0;0;0;0;0;0;0;0;0;0;36;0;56;5;2;2;2;2

416;Reference, Libraries, Digital;3;0;0;0;0;0;0;4;28;0;6;16;0;36;0;3;3;23;0;412;2;1;1;1;1

417;Reference, Libraries, National;3;2;0;0;1;0;0;7;16;1;10;28;2;4;2;1;158;19;0;782;2;1;1;1;1

418;Reference, Libraries, Presidential;8;2;32;3;6;25;2;0;0;0;0;0;0;5;2;2;3;36;0;267;3;3;3;2;3

419;Reference, Libraries, Research,
Associations;5;2;44;3;0;0;0;10;128;3;4;7;0;7;1;1;24;20;0;485;3;2;2;2;2

420;Reference, Maps, Collecting;7;2;26;2;1;4;2;1;2;0;6;50;2;0;0;5;55;28;0;623;5;3;3;3;3

421;Reference, Maps, OpenStreetMap;9;0;18;1;2;12;0;2;5;0;2;5;0;2;0;1;299;72;2;1304;1;2;2;2;2

422;Reference, Museums, Cultural;6;1;30;2;0;0;0;2;13;1;1;3;0;11;0;1;75;24;1;1008;1;3;3;3;3

423;Reference, Museums, Cultural;3;1;0;0;1;2;0;11;36;0;1;3;0;7;0;1;26;34;1;730;1;1;1;2;1

424;Reference, Museums, Military;13;0;0;0;7;18;0;1;6;0;0;0;0;0;0;0;69;21;0;422;0;1;1;2;1

425;Reference, Museums, Science;3;1;14;2;5;11;0;6;42;1;0;0;0;7;1;0;11;34;1;86;2;3;3;3;3

426;Reference, Museums,
Transportation;4;2;2;0;2;16;2;12;56;2;21;94;0;13;0;2;18;33;2;1534;2;2;2;2;2

427;Reference, Time, Clocks, Watches;5;1;16;2;0;0;0;1;1;0;1;2;0;3;1;2;4;38;1;350;2;3;3;2;3

428;Reference, Time, Horology;4;0;13;0;1;4;0;0;0;0;0;0;0;5;0;0;13;37;0;387;0;2;2;2;2

429;Reference, Time, Horology;9;1;50;1;0;0;0;0;0;0;1;1;0;5;0;1;7;43;0;171;1;2;3;3;3

430;Science, Astronomy, History, People, Galilei,
Galileo;5;1;0;0;2;8;1;0;0;0;9;39;1;0;0;1;15;48;1;888;1;2;2;2;2

431;Science, Biology, Bioinformatics;5;1;4;1;1;4;1;0;0;0;0;0;0;1;0;1;4;37;0;113;2;1;2;2;2

432;Science, Biology, Bioinformatics;5;0;3;0;2;4;0;3;60;2;4;11;0;6;0;0;4;31;0;429;4;1;2;2;2

433;Science, Biology, Botany;3;0;41;2;1;6;0;0;0;0;11;27;0;11;0;2;24;25;0;1302;1;3;2;3;3

434;Science, Biology, Ecology, Ecosystems;3;0;0;0;0;0;0;0;0;0;0;0;9;0;1;3;46;0;745;0;1;1;1;1

435;Science, Biology, Ecology, Ecosystems;5;1;21;1;0;0;0;0;0;0;0;0;2;0;1;2;63;0;222;2;2;2;2;2

436;Science, Biology, Education;1;0;0;0;0;0;0;32;135;6;0;0;0;38;4;5;119;35;0;1436;4;2;2;1;2

437;Science, Biology, Education;7;0;24;0;1;6;1;1;9;0;2;24;1;0;0;1;58;24;0;569;1;2;2;2;2

438;Science, Biology, Evolution;2;1;6;0;0;0;0;0;0;0;1;5;0;18;1;1;3;42;1;125;1;2;2;2;2

439;Science, Biology, Evolution;2;1;0;0;1;2;1;0;0;0;3;25;1;1;0;0;6;47;0;1008;2;1;1;2;1

440;Science, Biology, Flora, Fauna;4;0;162;0;0;0;0;0;0;0;0;0;0;18;0;0;1;32;0;263;0;1;2;1;1

441;Science, Biology, Flora, Fauna;2;0;21;0;0;0;0;0;0;0;0;0;0;24;0;0;10;29;0;154;0;2;2;2;2

442;Science, Biology, Genetics;6;0;0;0;1;3;0;7;11;0;3;17;0;2;0;3;2;25;0;324;1;2;2;1;2

443;Science, Biology, Immunology, Antibodies;6;1;0;0;1;5;1;3;5;0;18;50;1;1;0;4;40;115;1;881;1;1;2;1;1

444;Science, Biology, Methods, Techniques;2;0;0;0;12;55;0;1;8;1;1;1;0;3;0;2;77;32;0;453;1;1;2;1;1

445;Science, Biology, Neurobiology;1;0;0;0;0;0;0;0;0;0;0;0;0;4;0;0;2;38;0;152;0;1;1;1;1

446;Science, Biology, Neurobiology;8;0;13;0;0;0;0;0;0;0;0;0;0;0;0;0;4;25;0;175;0;1;1;2;1

447;Science, Biology, Physiology;11;0;0;0;2;8;0;5;23;1;0;0;0;12;1;1;70;18;0;1407;0;1;1;2;1

448;Science, Chemistry, Chemists;1;0;0;0;0;0;0;0;0;0;0;0;0;2;0;0;3;28;0;58;2;1;1;1;1

449;Science, Chemistry, Organic;6;2;0;0;11;39;2;0;0;0;4;5;0;50;0;2;3;37;1;1791;4;2;2;2;2

450;Science, Chemistry, Organizations;7;0;0;0;0;0;0;3;9;0;18;61;1;8;0;1;27;30;0;433;1;1;2;1;1

451;Science, Environment, Air, Quality;9;3;22;3;0;0;0;0;0;0;0;0;0;1;1;2;2;24;1;92;2;2;3;3;3

452;Science, Environment, Air, Quality;5;1;20;1;1;0;0;0;0;0;14;57;1;25;1;0;13;20;0;445;2;3;3;3;3

453;Science, Math, Calculus;2;0;0;0;1;4;0;9;41;2;0;0;0;2;0;1;30;24;1;192;3;1;2;2;2

454;Science, Math, Operations, Research;2;0;0;0;1;0;0;1;5;0;0;0;0;0;0;0;84;45;0;606;2;1;1;2;1

455;Science, Math, Operations, Research;4;0;0;0;1;4;0;7;48;0;0;0;0;1;0;3;18;59;1;1494;1;1;1;2;1

456;Science, Social, Sciences, Archaeology;9;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;49;0;0;0;1;1;1;1

457;Science, Social, Sciences, Communication, Rhetoric;5;1;6;1;1;4;1;10;25;2;0;0;0;1;1;1;14;58;1;3223;2;2;3;3;3

458;Science, Social, Sciences, Communication, Rhetoric;9;1;0;0;1;5;1;10;29;1;7;36;1;6;0;1;20;51;1;16589;1;2;2;2;2

459;Science, Technology, Civil, Engineering;9;2;79;3;1;2;2;8;20;0;0;0;0;1;2;2;46;58;2;248;2;3;3;3;3

460;Science, Technology, Energy, Conservation;5;2;32;1;1;3;1;46;117;2;7;56;1;11;0;2;476;63;1;3213;1;3;3;2;3

461;Science, Technology, Energy, Conservation;6;1;85;1;1;1;1;1;3;0;0;0;0;1;0;2;62;41;0;1213;1;3;3;2;3

462;Science, Technology, Materials, Ceramics;2;0;0;0;0;0;0;0;0;0;0;0;1;0;0;1;28;0;10;0;1;1;1;1

463;Science, Technology, Materials, Ceramics;27;2;0;0;0;0;0;1;3;1;7;31;1;0;0;3;34;22;0;357;2;2;2;2;2

464;Science, Technology, Mining;9;1;0;0;1;3;1;1;5;0;49;116;2;0;0;1;21;86;0;1477;0;1;1;2;1

465;Science, Technology, Mining;2;2;26;2;4;35;2;40;129;3;188;238;0;1;0;3;200;34;1;2174;2;3;3;2;3

466;Science, Technology, Pyrotechnics;4;1;19;1;1;4;0;4;6;0;2;4;0;7;1;0;5;20;0;35;1;2;2;2;2

467;Science, Technology, Television;11;0;0;0;1;0;0;1;6;1;1;4;0;1;1;0;40;28;0;710;0;1;1;2;1

468;Science, Technology, Transportation;6;0;7;0;0;0;0;0;0;0;0;0;0;7;0;0;3;29;0;386;1;2;2;2;2

469;Shopping, Antiques, Collectibles, Beer, Steins, Glasses;11;2;25;2;1;2;0;1;3;0;8;16;2;12;2;2;19;31;1;292;4;2;3;3;3

470;Shopping, Antiques, Collectibles, Cameras, Photographs;9;0;65;3;0;0;0;0;0;0;0;0;0;1;0;2;3;30;0;188;2;2;2;3;2

471;Shopping, Antiques, Collectibles, Cameras, Photographs;7;1;34;1;2;5;0;0;0;0;0;0;0;4;0;1;1;29;0;263;1;2;2;2;2

472;Shopping, Antiques, Collectibles, Ceramics, Pottery;8;0;24;1;21;61;0;0;0;0;4;5;1;11;0;1;96;31;0;736;1;2;2;2;2

473;Shopping, Antiques, Collectibles, Ceramics, Pottery;9;3;21;3;0;0;0;0;0;0;0;0;37;4;0;68;27;0;1802;2;2;3;3;3

474;Shopping, Antiques, Collectibles, Ceramics, Pottery;12;1;24;2;7;32;2;3;22;0;0;0;0;0;0;2;4;24;0;204;2;2;2;2;2

475;Shopping, Antiques, Collectibles, Clocks, Watches;3;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;34;2;90;4;1;1;1;1

476;Shopping, Antiques, Collectibles, Clocks, Watches;18;3;0;0;0;0;0;0;0;0;0;0;82;4;2;0;25;1;958;7;2;2;2;2

477;Shopping, Antiques, Collectibles, Clocks, Watches;13;3;0;0;2;5;1;7;31;3;3;13;2;5;1;2;59;37;2;322;5;2;2;2;2

478;Shopping, Auctions, Autos;9;1;29;1;1;2;1;3;6;0;7;63;1;54;3;1;74;21;0;1308;3;3;3;3;3

479;Shopping, Auctions, Autos;22;2;22;2;1;7;1;3;10;1;0;0;0;4;0;0;43;23;0;652;4;3;2;3;3

480;Shopping, Auctions, Boats;7;2;23;1;2;17;2;2;11;2;0;0;0;3;1;5;11;30;1;252;4;3;3;3;3

481;Shopping, Auctions, Guns;6;1;26;4;1;5;2;6;30;3;10;24;1;55;1;4;94;35;1;1882;4;3;3;3;3

