Usporedba korijenovatelja, lematizatora i obilježivača vrsta riječi

Krušić, Lucija

Undergraduate thesis / Završni rad

2017

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka, Faculty of Humanities and Social Sciences / Sveučilište u Rijeci, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:186:496712

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-12-19**

Repository / Repozitorij:

Repository of the University of Rijeka, Faculty of Humanities and Social Sciences - FHSSRI Repository

Sveučilište u Rijeci – Odjel za informatiku

Preddiplomski studij engleskog jezika i književnosti i informatike

Lucija Krušić

Usporedba korijenovatelja, lematizatora i obilježivača vrsta riječi

Završni rad

Mentor: izv. prof. dr. sc. Sanda Martinčić-Ipšić, dipl. ing

Rijeka, rujan 2017.

Sadržaj

Sažetak

Korijenovatelji, lematizatori i obilježivači vrsta riječi su sastavni dio alata za obradu teksta i važan dio jezičnih tehnologija. Spadaju među korake obrade prirodnoga jezika koji se bave sintaksom (polje lingvistike koje proučava pravila i procese koji određuju strukturu rečenica). Obrada prirodnog jezika je interdisciplinarno područje koje se bavi interakcijom između računala i prirodnog jezika u smislu računalne obrade prirodnoga jezika. Cilj ovoga završnoga rada je usporediti dva korijenovatelja i lematizator te tri obilježivača vrsta riječi za engleski jezik kako bi se dobile informacije o njihovoj uspješnosti. U radu bit će prikazan teorijski pregled algoritama za korijenovanje i lematizaciju te pristupa za obilježavanje vrsta riječi. Porterov i Paice-Huskov korijenovatelj te lematizator su praktično uspoređeni na osnovi tekstova kao i Stanford, NlpDotCom i Genia obilježivač riječi. Rezultati su prikazani u postocima točnosti, s obzirom na količinu pogrešaka na ukupan broj riječi. Prema tome, WordNet lematizator je bio najuspješniji sa prosječno 76.34% ispravno pronađenih lema, dok je na drugom mjestu Paice-Huskov korijenovatelj sa prosječno 64.09% ispravno korijenovanih riječi a Porterov korijenovatelj je ispravno skratio 62.42% unesenih riječi. Između tri uspoređena obilježivača vrsta riječi, Stanfordov se pokazao najbolji sa 91.06% ispravno označenih riječi, sljedeći je bio Genia sa 87% i NlpDotNet sa 86.18%.

Ključne riječi : korijenovanje, lematizacija, obilježavanje vrsta riječi, obrada prirodnog jezika, računalna analiza prirodnog jezika

Comparative analysis of stemmers, lemmatizers and POS taggers for English language

Abstract

Stemming algortihms, lemmatizators and Part-of-speech taggers are a crucial part of text processing tools and an important part of speech technologies. They are part of the subcategory of tasks of Natural language processing that deals with syntax (field of linguistics that studies the rules and processes that determine the structure of sentences). Natural language processing is an interdisciplinary field of studies that deals with the interaction between human languages and computors. The main goal of this bachelor thesis is to compare two stemming algorithms with a lemmatizer as well as to compare three POS taggers for english language. The thesis includes a theoretical overview of stemming algorithms and lemmatizers as well as various approaches to POS tagging. The thesis also includes a practical comparison between the Porter stemmer, Paice-Husk stemmer and WordNet lemmatizator and between Stanford, NlpDotNet and Genia POS taggers. Furthermore, the success rates of Stanford's POS tagger, NlpDotNet tagger and Genia Tagger are measured. The results are displayed in percentages, based on the quantity of errors in a given text. The results show that the WordNet lemmatizator was the most successful with an average of 76.34%, followed by the Paice-Husk stemmer which correctly stemmed 64.09% of words and Porter stemmer which had a 62.42% success rate. Among the POS taggers, Stanford's tagger proved to be the most successful with 91.06% of correctly tagged words, followed by Genia with 87% and NlpDotNet with 86.18%.

Key words: stemming, lemmatization, POS tagging, Natural language processing

1. Uvod

Korijenovanje, lematizacija i obilježavanje vrsta riječi su postupci koji se koriste u okviru obrade prirodnog jezika (engl. Natural Language Processing, NLP), što je interdisciplinarno područje koje se bavi računalnom obradom prirodnog jezika kroz interakciju između prirodnog jezika i računala. Primjene NLP-a su između ostalog, ekstrakcija informacija iz prirodnog jezika, sumarizacija, pretraživanje informacija, kategorizacija teksta, generiranje prirodnog jezika i mnoge druge. Obrada prirodnog jezika u obzir uzima sintaksu, semantiku, diskurs i govor. Korijenovanje, lematizacija i obilježavanje vrsta riječi spadaju među korake koji se koriste u primjenama obrade prirodnog jezika vezanima za sintaksu (polje lingvistike koje se bavi gramatičkim odnosima između sastavnica rečenice). Korijenovanje se bavi skraćivanjem gramatičkih riječi na njihov „pseudo-korijen“ a lematizacija skraćuje riječi na njihov pravi korijen. Lematizatori, za razliku od korijenovatelja, rade morfološku analizu riječi da bi osigurali da je korijen (lema) pravilna riječ. Obilježavanje vrsta riječi je postupak kojim se riječima dodijeljuje oznaka koja ukazuje na njihovu vrstu (engl. word form, part of speech). Vrsta riječi je kategorija kojoj neka riječ pripada s obzirom na njezinu sintaksnu ulogu u rečenici (primjerice imenica, glagol, pridjev). U ovome završnom radu biti će proučeno više algoritama koji se koriste za korijenovanje i lematizaciju te poznate pristupe obilježavanju vrsta riječi za engleski jezik.

U nastavku završnog rada cilj je praktično usporediti poznate korijenovatelje i lematizatore za engleski jezik kako bi se analizirala njihova uspješnost i primjerenost te otkrilo koji najuspješnije pronalazi korijene riječi. Nadalje u radu će se usporediti i tri obilježivača vrste riječi za engleski kako bi se odredilo koji je najuspješniji, odnosno najprimjereniji za uporabu u jezičnim aplikacijama.