482;Shopping, Auctions, Jewelry;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;26;0;0;0;1;1;1;1

483;Shopping, Auctions, Jewelry;7;1;38;1;3;9;1;3;5;1;14;45;1;0;0;2;23;39;1;528;1;2;3;3;3

484;Shopping, Auctions, Sports;11;2;8;1;0;0;0;3;7;0;6;13;2;2;2;0;21;30;2;481;3;3;3;3;3

485;Shopping, Clothing, Leather, Flight, Jackets;13;3;0;0;1;29;3;4;16;3;0;0;0;1;0;4;97;30;0;603;2;2;2;2;2

486;Shopping, Clothing, Leather, Flight, Jackets;7;3;28;5;1;7;1;1;3;0;8;14;1;13;3;4;5;36;2;439;4;3;3;3;3

487;Shopping, Clothing, Natural, Fiber;4;1;18;1;1;2;0;0;0;0;0;0;0;10;0;1;316;33;0;658;1;2;2;2;2

488;Shopping, Clothing, Natural, Fiber;7;1;19;0;11;28;0;5;11;0;0;0;0;7;1;0;40;32;0;133;1;2;2;2;2

489;Shopping, Clothing, Swimwear;2;0;0;0;0;0;0;0;0;0;0;0;0;1;0;0;2;29;0;499;0;1;1;1;1

490;Shopping, Clothing, Swimwear;9;3;25;2;1;9;2;8;25;2;0;0;0;10;0;3;67;25;0;496;3;2;3;3;3

491;Shopping, Clothing, Uniforms, Aviation;5;0;13;1;1;0;0;6;12;0;2;5;0;1;0;1;7;42;1;43;0;2;2;2;2

492;Shopping, Clothing, Uniforms, Aviation;2;0;21;2;1;7;1;10;31;2;5;11;1;2;0;3;14;30;0;199;3;2;2;2;2

493;Shopping, Clothing, Uniforms, Postal;8;3;27;3;1;5;2;1;4;0;2;35;4;10;4;3;1;41;2;549;4;2;3;3;3

494;Shopping, Clothing, Uniforms, School;8;2;23;3;0;0;0;2;8;2;3;12;2;17;3;4;0;29;0;494;3;2;2;2;2

495;Shopping, Flowers, Fresh, Cut;2;0;31;2;3;7;0;5;23;1;1;11;1;8;1;2;6;30;1;94;2;2;2;2;2

496;Shopping, Flowers, Fresh, Cut;3;1;6;0;1;5;1;0;0;0;0;0;0;0;0;1;0;31;1;279;2;2;2;2;2

497;Shopping, Gifts, Clubs;19;3;23;4;1;5;2;1;11;1;1;4;0;84;3;4;9;29;1;966;2;2;3;3;3

498;Shopping, Jewelry, Diamonds;9;2;17;4;0;0;0;0;0;0;0;0;0;0;0;0;2;44;28;0;2056;3;2;2;2;2

499;Shopping, Music, Equipment, Karaoke;22;1;11;1;0;0;0;0;0;0;0;0;0;0;0;1;16;29;1;896;2;2;2;2;2

500;Shopping, Music, Equipment, Karaoke;9;3;25;3;1;0;0;0;0;0;9;461;1;42;2;3;10;28;1;2269;2;2;2;2;2

501;Shopping, Music, Equipment,
Karaoke;2;1;17;2;0;0;0;1;2;0;17;51;0;19;1;1;140;27;0;590;1;2;2;2;2

502;Shopping, Music, Instruments;16;3;1;0;0;0;0;0;0;0;0;0;0;0;0;1;0;26;0;142;3;2;2;1;2

503;Shopping, Music, Instruments;5;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;2;38;1;390;3;1;2;1;1

504;Shopping, Music, Instruments;4;2;15;3;6;25;1;8;35;1;10;34;0;2;1;2;199;26;1;287;3;3;3;3;3

505;Shopping, Publications, Books,
Military;12;2;29;2;9;18;1;1;3;0;1;7;1;24;2;3;10;37;0;711;3;2;2;2;2

506;Shopping, Publications, Books, Military;4;0;0;0;0;0;0;0;0;0;0;0;0;3;1;4;15;25;0;338;4;1;1;1;1

507;Shopping, Publications, Books, Pets;3;1;27;2;1;13;1;5;53;2;1;12;1;1;1;0;8;30;1;2398;2;3;3;2;3

508;Shopping, Tobacco, Pipes;4;1;24;1;0;0;0;15;66;3;8;15;0;18;3;3;5;35;1;247;4;2;2;2;2

509;Shopping, Tobacco, Pipes;10;2;0;0;0;0;0;0;0;0;1;14;1;5;2;3;1;25;1;270;2;1;2;2;2

510;Shopping, Tools, Woodworking;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;39;1;181;3;2;2;1;2

511;Shopping, Tools, Woodworking;3;0;0;0;1;141;2;3;7;1;4;9;0;9;0;3;53;27;0;476;3;1;2;2;2

512;Shopping, Toys, Games, Educational;16;2;23;3;1;3;0;5;15;0;0;0;0;24;0;0;207;28;0;497;1;3;3;2;3

513;Shopping, Toys, Games, Educational;3;2;56;4;0;0;0;6;37;3;2;8;0;16;4;5;26;39;2;974;3;2;2;2;2

514;Shopping, Travel, Luggage;8;2;16;2;5;24;1;1;2;0;1;9;0;13;3;2;9;30;1;307;3;3;3;2;3

515;Shopping, Travel, Luggage;13;3;27;2;0;0;0;74;165;3;7;21;0;18;2;3;19;22;0;917;1;2;3;3;3

516;Shopping, Vehicles, Aircraft;3;0;0;0;0;0;0;4;5;0;0;0;0;2;0;1;1;29;0;419;1;1;2;1;1

517;Shopping, Vehicles, Aircraft;14;1;0;0;2;5;0;1;10;1;0;0;0;0;0;1;20;29;0;341;1;2;2;2;2

518;Shopping, Vehicles, Aircraft;20;1;22;1;2;24;1;4;7;1;15;47;1;21;1;1;6;31;1;267;1;3;3;3;3

519;Shopping, Vehicles, Aircraft,
Paramotor;8;0;25;0;1;1;0;14;34;0;8;11;0;1;0;0;126;42;0;567;0;1;1;1;1

520;Shopping, Vehicles, Aircraft, Parts,
Accessories;2;0;17;1;6;196;0;6;22;0;0;0;0;0;0;1;22;25;0;369;0;2;1;1;1

521;Shopping, Vehicles, Aircraft, Parts,
Accessories;3;1;0;0;0;0;0;0;0;0;0;0;0;0;0;1;0;25;0;659;0;1;1;1;1

522;Shopping, Vehicles, Aircraft, Parts,
Accessories;7;0;0;0;1;4;1;1;7;0;0;0;0;10;1;1;3;24;0;239;1;2;2;1;2

523;Shopping, Vehicles, Aircraft, Ultralight;3;0;13;0;0;0;0;0;0;0;0;0;0;18;0;0;2;34;0;303;0;2;1;1;1

524;Shopping, Vehicles, Aircraft, Ultralight;20;1;43;1;0;0;0;0;0;0;0;0;0;1;0;1;1;23;0;302;1;2;2;2;2

525;Shopping, Vehicles, Autos, Appraisers;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;22;0;0;0;1;1;1;1

526;Shopping, Vehicles, Autos, Appraisers;6;3;75;3;1;0;0;2;2;0;4;8;0;14;2;4;145;33;2;912;6;2;2;2;2

527;Shopping, Vehicles, Autos, Appraisers;12;1;44;1;0;0;0;0;0;0;0;0;0;4;2;1;0;27;0;287;7;2;2;2;2

528;Society, Crime, Internet, Crime;7;0;25;0;1;7;0;5;7;0;0;0;0;8;0;1;2;50;0;1430;0;1;1;2;1

529;Society, Crime, Prisons;10;1;29;1;2;7;0;1;4;0;0;0;0;10;1;2;2;28;0;839;2;3;2;2;2

530;Society, Death, Death, Care, Ash, Scattering;9;3;0;0;1;3;1;0;0;0;1;7;2;0;0;0;0;29;1;203;4;2;2;2;2

531;Society, Death, Suicide;3;1;32;2;0;0;0;0;0;0;0;0;0;4;1;2;192;35;1;759;4;2;2;2;2

532;Society, Folklore, Weather, Beliefs;4;2;13;2;0;0;0;4;25;1;23;302;0;1;0;0;5;35;0;439;3;2;3;3;3

533;Society, Genealogy, Royalty;8;1;39;1;0;0;0;0;0;0;0;0;0;4;1;0;19;38;0;3106;1;2;3;3;3

534;Society, History, Education;2;0;0;0;1;10;2;1;11;0;15;36;2;8;1;3;4;28;0;564;3;1;2;2;2

535;Society, History, Education;3;0;29;2;0;0;0;0;0;0;0;0;0;15;1;1;3;52;0;159;2;1;2;2;2

536;Society, History, On, This, Day, in, History;9;3;18;3;1;0;0;0;0;0;0;0;0;1;1;0;0;32;2;87;9;2;2;2;2

537;Society, Law, Legal, Information;7;3;14;3;0;0;0;1;5;0;0;0;0;8;1;4;151;24;1;671;4;2;2;2;2

538;Society, Law, Legal, Information, Bankruptcy;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;25;1;0;0;1;2;1;1

539;Society, Law, Legal, Information, Drunk, Driving;5;0;0;0;1;5;0;1;5;0;0;0;0;0;0;0;12;76;2;82;0;1;1;2;1

540;Society, Law, Legal, Information, Drunk, Driving;3;2;0;0;12;56;3;5;7;1;4;8;2;10;4;5;40;50;1;1922;3;1;2;1;1

541;Society, Law, Legal, Information, Drunk, Driving;2;0;29;4;1;9;1;1;2;0;0;0;0;31;4;6;0;69;0;901;6;2;2;2;2

542;Society, Law, Legal, Information, Family, Law, Divorce;8;8;6;6;0;0;0;1;8;2;6;27;10;14;4;7;47;27;1;616;9;3;3;1;3

543;Society, Law, Legal, Information, International, Law;10;2;45;3;0;0;0;0;0;0;0;0;0;0;3;0;34;2;249;4;2;3;3;3

544;Society, Law, Legal, Information, Tax;3;1;8;2;0;0;0;3;16;1;3;8;1;6;1;1;31;20;1;679;2;2;2;2;2

545;Society, Military, Veterans;26;1;0;0;1;13;1;11;35;0;9;15;0;3;0;1;39;28;0;249;1;2;2;2;2

546;Society, Military, Weapons, Equipment, Small, Arms;4;2;3;2;1;5;2;0;0;0;4;7;0;11;4;2;8;32;2;410;2;2;2;2;2

547;Society, Paranormal, Prophecies;19;1;17;1;1;17;1;1;32;0;13;27;1;5;0;2;665;30;0;3978;2;2;2;2;2

548;Society, Paranormal, UFOs;2;0;0;0;0;0;0;0;0;0;0;0;5;1;1;9;31;0;392;1;1;1;1;1

549;Society, Philanthropy, Refugees;8;1;22;0;1;2;0;39;72;2;11;62;0;0;0;2;111;22;0;853;1;3;3;3;3

550;Society, Philanthropy, Volunteering;2;1;26;3;1;1;1;9;20;3;8;25;2;25;2;3;54;42;1;821;3;3;3;3;3

551;Society, Politics, Christian, Democracy;4;2;20;2;1;2;2;5;9;0;0;0;0;2;0;2;1;43;0;2671;2;3;3;3;3

552;Society, Politics, Globalization;8;1;0;0;2;3;0;8;21;2;6;36;2;3;0;2;30;22;0;326;2;2;2;2;2

553;Society, Politics, Liberalism, Parties;4;1;0;0;1;4;1;2;6;0;0;0;0;15;1;3;126;21;0;1464;2;2;2;2;2

554;Society, Religion, Spirituality, Hinduism;5;1;9;0;3;4;1;0;0;0;0;0;0;4;1;0;1;36;1;717;1;2;2;2;2

555;Society, Religion, Spirituality, Judaism, Camps;3;0;3;0;4;40;1;4;24;0;0;0;0;12;0;1;28;24;0;358;1;1;2;2;2

556;Society, Religion, Spirituality, Taoism;12;1;0;0;5;10;0;1;10;1;1;5;0;1;0;3;28;23;0;382;2;1;2;2;2

557;Society, Religion, Spirituality, Taoism;1;1;5;2;3;5;1;3;6;1;3;19;1;17;0;3;4;45;1;216;3;3;2;3;3

558;Society, Transgendered, Crossdressing;2;0;22;1;0;0;0;14;58;0;26;128;0;40;0;1;290;36;1;1526;1;2;2;2;2

559;Sports, Badminton, Associations, Ireland;4;2;0;0;2;5;2;2;5;2;4;10;1;0;0;3;7;32;1;220;4;2;2;2;2

560;Sports, Basketball, Coaching;8;2;42;3;1;6;2;0;0;0;0;0;0;14;1;1;1;30;1;104;4;3;2;3;3

561;Sports, Basketball, Regional, Australia, Victoria;4;1;0;0;0;0;0;9;25;1;20;148;1;45;2;1;100;39;1;376;2;1;2;1;1

562;Sports, Basketball, Wheelchair;7;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;29;0;0;0;1;1;1;1

563;Sports, Basketball, Wheelchair;5;2;0;0;1;3;2;0;0;0;4;9;2;6;2;2;11;40;0;312;2;1;2;2;2

564;Sports, Boomerang, Design, Construction;2;1;0;0;0;0;0;0;0;0;0;0;0;0;12;0;1;20;73;1;267;2;1;2;2;2

565;Sports, Bowling, Lawn, Bowling;17;3;53;3;1;8;3;0;0;0;0;0;0;7;3;3;7;34;3;1286;3;2;2;2;2

566;Sports, Boxing, History;7;2;19;2;0;0;0;0;0;0;0;0;0;0;0;0;1;36;1;107;4;1;3;3;3

567;Sports, Cricket, Blind;4;2;0;0;1;4;2;0;0;0;1;22;2;2;0;2;2;46;2;206;2;2;2;2;2

568;Sports, Cricket, Blind;6;2;0;0;1;1;0;0;0;0;9;53;2;4;0;2;7;23;0;599;2;1;2;2;2

569;Sports, Cricket, History;8;2;8;2;3;8;2;0;0;0;0;0;0;0;0;2;20;57;0;2021;2;2;2;2;2

570;Sports, Cricket, Players;18;2;17;1;0;0;0;0;0;0;1;3;0;103;1;3;4;40;2;513;3;1;2;2;2

571;Sports, Cue, Sports, English, Billiards, Associations;1;0;0;0;0;0;0;55;349;2;5;12;0;4;2;2;134;23;0;8485;2;1;1;2;1

572;Sports, Cycling, Bike, Shops, Kansas;5;2;65;5;0;0;0;2;11;0;0;0;0;0;0;3;14;33;1;193;4;3;3;3;3

573;Sports, Fencing, Clubs, New, Zealand;5;2;29;7;1;4;2;0;0;0;0;0;0;0;0;3;2;39;1;53;3;3;2;3;3

574;Sports, Fencing, College, University;6;1;0;0;1;1;0;1;6;0;1;0;0;1;0;0;0;33;1;208;2;1;2;2;2

575;Sports, Fencing, Wheelchair;14;3;9;3;1;1;0;1;5;2;6;24;2;11;0;0;74;20;0;543;4;3;3;3;3

576;Sports, Football, American, High, School;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0;1;3;30;1;258;2;1;2;1;1

577;Sports, Football, American, High, School;4;1;0;0;0;0;0;1;2;0;9;10;1;0;0;2;102;44;0;336;1;1;1;1;1

578;Sports, Golf, Courses, Maryland;4;2;0;0;0;0;0;0;0;0;0;0;0;2;1;2;4;36;2;256;4;1;2;1;1

579;Sports, Golf, Disabled, Blind;5;2;0;0;1;0;0;2;4;0;0;0;0;3;3;2;44;28;2;184;2;2;2;2;2

580;Sports, Gymnastics, Artistic;7;1;0;0;1;6;1;0;0;0;3;8;1;2;1;2;1;20;0;872;2;2;2;2;2

581;Sports, Gymnastics, Rhythmic;5;2;0;0;0;0;0;0;0;0;0;0;0;1;0;2;3;38;0;16;2;1;2;1;1

582;Sports, Lacrosse, Camps;9;2;0;0;18;78;4;0;0;0;0;0;0;3;3;2;23;26;0;792;3;1;2;2;2

583;Sports, Lacrosse, Clubs;2;0;0;0;0;0;0;0;0;0;0;0;0;2;0;1;7;31;0;64;2;1;2;1;1

584;Sports, Martial, Arts, Kickboxing;13;0;0;0;2;3;0;6;39;1;16;48;3;1;0;4;52;30;1;419;3;2;2;2;2

585;Sports, Paintball, Tournaments;11;1;1;0;1;10;1;4;16;2;2;24;1;2;1;1;24;26;1;698;3;2;2;2;2

586;Sports, Rodeo, Schools;10;1;27;1;1;0;0;0;0;0;4;7;0;9;1;0;5;29;0;128;2;2;2;2;2

587;Sports, Skating, Roller, Skating;6;3;6;3;0;0;0;2;5;0;0;0;0;16;7;0;1;36;1;157;7;2;2;2;2

588;Sports, Soccer, CONCACAF, Canada, Youth;6;0;0;0;0;0;0;4;11;2;0;0;0;10;2;2;48;25;0;421;3;1;2;2;2

589;Sports, Soccer, CONCACAF, Canada, Youth;4;1;0;0;0;0;0;1;5;1;7;64;1;0;0;1;53;36;2;210;1;1;2;2;2

590;Sports, Soccer, Freestyle;8;1;13;1;1;7;1;5;26;1;2;4;0;2;1;1;26;22;1;219;1;2;2;2;2

591;Sports, Soccer, UEFA, England, Clubs, M, Marine;3;3;14;4;0;0;0;1;1;0;9;20;6;72;13;19;179;38;4;629;17;2;2;2;2

592;Sports, Softball, Deaf;6;2;0;0;4;14;2;0;0;0;0;0;0;0;0;7;21;0;49;2;2;2;2;2

593;Sports, Squash, Clubs;13;2;21;2;13;45;2;0;0;0;1;3;0;7;1;2;12;33;1;475;2;1;1;1;1

594;Sports, Squash, Clubs;6;2;4;2;1;4;0;0;0;0;5;19;2;10;0;2;3;23;0;295;2;2;3;3;3

595;Sports, Table, Tennis, Clubs;4;3;4;3;15;112;2;13;160;0;13;53;0;11;2;4;8;42;2;706;3;3;2;3;3

596;Sports, Tennis, Juniors;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;36;1;0;0;1;1;1;1

597;Sports, Volleyball, Tournaments;17;2;0;0;1;4;2;3;10;0;4;8;1;4;0;2;31;23;0;292;2;2;2;2;2

598;Sports, Water, Sports, Swimming, Diving, Events;13;1;0;0;1;0;0;19;131;1;5;18;1;42;1;1;164;30;0;520;3;2;2;2;2

599;Sports, Winter, Sports, Snowboarding, Resorts;3;1;0;0;9;21;1;8;15;3;0;0;0;12;1;4;11;34;1;447;7;1;2;2;2

600;Sports, Wrestling, Amateur, Youth, Wrestling;5;2;0;0;0;0;0;1;6;2;0;0;0;2;3;2;11;74;5;238;4;2;2;2;2

Prilog br. 2: SEO stručnjaci i upute za ocjenjivanje

Skup podataka od 600 stranica ocjenili su sljedeći SEO stručnjaci:

1) Branko Mejak, Mijena d.o.o., Pula, Hrvatska
2) Anju Agrawal, Unified Business Web Solutions Pvt, Jaipur, Indija
3) Nihal Gupta, Twigital Pvt, Indore, Madhya Pradesh, Indija

Prije ocjenjivanja stručnjaci su dobili sljedeće upute:

Na hrvatskom jeziku:

1) Mrežnu stranicu svrstajte u jednu od sljedeće tri moguće kategorije:
    - „neprilagođeno“ – mrežne stranice koje nemaju elemente „on-page“ SEO ili imaju, ali nisu prilagođene ključnim riječima ili imaju jako malu prilagođenost
    - „djelomično prilagođeno“ – mrežne stranice koje imaju određene elemente „on-page“ SEO i prilagođene su određenim ključnim riječima, no nedostaje još elemenata i rada do pune prilagođenosti,
    - „prilagođeno“ – mrežne stranice koje su prilagođene ključnim riječima i izrađene su po SEO preporukama. Mogu sadržavati male nedostatke.

    Svrstavanje stranica u navedene kategorije vršite s obzirom na Vaše mišljenje koliko je mrežna stranica prilagođena po SEO pravilima datim ključnim riječima. Koristite svoje znanje, iskustvo i slobodnu procjenu.
2) Ukoliko mrežna stranica ima sve potrebne elemente i karakteristike SEO ali za neku drugu ključnu riječ (ili riječi) koja nije na popisu ključnih riječi za tu stranicu, i ako se riječi s popisa znatno razlikuju od onih za koje je stranica prilagođena, svrstajte tu stranicu u grupu „neprilagođeno“. Ako je razlika u ključnim riječima manja, tj. riječi su iz iste domene ali se ne pojavljuju na stranici, onda tu stranice svrstajte u kategoriju „djelomično prilagođeno“.
3) Vizualni dizajn mrežne stranice nije bitan kod označavanja.
4) Kod označavanja promatrajte samo tu stranicu koja se učita u pregledniku. Ostale mrežne stranice s te lokacije ne smiju utjecati na Vašu odluku.
5) Kod označavanja promatrajte samo „on-page“ SEO karakteristike – „off-page“ faktori ne ulaze u ovo istraživanje i ne smiju utjecati na Vašu odluku.