2. Teorijski opis korijenovatelja, lematizatora i obilježivača vrsta riječi

2.1. Korijenovatelji

Korijenovatelji (engl. stemmers, stemming algorithms) su jedan od alata računalne obrade prirodnog jezika koji se bave sintaksom prirodnoga jezika. Njihova svrha je uklanjanje sufiksa (nastavaka) od riječi i time stvaranje korijena riječi. Pretraživanje informacija (engl. information retreival) je postalo jedna od primjena računalne obrade prirodnog jezika (engl. natural language processing) te su korijenovatelji jedan od alata koji se koriste kao važan dio sustava za pretraživanje informacija.

Glavni ciljevi pretraživanja informacija su analiza i tretiranje dokumenata. Na primjer, da korisnik traži određeni dokument naslovljen sa: „How to drive?“ (engl. „Kako voziti?“), a u tražilicu upiše riječ „driving“ (engl. „voziti“), ako se koristi korijenovatelj, ta riječ će se skratiti na riječ „drive“, što je korijen te riječi i na taj način će korisnik lakše doći do traženog dokumenta. Korijenovanje se koristi za pretraživanje u brojnim jezicima te ovisno o jeziku poboljšava rezultate pretrage. Otkriveno je da u npr. hebrejskom jeziku, korijenovanje povećava broj pronađenih dokumenata za 10 do 50 puta, dok su u engleskom jeziku te brojke znatno manje. (Dennis).

Način na koji korjenovatelji rade je da traže korijen (engl. stem) tako što uklanjaju sufikse (nastavke) koji nose različita gramatička ili leksička značenja, od korijena riječi. Korijenovanje zapravo funkcionira tako da se „odreže“ drugi dio riječi, odnosno sufiks ili nastavak, u nadi da će se na taj način doći do korijena riječi. To je ponekad uspješno, dok ponekad preostali morfemi ne tvore pravi morfološki korijen već pseudo-korijen riječi. Rješenje za to je postupak nazvan lematizacija, nakon kojeg se dobiva ispravan, gramatički točan korijen kojeg se može pronaći u

rječniku i koji se s obzirom na postupak naziva lema. tvore ostale pomoću inflekcijskih sufiksa[1] (kao npr. „walk“ (engl. hodati) je lema, a riječi „walking“, „walked“ i „walks“ su tvorene od nje pomoću sufiksa). Po primjeru riječi „saw“ možemo lako uočiti razliku između korijenovanja i lematizacije, jer bi korijenovanje „odrezalo“ morfeme „aw“ a lematizator bi „odlučio“ je li ta riječ glagol „vidjeti“ u prošlom vremenu ili imenica, sa značenjem „pila“. U prvom slučaju bi lema bila „see“, a u drugome bi ostala „saw“. Time se dolazi do preciznijih rezultata.

Na temelju skraćivanja riječi na njihov korijen, korijenovatelji spajaju riječi u skupine ili jata (engl. cluster) s obzirom na njihove korijene. Pretpostavka je da ako dvije riječi imaju isti korijen, njihovo značenje mora biti srodno. Na taj način je mnogo lakše indeksirati dokumente prema ključnim riječima ali i proširiti upite tako da se dođe do preciznijih rezultata.

Primjeri teksta na kojem je obavljeno korijenovanje i lematizacija:

am, are, is $\Rightarrow$ be,
car, cars, car's, cars' $\Rightarrow$ car,

the boy's cars are different colors $\Rightarrow$
the boy car be differ color (Manning, 2008).

Istraživanje i razvoj algoritama za korijenovanje se u zadnjim desetljećima podijelio na dva pristupa- jedan je algoritamski dok je drugi lingvistički pristup uz koji se najviše veže lematizacija. Lingvistički pristup se bazira na semantičkim i leksičkim karakteristikama riječi koje se koriste da bi se riječi grupirale u skupine po sličnosti. Algoritamski pristup ne uzima u obzir lingvističke karakteristike riječi, kao što su rod, broj i vrijeme. S obzirom na to, događa se da stvoreni korijen nije ispravna riječ koja može prenositi značenje.

Međutim, korijenovatelji često proizvode i pogrešne korijene. Prva i najčešća greška je pretjerano skraćivanje (engl. overstemming) - do kojeg dolazi kada se dvije različite riječi, koje u stvarnosti nemaju isti korijen, svedu na isti korijen. Druga, potpuno suprotna, je kada se dvije riječi koje

[1] Sufiksi koji nose gramatičko značenje riječi. Dodavanjem tih sufiksa kao npr –s, -ed u engleskom jeziku gramatički se transformira osnovna riječ, mijenajući joj vrijeme ili rod, dok derivacijski sufiksi kao –ation ili –ous u potpunosti mijenjaju vrstu riječi

potječu od istog korijena skrate na različite te to zovemo greškama premalog skraćivanja (engl. understemming).

Korijenovatelje možemo podijeliti na tri vrste (Jivani, 2016): na korijenovatelje skraćivanja (engl truncating stemmers), statističke stemere (engl. statistical stemmers) i one s miješanim pristupom (engl. mixed stemmers).

2.1.1. Korijenovatelji skraćivanja

Među korijenovatelje skraćivanja ubrajaju se Porterov (Porter, Tartarus, 1980), Lovinsin (Porter, 2001), Paice-Huskov (Moral, 2014) i Dawsonov korijenovatelj (Jivani, 2016).

Porterov je jedan od najčešće korištenih korijenovatelja, a svakako jedan od najpoznatijih, koji se koristi posljednjih dvadesetak godina. Dizajnirao ga je Martin Porter, 1980. te je u početku bio kodiran u BCPLU-u („Basic Combined Programming Language“), netipiziranom programskom jeziku koji se danas rijetko koristi. Njegov cilj je otklanjanje morfoloških i inflekcijskih nastavaka, tj sufiksa riječima u engleskome jeziku kao dio procesa normalizacije. On se sastoji od pet faza skraćivanja riječi koje se odvijaju jedna za drugom, a svaka od njih se sastoji od seta pravila, kojima se iterativno odmiču sufiksi od riječi sve dok se nijedno od tih pravila više ne može primijeniti.

U Porterovom algoritmu, riječ se definira kao slijed parova samoglasnika i suglasnika - [C](VC)m[V]. U ovoj formuli C i V predstavljaju jedan ili više suglasnika (C-consonants) ili samoglasnika (V-vowels), a m je „mjera“ odnosno duljina (engl. measure) riječi, koja se otprilike bazira na broju slogova u riječi- cilj je saznati je li riječ dovoljno dugačka da bi odgovarala pravilu kao sufiks, a ne kao korijen riječi.