Na engleskom jeziku:

1) Web page should be classified in one of these categories:
    - „low on-page SEO“ – web page that doesn't have or have very few on-page SEO elements, or are not adjusted according given keywords,

- „medium on-page SEO“ – web page with few on-page SEO elements adjusted with few keywords, but are missing some elements and further work is needed to full optimization,
- „high on-page SEO“ – web page with all common on-page SEO elements adjusted to given keywords and built according SEO guidelines. They may contain minor flaws.

You should classify given web pages according to your subjective opinion about on-page SEO utilization against given keywords and SEO guidelines. Use your knowledge, experience and free estimation.

2) If a web page contains all on-page SEO elements, but optimized for another keyword, or keyword that is not on the list, and if list of keywords differ significantly from keywords for which the page is optimized, you should classify this page as „low on-page SEO“. If the difference in keywords is lower, for example keywords are from same domain, but does not appear on page, the page should be classfied in „medium on-page SEO“ category.
3) Visual appearance and design is not important for this classification.
4) While classifying observe only the page loaded in browser. Other pages from same site should not affect your decision.
5) While classifying observe only „on-page“ SEO factors. „Off-page“ factors should not affect your decision.

Prilog br. 3: Stablo odlučivanja

```
Mkw <= 0
|   txtKw <= 0
|   |   Mlen <= 7
|   |   |   h2kw <= 0
|   |   |   |   Tlen <= 10: 1 (40.0)
|   |   |   |   Tlen > 10
|   |   |   |   |   urlLen <= 26: 1 (3.0)
|   |   |   |   |   urlLen > 26: 2 (2.0)
|   |   |   h2kw > 0
|   |   |   |   h3 <= 1: 1 (3.0)
|   |   |   |   h3 > 1: 2 (2.0)
|   |   Mlen > 7
|   |   |   urlLen <= 30
|   |   |   |   h1 <= 0
|   |   |   |   |   txtLen <= 164: 2 (4.0)
|   |   |   |   |   txtLen > 164: 1 (2.0)
|   |   |   |   h1 > 0: 2 (5.0)
|   |   |   urlLen > 30: 1 (13.0/3.0)
|   txtKw > 0
|   |   Mlen <= 6
|   |   |   h2kw <= 1
|   |   |   |   Tlen <= 9
|   |   |   |   |   h1kw <= 0
|   |   |   |   |   |   Tkw <= 0
|   |   |   |   |   |   |   h3kw <= 1
|   |   |   |   |   |   |   |   txtKw <= 4
|   |   |   |   |   |   |   |   |   h2 <= 1: 1 (35.0/3.0)
|   |   |   |   |   |   |   |   |   h2 > 1
|   |   |   |   |   |   |   |   |   |   h1 <= 0: 1 (5.0)
|   |   |   |   |   |   |   |   |   |   h1 > 0
|   |   |   |   |   |   |   |   |   |   |   linkKw <= 2: 2 (4.0)
|   |   |   |   |   |   |   |   |   |   |   linkKw > 2: 1 (3.0/1.0)
|   |   |   |   |   |   |   |   txtKw > 4
|   |   |   |   |   |   |   |   |   txtKw <= 6: 2 (6.0/1.0)
|   |   |   |   |   |   |   |   |   txtKw > 6: 1 (2.0)
|   |   |   |   |   |   |   h3kw > 1: 2 (4.0/1.0)
|   |   |   |   |   |   Tkw > 0
|   |   |   |   |   |   |   h3 <= 9
|   |   |   |   |   |   |   |   h2len <= 2
|   |   |   |   |   |   |   |   |   h1len <= 0
|   |   |   |   |   |   |   |   |   |   Tlen <= 2: 2 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   Tlen > 2
|   |   |   |   |   |   |   |   |   |   |   linkKw <= 1
|   |   |   |   |   |   |   |   |   |   |   |   altKw <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   urlKw <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   alt <= 0: 2 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   alt > 0: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   urlKw > 0: 1 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   altKw > 0: 2 (2.0)
|   |   |   |   |   |   |   |   |   |   |   linkKw > 1: 1 (12.0/1.0)
|   |   |   |   |   |   |   |   |   h1len > 0: 2 (4.0)
|   |   |   |   |   |   |   |   h2len > 2: 2 (21.0/3.0)
|   |   |   |   |   |   |   h3 > 9: 1 (4.0)
|   |   |   |   |   h1kw > 0
|   |   |   |   |   |   Tlen <= 1: 1 (2.0)
|   |   |   |   |   |   Tlen > 1: 2 (43.0/9.0)
|   |   |   |   Tlen > 9: 2 (26.0/1.0)
|   |   |   h2kw > 1: 2 (31.0/2.0)
|   |   Mlen > 6
```

```
|   |   |   urlLen <= 46: 2 (27.0/2.0)
|   |   |   urlLen > 46: 3 (5.0/1.0)
Mkw > 0
|   Tkw <= 0
|   |   h1kw <= 1
|   |   |   alt <= 0: 1 (3.0/1.0)
|   |   |   alt > 0: 2 (42.0/4.0)
|   |   h1kw > 1: 3 (2.0)
|   Tkw > 0
|   |   h2 <= 0
|   |   |   h1 <= 3
|   |   |   |   txtKw <= 6
|   |   |   |   |   txtKw <= 0: 2 (5.0/1.0)
|   |   |   |   |   txtKw > 0
|   |   |   |   |   |   Mlen <= 3: 2 (3.0/1.0)
|   |   |   |   |   |   Mlen > 3
|   |   |   |   |   |   |   altKw <= 3
|   |   |   |   |   |   |   |   Tlen <= 14
|   |   |   |   |   |   |   |   |   Tkw <= 1
|   |   |   |   |   |   |   |   |   |   linkOut <= 21
|   |   |   |   |   |   |   |   |   |   |   Mkw <= 2
|   |   |   |   |   |   |   |   |   |   |   |   urlKw <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   linkKw <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   altKw <= 0: 3 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   altKw > 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   linkOut <= 11:
2 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   linkOut > 11: 3
(2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   linkKw > 0: 2 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   urlKw > 0: 3 (6.0)
|   |   |   |   |   |   |   |   |   |   |   Mkw > 2: 2 (2.0)
|   |   |   |   |   |   |   |   |   |   linkOut > 21: 2 (7.0)
|   |   |   |   |   |   |   |   |   Tkw > 1
|   |   |   |   |   |   |   |   |   |   txtLen <= 1407
|   |   |   |   |   |   |   |   |   |   |   Mkw <= 2
|   |   |   |   |   |   |   |   |   |   |   |   urlKw <= 1
|   |   |   |   |   |   |   |   |   |   |   |   |   alt <= 0: 2 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   alt > 0: 3 (8.0)
|   |   |   |   |   |   |   |   |   |   |   |   urlKw > 1: 2 (2.0)
|   |   |   |   |   |   |   |   |   |   |   Mkw > 2: 3 (7.0)
|   |   |   |   |   |   |   |   |   |   txtLen > 1407: 2 (3.0)
|   |   |   |   |   |   |   |   |   Tlen > 14: 2 (10.0/1.0)
|   |   |   |   |   |   |   |   altKw > 3: 3 (3.0)
|   |   |   |   |   txtKw > 6: 2 (7.0)
|   |   |   |   h1 > 3: 3 (3.0/1.0)
|   |   |   h2 > 0
|   |   |   |   h1 <= 0
|   |   |   |   |   Tlen <= 3: 2 (7.0)
|   |   |   |   |   Tlen > 3
|   |   |   |   |   |   linkOut <= 2: 2 (4.0)
|   |   |   |   |   |   linkOut > 2
|   |   |   |   |   |   |   Tlen <= 7
|   |   |   |   |   |   |   |   h2kw <= 2
|   |   |   |   |   |   |   |   |   h2kw <= 0
|   |   |   |   |   |   |   |   |   |   Tlen <= 6: 3 (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   Tlen > 6: 2 (2.0)
|   |   |   |   |   |   |   |   |   h2kw > 0: 3 (4.0)
|   |   |   |   |   |   |   |   h2kw > 2: 2 (4.0/1.0)
|   |   |   |   |   |   |   Tlen > 7: 3 (15.0/1.0)
|   |   |   h1 > 0
|   |   |   |   Mkw <= 1
```

```
|   |   |   |   |   h2kw <= 1
|   |   |   |   |   |   h3len <= 37
|   |   |   |   |   |   |   Tlen <= 9
|   |   |   |   |   |   |   |   Mlen <= 4: 2 (3.0)
|   |   |   |   |   |   |   |   Mlen > 4
|   |   |   |   |   |   |   |   |   altKw <= 1: 3 (12.0/3.0)
|   |   |   |   |   |   |   |   |   altKw > 1: 2 (3.0/1.0)
|   |   |   |   |   |   |   Tlen > 9: 2 (5.0)
|   |   |   |   |   |   h3len > 37: 3 (7.0)
|   |   |   |   |   h2kw > 1: 3 (10.0)
|   |   |   |   Mkw > 1
|   |   |   |   |   h2 <= 21: 3 (68.0/4.0)
|   |   |   |   |   h2 > 21: 2 (3.0/1.0)
```

Prilog br 4. Python skripta za ispravak stranica

```
import nltk
from nltk.corpus import stopwords
from nltk.corpus import wordnet
from nltk.stem import PorterStemmer
from transforms import filter_insignificant
from taggers import WordNetTagger

def main():
    # kreiraj xls datoteku koja će sadržavati rezultate
    import xlsxwriter
    workbook = xlsxwriter.Workbook('klasa1Ispravak.xlsx')
    worksheet = workbook.add_worksheet()

xlswrite=['pageID','Title','Tlen','Tkw','Meta','Mlen','Mkw','h1','h1Len','h
1kw','txt','txtLen','txtKw']
    col=0
    row=0
    for x in xlswrite:
        worksheet.write(0, col,  x)
        col+=1
    # učitaj podatke o uzorcima u klasi "neprilagođeno"
    from xlrd import open_workbook
    wb=open_workbook('klasa1.xls')
    for s in wb.sheets():
        # prođi kroz svaki redak (svaki uzorak)
        rowCounter=0
        for row in range(1,s.nrows):
            rowCounter+=1
            # za svaki redak čitaj podatke iz određenih stupaca
            kw=[] # ključne riječi
            kw1=s.cell(row,1).value
            kw=kw1.split(',')
            kw=kw[:-1]
            kw = [element.strip() for element in kw]
            kw = [element.lower() for element in kw]
            kw_clean=cleanText(s.cell(row,1).value,True)
            kw_clean = [element.strip() for element in kw_clean]
            kw_clean = [element.lower() for element in kw_clean]
            url=s.cell(row,2).value # url stranice
            title=s.cell(row,4).value # page title stranice
            Tlen=s.cell(row,5).value # dužina page titlea
            Tkw=s.cell(row,6).value # broj ključnih riječi u title
            meta=s.cell(row,7).value # tekst meta oznake
            Mlen=s.cell(row,8).value # dužina meta oznake
            Mkw=s.cell(row,9).value # broj ključnih riječi u meta oznaci
            h1=s.cell(row,11).value # tekst(ovi) h1 oznake (razdvojeni sa ;
ako ih ima više)
            h1len=s.cell(row,12).value # dužina h1
            h1kw=s.cell(row,13).value # broj ključnih riječi u h1
            text=s.cell(row,30).value # tekst stranice (bez HTML-a)
            txtLen=s.cell(row,31).value # dužina teksta
            txtKw=s.cell(row,32).value # broj ključnih riječi

            worksheet.write(rowCounter,0,s.cell(row,0).value) #spremi
pageID

            print("Procesiram PageID " + str(s.cell(row,0).value))
            # provjera što treba ispravljati i pozivanje funkcija za
ispravak
```

```
            # Najprije ispravi tekst jer je on ulaz u ostale funkcije
            noviText=''
            if txtLen>0:
                noviText=Obogati(text,kw,float(0.9),float(0.9))
                # spremi nove vrijednosti
                worksheet.write(rowCounter,10,noviText)
                worksheet.write(rowCounter,11,countwords(noviText))
                worksheet.write(rowCounter,12,
                       countkeywords(noviText,kw_clean))


            # Ispravak page titlea
            if Tlen<3 and txtLen>0:
                title=''
                if h1len>0:
                    # predloži h1 kao page title ako sadrži ključne riječi
                    # ako sadrži više h1, uzmi onaj s najvećom frekvencijom
ključnih riječi
                    # najprije obogati h1
                    h1text=Obogati(h1,kw,float(0.8),float(1))
                    if countkeywords(h1text,kw_clean)>0:
                        dijelovi=h1text.split(';')
                        rbrDio=0
                        maxC=0
                        rbr=0
                        for dio in dijelovi:
                            rbr+=1
                            cc=countkeywords(dio,kw_clean)
                            if cc>maxC:
                                maxC=cc
                                rbrDio=rbr
                        if rbrDio>0: title=dijelovi[rbrDio-1]
                if title=='':
                    # nije uspjelo izvlačenje iz h1 ili h1 ne sadrži
ključne riječi, dodaj ključne riječi na postojeće
                    title= ", ".join(kw)
            else:
                title=Obogati(title,kw,float(0.8),float(1))
            if countwords(title)>20: title=Skrati(title)
            # spremi nove vrijednosti
            worksheet.write(rowCounter,1,title)
            worksheet.write(rowCounter,2,countwords(title))
            worksheet.write(rowCounter,3,countkeywords(title,kw_clean))



            # Ispravak meta opisa
            if Mlen<5 and txtLen>0:
                meta=Sažimanje(noviText,kw,float(50),1)
            else:
                meta=Obogati(meta,kw,float(0.8),float(0.7))
            if countwords(meta)>30: meta=Skrati(meta,float(30))
            # spremi nove vrijednosti
            worksheet.write(rowCounter,4,meta)
            worksheet.write(rowCounter,5,countwords(meta))
            worksheet.write(rowCounter,6,countkeywords(meta,kw_clean))


            # Ispravak h1
            if h1len==0 and txtLen>0:
```

```
                h1text=Sažimanje(noviText,kw,float(50),2)
            else:
                h1text=Obogati(h1,kw,float(0.8),float(1))
            # spremi nove vrijednosti
            worksheet.write(rowCounter,7,h1text)
            worksheet.write(rowCounter,8,countwords(h1text))
            worksheet.write(rowCounter,9,countkeywords(h1text,kw_clean))

            continue
    workbook.close()

def Sažimanje (T, K, C, vrsta):
    # Sažimanje teksta T bazirana na upitu (ključne riječi K) i max.
dužinom C (br.riječi)
    # vrsta=1--> vrati summary ili vrsta=2--> vrati 1 rečenicu na javećom
frekvencijom klj.riječi
    import operator
    stemmer=PorterStemmer()
    stops=set(stopwords.words('english'))
    wnt=WordNetTagger()
    sent=nltk.sent_tokenize(T)
    freq={}
    idSent=0
    for s in sent:
        kFreq=0
        words=nltk.word_tokenize(s)
        words_s=[word for word in words if word not in stops]
        words_t=wnt.tag(words_s) # pos
        words_f=set(filter_insignificant(words_t,['JJ','UNKNOWN']))
        for w in words_f:
            w1 = stemmer.stem(str(w[0]).lower())
            for k in K:
                k1 = stemmer.stem(str(k).lower())
                if w1.lower() == k1.lower():
                    kFreq = kFreq + 1
        if kFreq>0: freq[idSent]=kFreq
        idSent=idSent+1
    freq_s = sorted(freq.items(), key=operator.itemgetter(1), reverse=True)
# sortiraj po važnosti

    summary=''
    if vrsta==1:
        for k in freq_s:
            if countwords(summary)<C:
               summary = summary + ' ' + sent[k[0]]
            else:
                break
    else:
        if len(freq_s)>0:
            sent_id=freq_s[0]
            summary=sent[sent_id[0]]

    return summary.strip()

def Obogati(T,K,S,C):
    # Obogaćivanje teksta T s ključnim riječima K uz graničnu sličnost S i
maksimalnom gustoćom C
    cT=cleanText(T,False)
    stemmer=PorterStemmer()
    # stemaj ključne riječi
    K1=[]
```

```
    for k in K:
        K1.append(stemmer.stem(k).lower())
    # spremi trenutnu gustoću
    ck=float(countkeywords(T,K1))
    cw=float(countwords(T))
    for f in cT: # za svaku riječ izračunaj sličnost
        # print("********** Ispitujem riječ: " + f)
        steamF=stemmer.stem(f).lower()
        for k in K:
            k=str(k).lower().strip()
            f=str(f).lower().strip()
            steamK=stemmer.stem(k).lower()
            if k!=f and len(f)>2 and steamF!=steamK:
                sin1=wordnet.synsets(f)
                sin2=wordnet.synsets(k)
                if sin1 and sin2:
                    sim=sin1[0].wup_similarity(sin2[0])
                    #print(sim)
                    if sim is not None:
                        if float(sim)>=S:
                            T=T.replace(f," [[ " + f + ", " + k + " ]] ",2)
                            print("Zamijenio sam riječ " + f + " s riječju
" + k)
                            ck+=T.count(" " + k + " ")
                # provjeri da li se premašila max. gustoća, ako je, izađi
iz petlje
                if float(ck/cw)>C: break

    return T


def cleanText(ctext,stemm):
    # tokenizacija, stemanje i otklanjanje zaustavne riječi
    words=nltk.word_tokenize(ctext)
    stops=set(stopwords.words('english'))
    words_s=[word for word in words if word not in stops]
    wnt=WordNetTagger()
    w_tags=wnt.tag(words_s) # pos
    filtered=set(filter_insignificant(w_tags,['JJ','VB','UNKNOWN'])) #
ostavi samo imenice
    stemmer=PorterStemmer()
    s=[]
    for f in filtered: #stemanje
        if stemm:
            s.append(stemmer.stem(f[0]))
        else:
            s.append(f[0])
    map(str.strip, s)
    return s

def countkeywords(wtext,KW):
    # broji koliko ključnih riječi sadrži tekst
    if len(wtext)>1000:
        wtext=wtext[0:1000]
    c=cleanText(wtext,True)
    brojac=0
    for k in KW:
        brojac+=c.count(k)
    return brojac

def countwords(wtext):
```

```
    # broji riječi u tekstu ignorirajući interpunkciju
    from nltk.tokenize import RegexpTokenizer
    tokenizer=RegexpTokenizer("[\w']+")
    return len(tokenizer.tokenize(wtext))

if __name__ == "__main__":
    main()
```

Prilog br. 5. Skup podataka ispravljenih stranica

ID;Kw;Tlen;Tkw;Mlen;Mkw;h1;h1len;h1kw;h2;h2len;h2kw;h3;h3len;h3kw;alt;altKw;linkKw;linkOut;urlLen;urlKw;txtLen;txtKw

20;Arts, Architecture, History, Organizations;3;0;6;1;1;1;0;2;15;0;14;33;0;5;0;3;16;27;0;227;3

23;Arts, Architecture, Landscape;7;2;39;2;4;8;2;2;5;0;6;13;0;5;0;1;35;28;0;239;3

24;Arts, Architecture, Preservation;3;1;52;1;0;0;0;0;0;0;0;0;0;0;0;1;32;31;1;178;2

31;Arts, Comics, Comic, Strips, Panels;5;6;16;2;1;11;2;0;0;0;0;0;0;0;0;2;1;33;0;78;2

33;Arts, Comics, Conventions;3;0;53;3;0;0;0;0;0;0;0;0;0;17;3;1;18;39;0;303;3

35;Arts, Comics, Fan, Pages;4;3;36;1;0;0;0;0;0;0;2;91;2;6;0;3;30;38;0;455;0

37;Arts, Comics, Manga, Fandom;4;3;33;0;0;0;0;0;0;0;0;0;0;0;0;0;0;32;0;156;0

43;Arts, Costumes, Exhibitions;4;0;42;2;1;4;0;0;0;0;0;0;0;3;0;4;7;34;0;313;5

45;Arts, Crafts, Candles;4;1;36;2;1;10;2;1;7;1;0;0;0;15;1;2;18;33;0;757;3

50;Arts, Education, Educators;3;0;38;4;0;0;0;0;0;0;0;0;0;0;0;5;29;0;631;5

51;Arts, Education, Educators;3;5;32;7;1;10;3;1;1;0;0;0;0;4;0;1;19;25;0;122;8

56;Arts, Humanities, Scholarship, Technology;4;4;26;4;1;1;0;1;2;0;0;0;0;9;0;2;4;37;0;109;4

58;Arts, Illustration, Airbrush;6;1;21;1;1;4;0;0;0;0;0;0;0;0;0;2;2;34;0;96;3

60;Arts, Performing, Arts, Acting, Actors, Actresses, Galvin, Anna;5;1;32;3;0;0;0;0;0;0;0;0;0;9;1;2;7;45;2;131;8

71;Business, Automotive, Import, Export;4;0;52;1;1;9;1;0;0;0;0;0;0;0;0;0;1;28;0;185;1

83;Business, Biotechnology, Pharmaceuticals;7;1;37;0;1;15;1;0;0;0;0;0;0;1;0;0;19;24;0;2354;0

84;Business, Biotechnology, Pharmaceuticals;10;1;0;0;0;0;0;7;51;0;3;22;0;3;1;0;54;17;0;489;0

86;Business, Biotechnology, Pharmaceuticals;16;1;51;2;24;48;1;0;0;0;0;0;0;8;1;1;58;19;0;1031;1