Na primjer, ako se mjera riječi (odnosno dio riječi koji pretpostavljamo da bi mogao biti njezin korijen) sastoji od više morfema nego sufiks „EMENT“ (m > EMENT), tada se taj sufiks uklanja.

Prema tome riječ „replacement“ se skraćuje na „replac“ ali riječ „cement“ se ne skraćuje na „c“ (Manning, 2008).

U pet koraka Porterovog algoritma sadržano je sveukupno oko 60 pravila, a ona osiguravaju da da sufiks i korijen zadovoljavaju određene uvjete.

Primjeri nekih pravila su :

(F)	Rule			Example		
	SSES	→	SS	caresses	→	caress
	IES	→	I	ponies	→	poni
	SS	→	SS	caress	→	caress
	S	→		cats	→	cat

Slika 1. Pravila Porterovog korijenovatelja (Manning, 2008)

Još jedan primjer djelovanja ovog stemera je skraćivanje riječi „generalizations“, koja bi se skratila prvo na „generalization“ odmičući –s koje označava množinu, zatim, na „generalize“, „general“ i napokon na „gener“ što bi bila stem te riječi. To bi se izvodilo kroz četiri faze tog algoritma (Moral, 2014).

U definiciji Porterovog algoritma prvo se navodi razlika između suglasnika i samoglasnika. Kao samoglasnike se definiraju sva slova koja nisu A, E, I, O, U. Također, ako je ispred slova Y u riječi samoglasnik, Y se smatra suglasnikom, a u obrnutom slučaju Y bi se smatrao samoglasnikom. Kao primjer toga priložene su riječi ili nizovi morfema TOY i SYZYGY. U prvoj riječi, Y bi bio smatran suglasnikom, odnosno samoglasnik bi bio O, dok u drugom nizu bi Y bio samoglasnik, a suglasnici S, Z i G. U narednom tekstu, samoglasnici će se označavati sa slovom v (engl. vowel) a suglasnici sa slovom c (engl. consonant) (Porter, Snowball Tatarus, 1980).

U bilo kojoj riječi koja sadrži više uzastopnih samoglasnika ili suglasnika, taj niz će se označavati sv oznakama V ili C. Dakle, c>0 =C, v>0=V.

Također, koriste se zagrade [ ] kako bi se pokazala proizvoljna prisutnost C ili V, na način: [C]VCVC ... [V].

Koristimo $(VC)^m$ da bismo pokazali da se niz VC ponavlja m puta. M predstavlja mjeru riječi ili njezin korijen. U slučaju da je m=0, to označava null ili nepostojeću riječ. Ta mjera riječi otprilike označava broj slogova od kojih je riječ sastavljena.

Primjeri:

m=0 TR, EE, TREE, Y, BY.

m=1 TROUBLE, OATS, TREES, IVY.

m=2 TROUBLES, PRIVATE, OATEN, ORRERY (Porter, Snowball Tatarus, 1980)

„M“ ovdje pokazuje koliko samoglasnika sadrži mjera.

Pravilo za otklanjanje sufiksa je - (uvjet) S1->S2.

To pravilo nam pokazuje da ukoliko je S1 sufiks riječi a korijen riječi zadovoljava postavljeni uvjet, tada sufiks S1 mijenjamo sa S2.

U ovom primjeru „(m>1)EMENT->“ – „EMENT“ je sufiks, te se on otklanja ako riječ zadovoljava uvjet da je mjera veća od jedan. Dakle, u riječi REPLACEMENT, bio bi otklonjen taj sufiks i ostao bi samo korijen „replac“ jer je to riječ kojoj je m=2, tj riječ od dva sloga.

Dodatne oznake su:

*S – korijen završava na S (ovdje može biti bilo koje slovo),

v - korijen sadrži samoglasnik,

*d – korijen završava sa dvostrukim suglasnikom (kao tt, ss),

*o – korijen završava sa cvc, odnosno suglasnik-samoglasnik-suglasnik, gdje drugi suglasnik nije W, X ili Y (Porter, Tartarus, 1980).

Uvjet može sadržavati i izraze „I“, „ILI“ i „NE“, tj konjunkciju, disjunkciju i negaciju. Primjeri tih uvjeta su :

(m>1 and (*S or *T)) – provjera da li korijen riječi sa m>1 završava na s ili t,

(*d and not (*L or *S or *Z)) - korijen riječi mora završavati na dvostruki suglasnik koji nije L, S ili Z.

U prvom koraku Porterovog stemera, algoritam se bavi otklanjanjem nastavaka za množinu i nastavaka koji se koriste za tvorbu prošlih glagolskih vremena i glagolskih pridjeva. Primjerice, riječ "“caresses“, što je treće lice jednine glagola se pretvara u korijen „caress“, koje nam govori da se nastavak „sses“ skraćuje u „ss“.[2] Pravilnim množinama se nastavak „s“ oduzima (kids -> kid) a imenice koje u jednini završavaju na y kao npr. „sky“ (množina „skies“) se skraćuju na „ski“.

caress -> caress
cats -> cat

Glagoli koji završavaju morfemima „eed“, ako im je mjera jednaka 0, to se skraćuje na „ee“ U ovim primjerima se vidi da riječ "feed" ima null samoglasnik, nema promjene. Na "agreed" je ta promjena vidljiva. U nastavku slijedi nekoliko primjera:

feed	->	feed
agreed	->	agree
plastered	->	plaster
bled	->	bleed
motoring	->	motor
sing	->	sing

[2] Pravila Porterovog korijenovatelja su svrstana u poglavlje „Prilozi“

Isto tako „bled“ se ne mijenja jer ispred sufiksa nije samoglasnik, dok se „r“ u slučaju riječi „plastered“ smatra samoglasnikom. Isto je i sa redukcijom „motoring“ na „motor“, dok se „sing“ ne reducira jer „s“ nije samoglasnik.

Nakon toga se ispravljaju greške koje su učinjene dosadašnjim skraćivanjem; Riječima koje bi inače završavale na samoglasnik „e“- a on je pogrešno maknut, zajedno sa nastavkom za prošlost „ed“- se taj samoglasnik vraća, a riječi koje su nastavkom „ed“ dobile i dvostruki suglasnik se taj suglasnik skraćuje (osim za riječi u kojima je suglasnik „l“, „s“ ili „z“). Ako riječ ima jedan slog i završava na suglasnik-samoglasnik-suglasnik (osim ako je drugi suglasnik „w“, „x“ ili „y“) korijenu se dodaje slovo „e“. Tako je riječi „filing“, koja je skraćena na korijen „fil“, pridruženo slovo e i tako tvori ispravna riječ „file“ dok riječ „failing“ je skraćena na „fail“ a s obzirom da korijen završava na samoglasnik-samoglasnik-suglasnik, riječi se ne dodaje „e“ i to je ispravni korijen.