87;Business, Chemicals, Associations;6;0;42;0;1;6;0;1;2;0;11;3;0;9;0;2;166;49;0;732;2

105;Business, Construction, Maintenance, Restoration;5;1;44;4;1;4;1;0;0;0;4;7;0;0;0;2;52;21;0;495;3

112;Business, Education, Training, Manual, Skills;2;4;19;1;1;19;1;2;2;0;0;0;0;0;0;0;0;45;0;139;1

128;Business, Healthcare, Products, Services, Nutrition;4;0;52;2;1;0;0;6;14;0;9;55;0;1;1;1;64;28;0;401;0

134;Computers, Computer, Science, Database, Theory;5;7;0;0;1;3;0;0;0;0;0;0;0;2;0;0;10;38;0;277;0

135;Computers, Computer, Science, Database, Theory;3;0;44;3;1;8;0;0;0;0;0;0;0;0;0;0;14;21;0;281;1

139;Computers, Education, Courses;3;0;39;3;1;4;0;0;0;0;0;0;0;2;0;1;2;30;0;315;4

144;Computers, Hacking, Cryptography;3;2;30;3;1;1;0;6;19;0;0;0;0;1;0;0;9;44;1;692;1

145;Computers, Hacking, Cryptography;3;2;42;3;1;3;0;1;1;0;2;11;0;0;0;2;0;42;0;1513;2

149;Computers, Internet, Abuse;3;2;43;1;3;7;0;1;2;0;0;0;0;0;0;0;0;28;0;601;2

156;Computers, Open, Source, Software;4;1;53;1;7;22;1;0;0;0;0;0;0;2;3;0;8;26;0;494;1

158;Computers, Organizations, Non-Profit;4;2;0;0;7;12;0;13;35;1;1;1;0;5;0;0;15;23;0;285;0

159;Computers, Organizations, Non-Profit;9;0;62;1;1;16;2;0;0;0;0;0;0;7;0;1;8;48;1;225;3

160;Computers, Organizations, Non-Profit;9;0;51;3;1;6;0;7;36;1;23;78;3;4;0;2;40;25;0;1614;2

164;Computers, Programming, Contests;4;2;42;5;0;0;0;0;0;0;0;0;0;46;0;0;3;19;0;322;8

170;Computers, Robotics, Robots;3;5;36;6;1;22;4;0;0;0;0;0;0;0;0;0;2;28;0;336;6

186;Computers, Usenet, Software;4;0;44;1;1;23;1;13;56;0;4;5;0;22;0;0;143;30;0;3002;1

188;Computers, Virtual, Reality, Hardware;3;0;16;1;0;0;0;8;20;0;4;6;0;26;0;0;134;24;0;216;1

189;Computers, Virtual, Reality, Hardware;4;3;0;0;1;4;0;2;6;0;4;57;0;0;0;0;25;30;0;289;0

194;Games, Board, Games, War, Politics;4;0;48;2;1;8;1;0;0;0;0;0;0;12;0;2;2;30;1;244;4

196;Games, Card, Games, Trick, Capturing, Bridge;5;2;38;2;2;12;3;10;32;1;3;13;3;1;0;0;4;27;1;347;5

202;Games, Game, Studies, Conferences;4;6;34;1;0;0;0;0;0;0;0;0;0;0;0;0;2;33;1;229;2

204;Games, Hand, Games, Thumb, Wrestling;5;0;36;4;4;2;0;7;15;0;0;0;0;0;0;2;5;48;0;499;7

210;Games, Puzzles, Brain, Teasers, Riddles;6;0;39;3;1;1;0;0;0;0;0;0;0;1;0;0;2;31;1;355;3

222;Games, Video, Games, Simulation, Flight;3;2;37;3;1;3;2;0;0;0;0;0;0;13;1;0;1;39;0;497;3

224;Games, Video, Games, Word, Games;10;1;38;3;1;3;0;5;15;2;0;0;0;0;0;2;7;24;1;352;2

234;Health, Alternative, Herbs;3;0;38;2;1;3;1;0;0;0;0;0;0;1;0;2;1;31;1;427;2

236;Health, Alternative, Herbs;9;1;6;0;1;5;1;0;0;0;0;0;0;1;0;0;2;32;0;177;2

244;Health, Alternative, Reiki;3;0;58;1;3;8;0;22;106;0;0;0;0;12;0;0;4;31;1;4543;0

245;Health, Alternative, Reiki;3;0;0;1;1;4;1;0;0;0;0;0;0;0;0;0;0;11;41;1;695;1

247;Health, Alternative, Tibetan, Medicine;3;2;32;2;1;3;2;0;0;0;0;0;0;2;1;3;49;49;2;938;2

248;Health, Beauty, Hair;3;1;38;1;1;3;1;0;0;0;0;0;0;8;1;2;1;45;1;187;3

249;Health, Beauty, Salons, Spas;3;1;33;2;1;15;2;0;0;0;0;0;0;6;2;1;1;34;1;172;2

256;Health, Child, Health, Nutrition;4;2;35;2;1;6;2;0;0;0;0;0;0;12;1;2;1;23;0;187;7

257;Health, Child, Health, Special, Needs;11;1;33;2;0;0;0;0;0;0;1;2;0;3;0;1;25;24;0;451;3

263;Health, Conditions, Diseases, Eye, Disorders, Blindness;5;1;32;1;0;0;0;5;36;0;1;7;0;2;1;0;77;45;1;474;1

272;Health, Pharmacy, Nuclear, Pharmacy;9;1;32;2;1;10;1;0;0;0;0;0;0;0;0;0;7;13;44;1;113;8

274;Health, Public, Health, Safety, Emergency, Services;6;0;23;3;1;5;2;4;12;0;0;0;0;3;0;2;22;21;0;5309;4

277;Health, Senior, Health, Drugs;3;1;31;2;1;15;1;0;0;0;2;3;0;9;0;7;47;31;1;153;4

298;Home, Cooking, For, Children;8;1;19;2;0;0;0;1;4;0;1;3;0;18;1;1;5;56;0;147;2

303;Home, Family, Pregnancy;3;0;25;1;1;3;0;4;13;0;1;1;0;5;0;0;7;24;0;547;0

321;Home, Homeowners, Treehouses;9;0;19;1;1;4;0;2;7;0;0;0;0;4;1;1;6;34;1;424;0

329;Kids, Teens, Entertainment, Bands, Artists;5;0;20;1;1;2;0;3;11;0;6;19;0;1;0;0;24;21;0;324;1

331;Kids, Teens, Entertainment, Magazines, E-zines;6;4;5;0;1;0;0;1;4;0;26;154;0;0;0;0;7;21;0;785;2

342;Kids, Teens, School, Time, Homework, Help;6;5;0;0;1;8;0;5;23;0;1;6;0;15;0;0;7;28;0;148;0

360;Kids, Teens, Your, Family, Pets;3;0;43;2;0;0;0;0;0;0;0;0;0;0;12;0;1;10;51;2;1221;3

365;Recreation, Aviation, Model, Aviation;3;0;54;1;1;3;1;1;4;0;6;8;0;4;0;2;2;18;0;362;3

368;Recreation, Camps, Academic;3;2;378;1;0;0;0;0;0;0;2;4;0;0;0;0;18;33;0;846;1

369;Recreation, Camps, Academic;5;1;77;1;1;4;1;0;0;0;2;4;0;1;0;0;101;37;0;660;1

371;Recreation, Camps, Alumni;3;0;32;2;1;13;2;0;0;0;0;0;0;17;2;1;11;27;1;429;1

384;Recreation, Collecting, Stamps;5;1;34;2;1;5;1;1;1;0;0;0;0;0;0;1;213;33;1;1119;2

386;Recreation, Drugs, Cannabis;3;2;29;2;1;8;1;0;0;0;0;0;0;0;0;2;49;30;0;126;2

389;Recreation, Guns, Education;6;1;19;3;0;0;0;0;0;0;0;0;0;0;0;0;0;30;0;39;0

390;Recreation, Guns, Education;6;2;34;3;1;3;0;0;0;0;9;37;0;5;0;0;7;40;0;405;8

393;Recreation, Pets, Dogs;8;1;41;2;1;10;1;0;0;0;0;0;0;0;0;1;0;31;1;348;3

395;Recreation, Scouting, Campsites;3;3;56;1;1;2;0;1;8;0;0;0;0;3;0;0;7;62;0;249;2

396;Recreation, Scouting, Philanthropy;3;3;8;1;0;0;0;1;1;0;0;0;0;8;1;1;8;36;0;113;2

397;Recreation, Tobacco, Cigars;3;3;0;0;0;0;0;0;0;0;0;0;0;0;0;0;10;32;1;101;0

400;Reference, Almanacs, History;5;1;39;2;1;11;0;1;10;0;33;372;1;10;0;1;2;54;1;7172;1

404;Reference, Dictionaries, Pronouncing;4;1;37;1;1;4;2;0;0;0;4;27;1;0;0;1;10;45;0;419;2

405;Reference, Dictionaries, Rhyming;6;1;61;2;1;17;1;0;0;0;0;0;0;39;2;2;0;47;0;231;4

407;Reference, Encyclopedias, Subject, Encyclopedias, Paranormal;5;3;15;1;1;15;1;0;0;0;0;0;0;8;3;1;7;40;3;167;1

411;Reference, Flags, Vexillology;3;2;41;1;2;2;0;0;0;0;3;15;1;3;1;1;66;17;0;535;2

416;Reference, Libraries, Digital;3;0;14;2;1;7;1;4;28;0;6;16;0;36;0;3;3;23;0;412;2

417;Reference, Libraries, National;3;1;31;1;1;4;1;7;16;1;10;28;2;4;2;1;158;19;0;782;2

423;Reference, Museums, Cultural;3;1;69;1;1;2;0;11;36;0;1;3;0;7;0;1;26;34;1;730;1

424;Reference, Museums, Military;13;0;31;1;7;18;0;1;6;0;0;0;0;0;0;0;0;69;21;0;422;0

434;Science, Biology, Ecology, Ecosystems;3;0;43;2;0;0;0;0;0;0;0;0;0;0;9;0;1;3;46;0;745;1

439;Science, Biology, Evolution;2;1;33;1;1;2;1;0;0;0;3;25;1;1;0;0;6;47;0;1008;2

440;Science, Biology, Flora, Fauna;5;1;63;1;1;12;0;0;0;0;0;0;0;18;0;0;1;32;0;263;1

443;Science, Biology, Immunology, Antibodies;8;1;263;1;1;5;1;3;5;0;18;50;1;1;0;4;40;115;1;881;1

444;Science, Biology, Methods, Techniques;4;4;68;2;12;55;0;1;8;1;1;1;0;3;0;2;77;32;0;453;1

445;Science, Biology, Neurobiology;3;3;0;0;0;0;0;0;0;0;0;0;0;4;0;0;2;38;0;152;0

447;Science, Biology, Physiology;11;0;32;3;2;8;0;5;23;1;0;0;0;12;1;1;70;18;0;1407;2

448;Science, Chemistry, Chemists;3;3;4;1;1;4;1;0;0;0;0;0;0;2;0;0;3;28;0;58;2

450;Science, Chemistry, Organizations;7;0;85;1;1;9;1;3;9;0;18;61;1;8;0;1;27;30;0;433;1

454;Science, Math, Operations, Research;4;4;39;3;1;10;2;1;5;0;0;0;0;0;0;0;0;84;45;0;606;3

455;Science, Math, Operations, Research;4;0;31;3;1;4;0;7;48;0;0;0;0;1;0;3;18;59;1;1494;1

456;Science, Social, Sciences, Archaeology;10;3;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;49;0;0;0

462;Science, Technology, Materials, Ceramics;4;4;0;0;0;0;0;0;0;0;0;0;0;1;0;0;1;28;0;10;0

464;Science, Technology, Mining;9;1;42;3;1;3;1;1;5;0;49;116;2;0;0;1;21;86;0;1477;0

467;Science, Technology, Television;11;0;31;1;1;0;0;1;6;1;1;4;0;1;1;0;40;28;0;710;0

475;Shopping, Antiques, Collectibles, Clocks, Watches;3;0;21;1;1;2;0;0;0;0;0;0;0;0;0;0;0;34;2;90;4

489;Shopping, Clothing, Swimwear;3;1;38;1;1;7;1;0;0;0;0;0;0;1;0;0;2;29;0;499;0

503;Shopping, Music, Instruments;5;1;55;2;0;0;0;0;0;0;0;0;0;0;0;0;0;2;38;1;390;3

506;Shopping, Publications, Books, Military;4;0;38;3;1;8;2;0;0;0;0;0;0;3;1;4;15;25;0;338;4

516;Shopping, Vehicles, Aircraft;3;0;69;1;1;3;1;4;5;0;0;0;0;2;0;1;1;29;0;419;1

519;Shopping, Vehicles, Aircraft, Paramotor;8;1;25;0;1;1;0;14;34;0;8;11;0;1;0;0;126;42;0;567;4

520;Shopping, Vehicles, Aircraft, Parts, Accessories;5;4;18;2;6;196;0;6;22;0;0;0;0;0;0;1;22;25;0;369;0

521;Shopping, Vehicles, Aircraft, Parts, Accessories;4;1;226;1;0;0;0;0;0;0;0;0;0;0;1;0;25;0;659;1

540;Society, Law, Legal, Information, Drunk, Driving;3;0;34;0;12;56;3;5;7;1;4;8;2;10;4;5;40;50;1;1922;3

548;Society, Paranormal, UFOs;3;2;46;2;1;4;1;0;0;0;0;0;0;5;1;1;9;31;0;392;3

561;Sports, Basketball, Regional, Australia, Victoria;5;1;37;3;1;3;1;9;25;1;20;148;1;45;2;1;100;39;1;376;2

562;Sports, Basketball, Wheelchair;8;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0;29;0;0;0

571;Sports, Cue, Sports, English, Billiards, Associations;6;5;40;3;1;8;2;55;349;2;5;12;0;4;2;2;134;23;0;8485;2

576;Sports, Football, American, High, School;5;5;46;3;1;6;2;0;0;0;0;0;0;0;0;1;3;30;1;258;3

577;Sports, Football, American, High, School;5;2;40;1;1;4;1;1;2;0;9;10;1;0;0;2;102;44;0;336;1

578;Sports, Golf, Courses, Maryland;5;3;41;3;1;12;3;0;0;0;0;0;0;2;1;2;4;36;2;256;4

581;Sports, Gymnastics, Rhythmic;6;2;13;2;1;7;2;0;0;0;0;0;0;1;0;2;3;38;0;16;2

583;Sports, Lacrosse, Clubs;3;2;31;2;1;13;2;0;0;0;0;0;0;2;0;1;7;31;0;64;2

Prilog 6: R skripta za treniranje klasifikatora i prilagodbu hiperparametara

```
library(readxl)
#učitaj data set
podaci <- read_excel("podaci.xlsx")
#normaliziraj
podaci.norm <- as.data.frame(apply(podaci[, 1:21], 2, function(x) (x - min(x))/(max(x)-min(x))))
podaci.norm$klasa<-podaci$klasa

#pretvori klasu u factor
podaci$klasa<-factor(podaci$klasa)
podaci.norm$klasa<-factor(podaci.norm$klasa)

#machine learning - priprema
set.seed(1)
library(ParamHelpers)
library(mlr)
trainTask <- makeClassifTask(data = podaci,target = "klasa")
rCv10Strat <- makeResampleDesc("CV", iters = 10, stratify = TRUE) #10 folds cross validation stratificirani
rHoldStrat <- makeResampleDesc("Holdout", stratify = TRUE) #houldout metoda s 67% split i stratifikacijom

###################################
#stabla odlučivanja - WEKA j48
###################################
library(RWeka)
mytree <- makeLearner("classif.J48", predict.type = "response")
mymodel <- train(mytree,trainTask)
#mytree=setHyperPars(mytree, cp = 0.1)
#tuning hiperparametara C i M
#definiraj search space
num_ps = makeParamSet(
makeNumericParam("C", lower = 0.2, upper = 0.49),
makeIntegerParam("M", lower = 10, upper = 20))
#gird search
ctrl = makeTuneControlGrid()
res = tuneParams(mytree, trainTask, resampling = rCv10Strat,
```

```
            par.set = num_ps, control = ctrl, measures = acc)
#analiziraj rezultate tuninga
data = generateHyperParsEffectData(res)
plt=plotHyperParsEffect(data, x = "iteration", y = "acc.test.mean", plot.type = "line")
plt$labels$y="točnost"
plt$labels$x="iteracija"
plt
#vizualziraj s heat mapom
plt=plotHyperParsEffect(data, x = "C", y="M", z = "acc.test.mean", plot.type = "heatmap")
plt$labels$y="M"
plt$labels$x="C"
plt$labels$z="točnost"
plt
#primjeni parametre, treniraj i prikaži matricu
mytree=setHyperPars(mytree,par.vals=res$x)

#Cross validation 10 folds
rCV10<-resample(mytree,trainTask,rCv10Strat,measures=list(acc,mmce),models=FALSE)
print(rCV10)
print(calculateConfusionMatrix(rCV10$pred)) #konfuzijska matrica

#holdout
rHold<-resample(mytree,trainTask,rHoldStrat,measures=list(acc,mmce),models=FALSE)
print(rHold)
print(calculateConfusionMatrix(rHold$pred)) #konfuzijska matrica

#prikaži stablo
mymodel$learner.model

##################################
# SVM - ksvm iz kernlab
##################################
library("kernlab")
mySVM <- makeLearner("classif.ksvm", predict.type = "response")

#tuning hiperparametara C i Sigma
```

```
#definiraj search space
num_ps = makeParamSet(
  makeNumericParam("C", lower = -5, upper = 15, trafo = function(x) 10^x),
  makeNumericParam("sigma", lower = -10, upper = 5, trafo = function(x) 10^x)
)
#gird search
ctrl = makeTuneControlGrid()
res = tuneParams(mySVM, trainTask, resampling = rCv10Strat,
                 par.set = num_ps, control = ctrl, measures = acc)
#analiziraj rezultate tuninga
data = generateHyperParsEffectData(res)
plt=plotHyperParsEffect(data, x = "iteration", y = "acc.test.mean", plot.type = "line")
plt$labels$y="točnost"
plt$labels$x="iteracija"
plt
#vizualziraj s heat mapom
plt=plotHyperParsEffect(data, x = "C", y="sigma", z = "acc.test.mean", plot.type = "heatmap")
plt$labels$x="C"
plt$labels$y="sigma"
plt$labels$z="točnost"
plt

#primjeni parametre, treniraj i prikaži matricu
mySVM=setHyperPars(mySVM,par.vals=res$x)

#Cross validation 10 folds
rCV10<-resample(mySVM,trainTask,rCv10Strat,measures=list(acc,mmce),models=FALSE)
print(rCV10)
print(calculateConfusionMatrix(rCV10$pred)) #konfuzijska matrica

#holdout
rHold<-resample(mySVM,trainTask,rHoldStrat,measures=list(acc,mmce),models=FALSE)
print(rHold)
print(calculateConfusionMatrix(rHold$pred)) #konfuzijska matrica
```

```
##################################
# KNN - kknn
##################################
library("kknn")
myKNN <- makeLearner("classif.kknn", predict.type = "response")
#podaci moraju biti standardizirani za knn
trainTask <- makeClassifTask(data = podaci.norm,target = "klasa")
#tuning hiperparametara k i distance (minkowsi, 1=manhattan, 2=euclidean)
#definiraj search space
num_ps = makeParamSet(
  makeIntegerParam("k", lower = 1, upper = 100),
  makeNumericParam("distance", lower = 1, upper = 3)
)

#gird search
ctrl = makeTuneControlGrid()
res = tuneParams(myKNN, trainTask, resampling = rCv10Strat,
                 par.set = num_ps, control = ctrl, measures = acc)
#analiziraj rezultate tuninga
data = generateHyperParsEffectData(res)
plt=plotHyperParsEffect(data, x = "iteration", y = "acc.test.mean", plot.type = "line")
plt$labels$x="iteracija"
plt$labels$y="točnost"
plt

#vizualziraj s heat
plt=plotHyperParsEffect(data, x = "k", y="distance", z = "acc.test.mean", plot.type = "heatmap")
plt$labels$x="k"
plt$labels$y="p"
plt$labels$z="točnost"
plt

#primjeni parametre, treniraj i prikaži matricu
myKNN=setHyperPars(myKNN,par.vals=res$x)

#Cross validation 10 folds
```

```
rCV10<-resample(myKNN,trainTask,rCv10Strat,measures=list(acc,mmce),models=FALSE)
print(rCV10)
print(calculateConfusionMatrix(rCV10$pred)) #konfuzijska matrica

#holdout
rHold<-resample(myKNN,trainTask,rHoldStrat,measures=list(acc,mmce),models=FALSE)
print(rHold)
print(calculateConfusionMatrix(rHold$pred)) #konfuzijska matrica

##################################
# Bayes - NaiveBayes iz e1071
##################################
library("e1071")
myBayes <- makeLearner("classif.naiveBayes", predict.type = "response")
#podaci moraju biti standardizirani za knn
trainTask <- makeClassifTask(data = podaci,target = "klasa")

#tuning hiperparametara laplace
#definiraj search space
num_ps = makeParamSet(
  makeIntegerParam("laplace", lower = 1, upper = 1000)
)

#gird search
ctrl = makeTuneControlGrid()
res = tuneParams(myBayes, trainTask, resampling = rCv10Strat,
                 par.set = num_ps, control = ctrl, measures = acc)
#analiziraj rezultate tuninga
data = generateHyperParsEffectData(res)
plt=plotHyperParsEffect(data, x = "iteration", y = "acc.test.mean", plot.type = "line")
plt$labels$x="iteracija"
plt$labels$y="točnost"
plt

#vizualziraj s heat mapom
plt=plotHyperParsEffect(data, x = "iteration", y="laplace", z = "acc.test.mean", plot.type = "heatmap")
```

```
plt$labels$x="iteracija"
plt$labels$y="laplace"
plt

myBayes=setHyperPars(myBayes,par.vals=res$x)

#Cross validation 10 folds
rCV10<-resample(myBayes,trainTask,rCv10Strat,measures=list(acc,mmce),models=FALSE)
print(rCV10)
print(calculateConfusionMatrix(rCV10$pred)) #konfuzijska matrica

#holdout
rHold<-resample(myBayes,trainTask,rHoldStrat,measures=list(acc,mmce),models=FALSE)
print(rHold)
print(calculateConfusionMatrix(rHold$pred)) #konfuzijska matrica

###################################
# LiblineaRL2LogReg  - LiblineaR
# logistička reg. s L2
###################################
library("LiblineaR")
myLog <- makeLearner("classif.LiblineaRL2LogReg", predict.type = "response")
trainTask <- makeClassifTask(data = podaci,target = "klasa")

#tuning hiperparametara cost
#definiraj search space
num_ps = makeParamSet(
  makeIntegerParam("cost", lower = 1, upper = 10)
)

#gird search
ctrl = makeTuneControlGrid()
res = tuneParams(myLog, trainTask, resampling = rCv10Strat,
                 par.set = num_ps, control = ctrl, measures = acc)
#analiziraj rezultate tuninga
data = generateHyperParsEffectData(res)
```

```
plt=plotHyperParsEffect(data, x = "iteration", y = "acc.test.mean", plot.type = "line")
plt$labels$x="iteracija"
plt$labels$y="točnost"
plt

#vizualziraj s heat mapom
plt=plotHyperParsEffect(data, x = "iteration", y="cost", z = "acc.test.mean", plot.type = "heatmap")
plt$labels$x="iteracija"
plt$labels$y="cost"
plt

#primjeni parametre, treniraj i prikaži matricu
myLog=setHyperPars(myLog,par.vals=res$x)

#Cross validation 10 folds
rCV10<-resample(myLog,trainTask,rCv10Strat,measures=list(acc,mmce),models=FALSE)
print(rCV10)
print(calculateConfusionMatrix(rCV10$pred)) #konfuzijska matrica

#holdout
rHold<-resample(myLog,trainTask,rHoldStrat,measures=list(acc,mmce),models=FALSE)
print(rHold)
print(calculateConfusionMatrix(rHold$pred)) #konfuzijska matrica

######################
# BENCHMARK USPOREDBA
######################

lrns = list(mytree,mySVM,myKNN,myLog,myBayes)
rdesc <- makeResampleDesc("CV", iters = 10, stratify = TRUE)
meas = list(acc,mmce)
bmr = benchmark(lrns, trainTask, rdesc, meas)
bmr
getBMRPerformances(bmr)
library(ggplot2)
plt=plotBMRBoxplots(bmr, measure = acc, order.lrn = getBMRLearnerIds(bmr))
```

```
plt + ylab("točnost")


perf = getBMRPerformances(bmr, task.id = "podaci", as.df = TRUE)
df = reshape2::melt(perf, id.vars = c("task.id", "learner.id", "iter"))
df = df[df$variable == "acc",]
df = reshape2::dcast(df, task.id + iter ~ variable + learner.id)
library(GGally)
GGally::ggpairs(df, 3:7)
```