Primjerice:

conflat(ed)	->	conflate
troubl(ed)	->	trouble
tann(ed)	->	tan
fall(ing)	->	fall
fail(ing)	->	fail
fil(ing)	->	file

U sljedećem koraku algoritma se provjerava riječi koje završavaju na “y” a mjera im sadrži samoglasnik. Taj se “y” pretvara u “i” kao u riječima:

happy -> happi

sky -> sky

U drugom koraku vidimo da se sve imenice i pridjevi kojima korijen ima jedan slog skraćuju na njihove prave korijene. Primjerice:

relational	->	relate
conditional	->	condition
hesitanci	->	hesitance
analogousli	->	analogous
vietnamization	->	vietnamize
predication	->	predicate
hopefulness	->	hopeful
callousness	->	callous
formaliti	->	formal
sensitiviti	->	sensitive
sensibiliti	->	sensible

U trećem koraku se isto tako imenice i pridjevi s jednim slogom skraćuju, međutim za razliku od prethodnog koraka kada so ih zamijenjivali sa različitim nastavcima, ovdje ima i primjera gdje se sufiks jednostavno izbriše- kao u „formative“-„form“.

triplicate	->	triplic
formative	->	form
formalize	->	formal
electriciti	->	electric
electrical	->	electric
hopeful	->	hope
goodness	->	good

U četvrtom koraku se od riječi koje imaju više od jednog sloga i završavaju nastavcima „ement“, „al“, „ance“, „ence“, „er“, „ic“, „able“, „ible“, „ant“, „ment“, „ent“, „ou“, „ism“, „ate“, „iti“, „ous“, „ive“ i „ize“ te riječi kojima je sufiks „ion“ a korijen im završava na „s“ ili „t“ odbacuju sufiksi. Time uglavnom dobivamo pravilne, gramatičke riječi, dok je u nekim slučajevima potrebno dorađivanje.

allowance	->	allow
inference	->	infer
gyroscopic	->	gyroscop
adjustable	->	adjust
defensible	->	defens
replacement	->	replac
adjustment	->	adjust
dependent	->	depend
adoption	->	adopt
homologous	->	homolog
effective	->	effect
bowdlerize	->	bowdler

U sljedećem koraku vidimo da se svim riječima koje imaju više od jednog sloga,a korijen im završava na „e“, oduzimaju sufiksi – u slučaju „rate“ to se ne događa, jer je to jednosložan korijen. Kada korijen ima jedan slog i nema slijed suglasnik-samoglasnik-suglasnik, slovo "e" na kraju nestaje.

probate	->	probat
rate	->	rate
cease	->	ceas

U posljednjem koraku, vidimo da se sve riječi čiji korijen ima jedan slog ili više, kojima korijen sadržava samoglasnik i završava na „l“, to dvostruko zadnje slovo smanjuje na jedno.

controll -> control
roll -> roll

Porterov stemer ima i određene nedostatke, jedan od kojih je naravno, to što povremeno pogriješi u pronalaženju ispravnog korijena riječi. Primjerice, riječ „general“ se krati na „gener“ a „iteration“ na „iter“, što nisu riječi u engleskom jeziku. Isto tako, događa se da korijenovatelj krati riječi na osnove koje u stvarnosti imaju u potpunosti različito značenje od kraćene riječi kao „doing“- „doe“ (engl. „raditi“- „golubica“) i „punish“ –„pun“ (engl. „kazniti“ – „pošalica“). To dovodi do pogrešnog spajanja (engl. conflating) riječi u skupine pa se gubi smisao korijenovanja. Isto tako, problem je što u potpunosti ignorira prefikse, pa se tako riječi poput „uncertain“ i „invisible“ ne krate na njihov korijen jer se prefiksi „un“ i „in“ ne skrate. Porter te dvije vrste pogreška dijeli u skupine koje naziva pretjerano korijenovanje (engl. overstemming) i pogrešno stemanje (engl. mis-stemming).

Porterov algoritam se i dalje široko koristi i implementiran je u mnogim programskim jezicima, između ostalog, u Haskellu, Lispu, Prologu, SQL-u, C-u, itd.

Prvi stemer koji je bio u uporabi, Lovinsin korijenovatelj, je izrađen 1968. godine i on se sastoji od dva koraka. Prvi korak je eliminacija sufiksa, a drugi nošenje sa preostalim korijenom. Korijenovatelj funkcionira tako da se sufiksi miču nakon uspoređivanja posljednjih morfema riječi s listom od 293 sufiksa, nakon čega se moraju zadovoljiti jedan ili više od 29 uvjeta i 35 pravila transformacije (Moral, 2014). Lovinsin korijenovatelj radi na tome da uklanja najduži sufiks riječi. Nakon micanja sufiksa, riječ se preoblikuje koristeći drugi skup tablica kako bi se stvorila ispravna riječ. Ovaj korijenovatelj je uspješan u rješavanju određenih problema, kao što je uklanjanje dvostrukog slova nastalog skraćivanjem infinitiva glagola, kao npr od "hitting" nastaje „hit“ a ne „hitt“ (Moral, 2014). Nakon toga se spajanje (engl. conflation), odvija tako da se u grupe riječi združuju korijeni koji su slični ali ne nužno i isti, jer se uzima u obzir i

pretpostavka da se u odmicanju sufiksa mogla dogoditi i pogreška. To vodi i do nepravilnosti, s obzirom da povećava broj riječi koje nemaju isti korijen a stavljene su u istu grupu.

Lovinsin korijenovatelj se sastoji od četiri dijela- A, B, C i D. U prvom dijelu se nalazi 294 nastavka odnosno sufiksa, sa slovom pokraj njih koje označuje uvjet na koji se referira (Aylett, Tatarus).

Primjer:

alistically	B
arizability	A
izationally	B
s	W
y	B

(Aylett, Tatarus).

Ovdje su sufiksi poredani od najdužeg ka najkraćem. U slijedećem koraku – B, predstavlja se 29 uvjeta[3] koje bi određeni sufiks (ovisno o slovu navedenom pokraj sufiksa u koraku A) trebao zadovoljiti.

Primjerice, uvjet G nam kaže da je minimalna dužina korijena jednaka tri slova i da se sufiks jedino odmiče ako mu prethodi slovo „f“.

Dio C je set od 35 pravila koja služe popravljanju zadnjeg slova u nađenom korijenu riječi, za slučaj da je previše odcijepano.

Primjer nekoliko njih:

1)	Bb	->	B	rubb[ing] -> rub
	Ll	->	L	controll[ed] -> control
	Mm	->	M	trimm[ed] -> trim

[3] Uvjeti Lovinsinog korijenovatelja se nalaze u poglavlju „Prilozi“

	Rr	->	R	abhorr[ing] -> abhor
2)	Iev	->	ief	believ[e] -> belief
3)	Uct	->	uc	induct[ion] -> induc[e]
4)	umpt	->	um	consumpt[ion] -> consum[e]
5)	Rpt	->	Rb	absorpt[ion] -> absorb
6)	Urs	->	Ur	recurs[ive] -> recur
7a)	metr	->	meter	parametr[ic] -> paramet[er]
8)	Olv	->	olut	dissolv[ed] -> dissolut[ion]

(Porter, Tartarus, 1980)

Prva linija prvog pravila nam pokazuje da ako korijen nakon otklanjanja sufiksa i dalje završava dvostrukim „bb", jedno slovo „b" nestaje te tako nastaje pravilna riječ.

Četvrti, D dio, se i ne mora smatrati kao iznimno važan za rad korijenovatelja, jer se u njemu navodi samo par pravila vezanih za upite (engl. query terms) i indekse.

Pozitivne strane ovog korijenovatelja su da se dobro nosi s dvostrukim slovima i s nepravilnim množinama. Negativne strane su da je spor te da mnogi sufiksi nisu dostupni u tablici sa nastavcima, što dovodi do čestih nepravilnosti i nepouzdanosti korijenovatelja (Jivani, 2016).

Treći korijenovatelj koja valja spomenuti je Paice-Huskov korijenovatelj. On se sastoji od jedne tablice sa 120 pravila koja su sva indeksirana sa zadnjim slovom sufiksa. Ovaj korijenovatrelj izvodi algoritam iterativno, tj. na svakoj iteraciji pokušava naći odgovarajuće pravilo pomoću zadnjeg slova sufiksa- tada se sufiks ili uklanja ili mijenja. U slučaju da takvo pravilo ne postoji, dolazi do prekida izvođenja algoritma. Do prekida dolazi i ako riječ počinje samoglasnikom i ima još dva slova te ako počinje suglasnikom i ima četiri slova sveukupno. Velika prednost ovog korijenovatelja je njegova jednostavnost, dok je veliki nedostatak to što često može doći do pretjeranog korijenovanja (Jivani, 2016).

Downsov korijenovatelj je još jedan korijenovatelj koji radi na principu skraćivanja riječi- on je nadograda Lovinsinog korijenovatelja, jer radi na istom principu, samo što ima puno veći popis

sufiksa-čak njih 1200. Vrlo se brzo izvodi, međutim vrlo je složen i nema standardne implementacije (Jivani, 2016).

2.1.2. Statistički korijenovatelji

Sljedeća vrsta korijenovatelja su oni koji su zasnovani na statističkim analizama i tehnikama te se stog nazivaju statističkim korijenovateljima. Prvi od te skupine je N-gram korijenovatelj, koji funkcionira tako da se nizovi dužine n slova izdvajaju iz teksta, odnosno iz riječi. Prednosti ovog pristupa su što je neovisan o jeziku, dok je nedostatak nepraktičnost zbog velike količine memorije koje n-grami moraju zauzimati. N-gramski modeli funkcioniraju na način da izračunavaju vjerojatnost pojavljivanja nekog niza znakova w_n ako mu prethodi niz w_{n-1} (Martinčić-Ipšić, 2007).

Kod korijenovanja n-grami su nizovi znakova koja se pojavljuju jedni za drugima, a koji se nalaze unutar neke riječi. Ideja takvih korijenovatelja je taj da će slične riječi imati veći broj zajedničkih n-grama. Bigrami su pritom nizovi dva znaka, dok su trigrami nizovi triju znakova. Na primjer, riječ „idiom“ se sastoji od bigrama *I, ID, DI, IO, OM; te od trigrama **I, *ID, IDI, IOM, OM*, M**. Pritom „*“ označava prazan prostor. Dakle, ako se riječ sastoji od n znakova, postoje n+1 bigrama i n+2 trigrama unutar te riječi (Jivani, 2016).

HMM korijenovatelj je zasnovan na Skrivenim Markovljevim modelima (engl. Hidden Markov Models), što znači da je zasnovan na konačnim automatima gdje su prijelazi određeni s funkcijama vjerojatnosti (Jivani, 2016). Skriveni Markovljevi Modeli funkcioniraju na način da je u centru pretpostavka koja kaže da neki događaj i njegova vrijednost (s_t) ovisi isključivo o vrijednosti prethodnog događaja (s_{t-1}). To se naziva Markovljevom pretpostavkom i ona nam govori o ovisnosti stanja u odnosu isključivo na prethodno (Martinčić-Ipšić, 2007). Ona glasi

$$P(s_t \mid s_1^{t-1}) = P(s_t \mid s_{t-1}), \quad (1)$$

te zajedno sa pretpostavkom o neovisnosti simbola X uvjetuje SMM-e prvoga reda. Pretpostavka o neovisnosti simbola X govori da je u slijedu izlaznih simbola, simbol X opažen u vremenu t

ovisan samo o stanju s_t i uvjetno ovisan o prošlim izlazima. Formula za pretpostavku o neovisnosti simbola glasi :

$$P(X_t \mid X_1^{t-1},\ s_1^t) = P(X_t \mid s_t).$$ (Martinčić-Ipšić, 2007) (2)

Skriveni Markovljevi modeli su definirani izlaznom abecedom, skupom stanja, matricom vjerojatnosti prijelaza, matricom početnih stanja i vektorom početnih vjerojatnosti. On kroz četiri koraka radi: inicijalizaciju u kojoj se odabire početno stanje s obzirom na početnu distribuciju vjerojatnosti i postavlja vrijeme, generiranje izlaznog simbola s obzirom na vjerojatnost izlaznog znaka, prijelaz u novo stanje u skladu s vjerojatnostima prijelaza te završetak u kojem se ili ponavlja sve od drugog koraka ovisno o vremenu t ili se završava (Martinčić-Ipšić, 2007).