Prilog 7: R skripta za klasifikaciju ispravljenih stranica

```
library(readxl)

#učitaj novi data set (podaci za predviđanje)
podacinew <- read_excel("podaci_new.xlsx")
podacinew<-podacinew[,3:23] #izbaci prva dva stupca (id i kw)
#normaliziraj
podacinew.norm <- as.data.frame(apply(podacinew[, 1:21], 2, function(x) (x - min(x))/(max(x)-min(x))))

#podaci za treniranje
#pretvori klasu u factor
podaci$klasa<-factor(podaci$klasa)
podaci.norm$klasa<-factor(podaci.norm$klasa)

#machine learning - priprema
set.seed(1)
library(ParamHelpers)
library(mlr)
trainTask <- makeClassifTask(data = podaci,target = "klasa")

###################################
#stabla odlučivanja - WEKA j48
###################################
library(RWeka)
mytree <- makeLearner("classif.J48", predict.type = "response")
mytree=setHyperPars(mytree, C = 0.49, M=16)
mymodel <- train(mytree,trainTask)
pred = predict(mymodel, newdata=podacinew)
pred[["data"]][["response"]]

###################################
# SVM - ksvm iz kernlab
###################################
library("kernlab")
mySVM <- makeLearner("classif.ksvm", predict.type = "response")
```

```
mySVM=setHyperPars(mySVM, C = 7.74e+03, sigma=0.000464)
mymodel <- train(mySVM, trainTask)
pred = predict(mymodel, newdata=podacinew)
pred[["data"]][["response"]]

###################################
# KNN - kknn
###################################
library("kknn")
myKNN <- makeLearner("classif.kknn", predict.type = "response")
trainTask <- makeClassifTask(data = podaci.norm,target = "klasa")
myKNN=setHyperPars(myKNN, k = 45, distance=1)
mymodel <- train(myKNN, trainTask)
pred = predict(mymodel, newdata=podacinew.norm)
pred[["data"]][["response"]]

###################################
# Bayes - NaiveBayes iz e1071
###################################
library("e1071")
myBayes <- makeLearner("classif.naiveBayes", predict.type = "response")
trainTask <- makeClassifTask(data = podaci,target = "klasa")
mymodel <- train(myBayes, trainTask)
pred = predict(mymodel, newdata=podacinew)
pred[["data"]][["response"]]

###################################
# LiblineaRL2LogReg  - LiblineaR
# logistička reg. s L2
###################################
library("LiblineaR")
myLog <- makeLearner("classif.LiblineaRL2LogReg", predict.type = "response")
myLog=setHyperPars(myLog, cost = 2)
trainTask <- makeClassifTask(data = podaci,target = "klasa")
mymodel <- train(myLog, trainTask)
pred = predict(mymodel, newdata=podacinew)
```

```
pred[["data"]][["response"]]
```

ŽIVOTOPIS AUTORA

Goran Matošević rođen je 27.12.1977. god. u Puli, Republika Hrvatska. Osnovnu školu završio je u Sv.Lovreču a srednju Ekonomsku školu „M.Balote“ u Poreču. Nakon toga upisuje i 2003. godine završava sveučilišni studij smjer „Poslovna informatika“ na Ekonomskom fakultetu u Rijeci. Poslijediplomski znanstveni magistarski studij „Informatički management“ na Ekonomskom fakultetu u Zagrebu završava 2009. godine. Poslijediplomski doktorski studij informacijskih znanosti na Fakultetu organizacije i informatike u Varaždinu upisuje 2013. godine. U međuvremenu 2012. godine završio je program dopunskog pedagoško-psihološko obrazovanja na Sveučilištu J.J.Strossmayera u Osijeku, a 2013. god. program „E-learning akademije“ CARNET-a. Posjeduje certifikate: Microsoft certified professional, Microsoft certified application developer, Google adwords qualified individual. Govori hrvatski, engleski, njemački i talijanski jezik.

Od 2003. do 2008. godine radio je kao programer informacijskih sustava u tvrtci Edcom d.o.o. u Poreču, a potom godinu dana u Infobip d.o.o. Vodnjan. Kao vanjski suradnik radio je na SEO projektima za tvrtku Unline d.o.o. Pula. Od 2010. do 2011. godine radio je kao učitelj informatike u Srednjoj školi M.Balote u Poreču. Na Sveučilištu J.Dobrile u Puli počinje raditi 2011. godine gdje je izabran u zvanje asistenta iz područja društvenih znanosti, polje ekonomija, grana poslovna informatika. U zvanje predavača izabran je 2018. godine. Sudjeluje na izvođenju kolegija na Fakultetu ekonomije i turizma Dr.Mijo Mirković te na Fakultetu informatike u Puli.

Objavio je sljedeće znanstvene radove:

1) Matošević, Goran; Bevanda, Vanja. The comparison of social media usage in Croatian and UK SME companies. Conference proceedings of 4. international scientific conference / Merkač Skok, Marjana ; Cingula, Marijan (ur.). Celje: Faculty of commercial and business sciences, 2012. 182-190 (predavanje,međunarodna recenzija,objavljeni rad,znanstveni).
2) Matošević, Goran. The Adoption of Semantic Annotations of Products in Web Shops. International Journal of Computer and Communication Engineering. 3 (2014) , 1; 6-10 (članak, znanstveni).
3) Matošević, Goran. Towards a metric for on-page search engine optimization. CECIIS Proceedings - Central European Conference on Information and Intelligent Systems 2014 - 25th International conference / Hunjak, T., Lovrenčić, S., Tomičić, I. (ur.).

Varaždin : Faculty of Organization and Informatics, University of Zagreb, 2014. 194-199 (predavanje,međunarodna recenzija,objavljeni rad,znanstveni).

4) Bevanda, Vanja; Matošević, Goran. Mobilne aplikacije u turizmu. Suvremeni trendovi u turizmu / Gržinić, Jasmina ; Bevanda, Vanja (ur.). Pula : Sveučilište u Puli, 2014. Str. 1-20.
5) Matošević, Goran. Measuring the Utilization of On-Page Search Engine Optimization in Selected Domain. Journal of Information and Organizational Sciences. 39 (2015) , 2; 199-207 (članak, znanstveni).
6) Matošević, Goran. Using anchor text to improve web page title in process of search engine optimization. CECIIS Proceedings - Central European Conference on Information and Intelligent Systems 2015 / Tihomir Hunjak, Valentina Kirinić, Mario Konecki (ur.). Varaždin : Faculty of Organization and Informatics, University of Zagreb, 2015. 173-176 (predavanje,međunarodna recenzija,objavljeni rad,znanstveni).
7) Matošević, Goran. A review of massive online open courses features. New Possibilities of ICT in Education / Ružić Baf, Maja ; Žufić, Janko (ur.). Pula : Sveučilište Jurja Dobrile u Puli, 2016. Str. 1-15.
8) Matošević, Goran; Bevanda, Vanja. Mining Customer Behavior in Trial Period of a Web Application Usage-Case Study. Artificial Intelligence Perspectives in Intelligent Systems : Proceedings of the 5th Computer Science On-line Conference (CSOC2016). Vol. 1 / Silhavy, R., Senkerik, R., Oplatkova, Z.K., Silhavy, P., Prokopova, Z. (ur.). Switzerland : Springer International Publishing, 2016. Str. 335-346.
9) Matošević, Goran. Text Summarization Techniques for Meta Description Generation in Process of Search Engine Optimization. Artificial Intelligence and Algorithms in Intelligent Systems / Silhavy, Radek (ur.). Springer : Springer International Publishing, 2019. Str. 165-173.