Primijenjeno na algoritme korijenovanja, riječ je sekvenca slova, tj riječ se smatra nizom dva podniza – prefiksa i sufiksa. Ovaj proces se modelira kroz SMM tako da se stanja podijele u dva odvojena niza gdje bi prvi mogao biti sastavljen samo od korijena a drugi ili od korijena ili sufiksa. Promjene stanja na taj način određuju sastavljanje riječi. Početno stanje može biti pridruženo nizu korijena, pretpostavljajući da riječ uglavnom započinje korijenom. Tranzicije između stanja sufiksa i stanja korijena uvijek imaju null vjerojatnost s obzirom da riječ uvijek mora biti niz korijena i sufiksa. SMM model može za bilo koju riječ pronaći točku razdvajanja sufiksa i korijena pomoću najvjerojatnijeg puta te se tada niz znakova prije te točke razdvajanja smatra korijenom (Jivani, 2016).

Ova metoda korijenovanja ne zahtijeva prethodno poznavanje skupine podataka koju se korijenuje te ne ovisi o jeziku što su njezine prednosti. Međutim, vrlo je kompleksna za implementaciju a osim toga i često dolazi do pretjeranog skraćivanja riječi.

2.1.3. Metode korijenovanja s miješanim pristupom

Među ove metode spadaju korijenovatelji koji rade na principu flektivnih i derivacijskih pravila (engl. inflectional and derivational methods), kao što su Xeroxov (Jivani, 2016) i Krovetzov (Jivani, 2016), korijenovatelji zasnovani na korpusima (engl. corpus-based) i oni koji su osjetljivi na kontekst (engl. context-sensitive).

Krovetzov korijenovatelj je algoritam koji se bazira na sintaksi i na inflekcijskoj vrijednosti riječi. Bavi se uklanjanjem sufiksa –s u množini imenica, -ed za prošlo vrijeme i –ing u infinitivnim oblicima te ih uspješno otklanja stvarajući jedninu, sadašnje vrijeme itd. Korijenovatelj radi tako da najprije otklanja sufiks a onda kroz korpus riječi pretražuje postojeće riječi i pronalazi odgovarajući korijen. Velika prednost ovoga korijenovatllja je ta što sa točnošću proizvodi morfološki ispravne riječi, nosi se s neispravnostima a i otklanja i prefikse i sufikse. Ovaj korijenovatelj, nažalost ima par nedostataka, izrazito je spor ako se unese dugačak tekst koji treba analizirati te ne može obraditi riječi koje nisu sadržane u leksikonu. Osim toga, leksikon se mora ručno napraviti, što iziskuje mnogo vremena (Jivani, 2016).

Xeroxov inflekcijski i derivacijski analizator je leksička baza podataka koja može analizirati i generirati inflekcijske i derivacijske nastavke. Radi odlično s dugačkim dokumentima, osigurava da su svi pronađeni korijeni prave riječi te se uspijeva nositi i s prefiksima. Nedostatak ovoga korijenovatelja je ovisnost o leksikonu i riječima koje sadrži, stoga se ne može nositi s riječima koje su izvan njega (engl. out of vocabulary words) (Jivani, 2016).

Metode koje se zasnivaju na korpusima i na kontekstu će biti ukratko opisane.

Metoda korijenovanja zasnovana na korpusu je razvijena radi pogreška u spajanju (engl.conflation) riječi u grupe, koje je proizvodio Porterov korijenovatelj a on nastoji izbjeći i stvaranje korijena koji nisu istinske riječi u riječniku. Primjeri toga su spajanje riječi „policy“ i „police“ u grupu srodnih riječi, premda one to nisu („policy“ znači pravilo, zakon dok je „police“ policija) te skraćivanje riječi „general“ u „gener“, što nije ispravna riječ. Ovako se pomoću Porterovog i ponekad Krovetzovog korijenovatelja pronalaze korijeni te se onda statističkim

metodama provjerava jesu li riječi pravilno grupirane po značenju te da su svi korijeni pravilne riječi. Pozitivne strane ove metode su riješevanje problema grupiranja i nepravilnih korijena. Međutim, za svaki korpus bi se onda trebala izraditi nova statistička mjera što iziskuje više vremena.

Posljednja metoda su korijenovatelji koji ovise o kontekstu. To su složeni korijenovatelj koji kroz četiri koraka poboljšavaju pronalazak riječi na temelju upita (engl. query) te pronalaženje traženih riječi u dokumentu. Kao osnovu koriste Porterov korijenovatelj kako bi proizveo korijene za svaku riječ upita, pronalazi glavne riječi u imeničkim frazama kako bi otkrio glavni koncept nekoga upita i kako bi se otkrio kontekst te to radi lomljenjem upita na segmente. Nadalje, pomoću n-grama traže oblike glavnih riječi koji bi mogli biti najkorisniji, što su uglavnom množine ako se radi o imenicama a nakon toga se kontekst upita uspoređuje sa temom određenog dokumenta kako bi se otkrilo da li ga pokazati. Ova metoda, osim što je jako kompleksna te zato iziskuje mnogo vremena, ponekad radi i pogreške prilikom otkrivanja imeničkih fraza.

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Slika 2- Usporedba korijenovatelja (Manning, 2008)

2.2. Lematizacija

Lematizacija se definira kao postupak svođenja riječi na njihov osnovni oblik. Taj osnovni oblik, se još naziva i „lema“. Lema je lingvistički ispravan kanonski oblik riječi te je lematizacija različita od korijenovanja po tome što ona uvijek proizvodi morfološki ispravan korijen, dok se korijenovanjem rijeći svode na tzv. „pseudokorijen“ uklanjanjem afiksa. (Pandžić, 2015) Lematizatori su jedan od alata računalne obrade prirodnog jezika, koji za razliku od korijenovatelja rade morfološku analizu riječi da bi osigurali da je osnova riječi ili lemma prava riječ. To dovodi do puno kvalitetnijih rezultata nego korjenovanje, međutim niti to ne otklanja sve pogreške s obzirom da postoje određene kolokacije za koje nije dovoljno znati samo morfološko već i pragmatičko značenje riječi da bi ih pravilno razumijeli i normalizirali (Šnajder, 2011).

Lematizacija je ovdje proces koji stvara leme neke leksičke baze podataka. Ako uzmemo za primjer riječi „paying“, „paid“ i „pays“, lema od tih riječi je riječ „pay“ (platiti). Također, od riječi „is“, „were“ i „am“ – lema je glagol „be“ (biti). Korijenovatelj bi riječ „paid“ vjerojatno netočno skratio na „pai“ dok je lematizator uspješno prepoznaje da ona potječe od riječi „pay“. Za razliku od korijenovanja, lematizacija teži grupiranju riječi po njihovom značenju – te teži otklanjanju dvosmislenosti koje su česte u engleskom jeziku, primjerice: riječ „wake“ se može odnositi na glagol „wake up“ (probuditi se) ili na imenicu koja znači „pogreb“ (Elastic). U slučaju takvih riječi, kada bi se oslanjali samo na korijenovanje, riječ bi vjerojatno bila neispravno skraćena.

U hrvatskom jeziku je veća razlika između korijenovatelja i lematizatora nego u engleskome. To je stoga što se hrvatskim riječima otklanjanjem afiksa rijetko može doći do ispravnoga korijena, dok je u engleskom jeziku to puno češće. Zato se korijenovatelji za engleski jezik kao što su Krovetzov i Xeroxov mogu smatrati i lematizatorima, jer pripadaju skupini koja koristi flektivna i derivacijska pravila za korijenovanje a osim toga su zasnovani na leksičkoj bazi podataka i velikim korpusima engleskih riječi.

Primjer:

> She sat and sold shells on the sea shore. The little boy watched the lonely seagull disappear in the distance and cried. Old women told tales of times long past. –> originalni tekst
>
> She sit and sell shell on the sea shore . The little boy watch the lonely seagull disappear in the distance and cry . Old woman tell tale of time long past . –> tekst obrađen lematizatorom
>
> Korišteni alat: (Text analysis Online, 2016).

Iz primjera vidimo da su sve riječi svedene na lemu, svi glagoli su vraćeni u sadašnje vrijeme na točan način, a ne samo oduzimanjem nastavaka, što se pogotovo vidi na primjeru nepravilnih glagola "sat"-"sit" i "tell"-"told". Riječ "women" je stavljena u jedninu kao "woman".

Taj isti tekst proveden kroz Porterov algoritam bi izgledao ovako:

> She sat and sold shell on the sea shore The littl boi watch the lone seagul disappear in the distanc and cri. Old women told tale of time long past.
>
> Korišteni alat: (NTLK, 2017)

Iz toga vidimo da se riječ „sat'' nije promijenila, a većina ostalih riječi je netočno skraćena, kao npr. „distanc'' i „cri“ od „cried''. Riječ „times'' je ispravno skraćena na „time'' iz množine . Riječ „women“ nije stavljena u jedninu jer je to nepravilna množina što Porterov algoritam ne raspoznaje.

2.3. Obilježavanje vrsta riječi za engleski jezik

Proces obilježavanja vrsta riječi (engl. Part of speech tagging), koji se još naziva i gramatičkim obilježavanjem je označivanje riječi u zadanom tekstu njihovim pripadajućim oznakama (engl. tag). Svakoj riječi se pritom pridaje njezina oznaka ovisno o vrsti riječi i njenoj ulozi u rečenici (ako je ulazni tekst formiran u rečenice). Obilježivači vrsta riječi (engl. POS taggers) su alati kojima se olakšava obrada prirodnog jezika (engl. natural language processing). Obrada prirodnog jezika je polje informacijskih znanosti koje se bavi interakcijom između računala i prirodnog jezika (Algorithmia, 2016).

Ulazni podaci koji se unose u obilježivače vrsta riječi su ili samo skupine riječi ili tekstovi formirani u rečenice i paragrafe te skupina oznaka (engl. tagset). Izlazni podatak koji dobivamo je optimalna oznaka vrste riječi za svaku riječ na ulazu. Vrste riječi mogu biti primjerice, imenice, glagoli, pridjevi, atributi, itd. U engleskom jeziku riječi su podijeljene u dvije kategorije- otvorena skupina vrsta riječi i zatvorena. Razlika između te dvije kategorije je ta što se u otvorenu skupinu riječi mogu dodavati nove riječi (koje ulaze u riječi iz drugih jezika, tj. posuđenice ili primjerice novoizmišljeni termini), dok su u zatvorenoj kategoriji uglavnom funkcijske riječi pa se toj skupini ne mogu dodavati nove riječi. Zatvorena kategorija riječi u engleskom se sastoji od prijedloga (engl. prepositions), kao što su preko, na, ispod (engl. over, on, under); čestica kao što su umjesto, bez (engl. instead, without); članova (kao npr. a, an, the); konjunkcija kao što su i i ili (engl. and, or); zamjenica i pomoćnih glagola (kao npr. can, may, should).

Otvorena kategorija riječi prihvaća dodavanje novih riječi, a sastoji se od imenica, koje mogu biti osobne i opće, glagola, pridjeva i priloga.

Sve te vrste riječi moraju biti označene pripadajućim oznakama kao i njihove potkategorije, npr. je li prilog mjesni ili vremenski.

Arhitektura obilježivača riječi je zasnovana na tri radnje-na tokenizaciji ili etiketiranju, pri čemu se cijelom tekstu (svakoj riječi i inetpunkcijskim znakovima) dodijele etikete, provjera

dvosmislenosti (koristi se analiza leksikona) i rješavanje dvosmislenosti (na temelju konteksta ili učestalosti uporabe neke riječi-npr. češće se koristi „power“ kao imenica nego kao glagol).

Naravno, i metoda obilježavanja ima svoje nedostatke. U engleskom jeziku to se uglavnom odnosi na višeznačnost nekih riječi, npr. riječ „show“ koja može biti i imenica i glagol. Cilj obilježivača je da tu riječ obilježi u kontekstu teksta u kojem se nalazi, stvarajući pretpostavke o riječima koje se nalaze u neposrednoj blizini analizirane riječi.

Glavni tipovi obilježivača su stohastički i obilježivači zasnovani na pravilima (Dresen, 2006). Stohastički se zasnivaju na maksimalnoj vjerodostojnosti (engl. likelihood) te na Skrivenim Markovljevim modelima (osnove spomenute u poglavlju 2.1.2. Statistički korijenovatelji) koji su jedan od glavnih stohastičkih pristupa obilježavanju. Pomoću njih određuje se najvjerojatnija oznaka za svaku riječ u rečenici na osnovu prethodnih ili narednih riječi.

S druge strane, obilježavanje temeljeno na pravilima radi pomoću formule:

Ako <neki uzorak> Onda ... <neka oznaka vrste riječi> (Dresen, 2006).

Pritom je omogućeno da se više puta prolazi kroz ovu petlju. Obilježivači zasnovani na pravilima rade na principu uzimanja pravila direktno iz rječnika te se odabire pravilna oznaka pomoću zapisivanja svih mogućih ograničenja i uvjeta sve dok se ne nađe samo jedna moguća oznaka. Neka od tih pravila su primjerice, da determinator ne može prethoditi glagolu već samo imenici i da glagolskim frazama prethode pomoćni glagoli, ali ne i obrnuto.

Obilježavanje temeljeno na pravilima je najstarija metoda obilježavanja te su sva pravila ručno napisana i koriste se kada postoje dvojbe o kategoriji kojoj riječ pripada- onda se analiziraju njezina lingvistička svojstva i donosi odluka o oznaci. Jedan od primjera pravila jest da ako je prethodna riječ član (a/an/the), onda riječ koju označujemo mora biti imenica. Jedan takav obilježivač je TAGGIT (Jurafsky & Martin, 2016) koji koristi kontekstno ovisna pravila, njih 3300 i 71 oznaku. On može uspješno obilježiti 77 posto riječi u korpusu Sveučilišta Brown (Robin, 2009).

2.3.1. Obilježavanje pomoću SMM-a

Za obilježavanje vrsta riječi upotrebom SMM-a koriste se modeli vjerojatnosti (engl. probabilistic models). Što znači da za riječi $w_1, ..., w_n$ nalazimo najvjerojatniji set oznaka $t_1,...t_n$. Uz to, „najvjerojatniji“ znači najčešće spomenut u određenom, promatranom korpusu (Meyers, 2012). Prvo se odabire oznaka dužine n, koja je najvjerojatnija s obzirom na uneseni niz znakova te se onda koristi Bayesov teorem –

$$P(A|B) = \frac{P(A)\ P(B|A)}{P(B)} \tag{3}$$

koji opisuje vjerojatnost nekog događaja zasnovan na prethodnom poznavanju uvjeta koji mogu biti povezani za taj događaj). Pri čemu su P(A) i P(B) vjerojatnost događaja A i vjerojatnost događaja B. P(A|B) je vjerojatnost događaja A ako je nastupio događaj B, a P(B|A) je vjerojatnost događaja B ako je nastupio događaj A.

Nakon toga se izračunava vjerojatnost da je riječ ovisna samo o svojoj oznaci i vjerojatnost da je oznaka neke riječi ovisna samo o oznaci prethodne riječi.

Kao rezultat tih pretpostavki, dobivamo formulu:

$$'t \approx \underset{t_1^n}{argmax} \prod_{i=1}^{n} P(w_i\,|t_i)P(t_i|t_{i-1}) \tag{4}$$

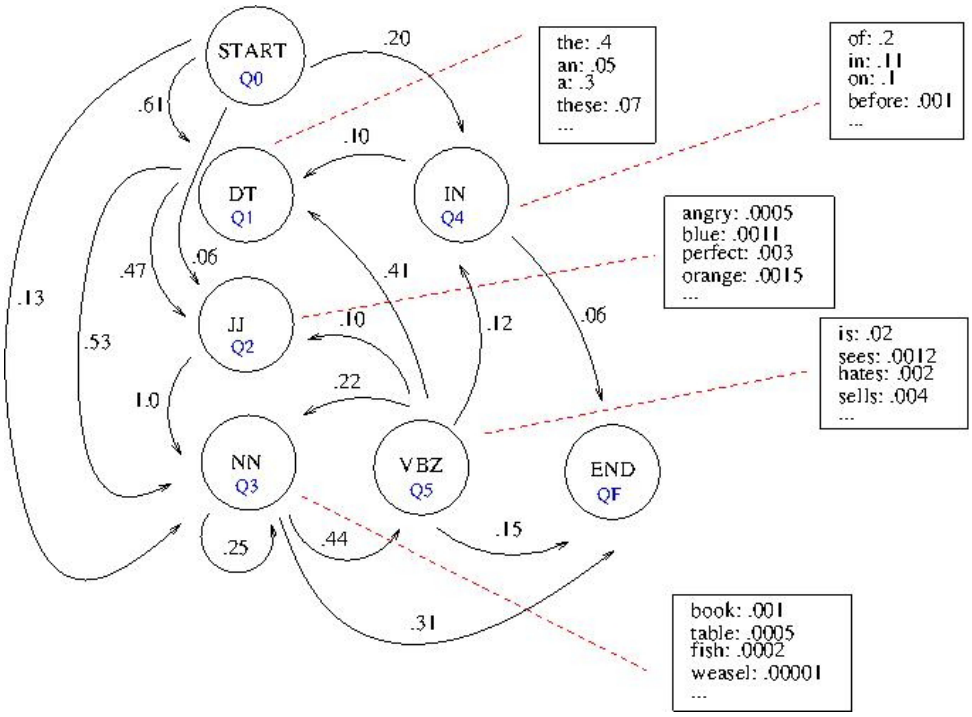
, koja je primjena Bayesovog teorema na vjerojatnost označivanja (eng. tag probability) (Meyers, 2012).

U formuli br. (4), t_1^n nam pokazuje odabiranje oznake t duljine n, s obzirom na unesene oznake, $P(w_i | t_i)$ označava vjerojatnost da riječ ovisi samo o svojoj oznaci a $P(t_i | t_{i-1})$ je vjerojatnost da neka oznaka ovisi samo o prethodnoj oznaci.

Pomoću izraza u formuli br. (4) dolazimo do frekvencija iz korpusa, odnosno do brojčane vrijednosti učestalosti pojavljivanja neke riječi sa danom oznakom, na primjer da se riječ "car" pojavi 20 puta u korpusu, od čega 10 puta kao NN.

SMM se sastoji od skupa stanja kao npr q0 (početno stanje) i qF(završno stanje), A- matrice vjerojatnosti prijelaza između bilo kojeg para od n stanja, O- niza riječi i B-niza vjerojatnosti (niza riječi sa njihovim oznakama).

Nakon konstrukcije SMM-a mora se primijeniti neki algoritam kako bi se SMM dekodirao, na primjer- Viterbi algoritam koji koristi dinamičko programiranje (Northwood, 2009).



Slika 3-Primjer SMM na obilježavanju riječi (Meyers, 2012)

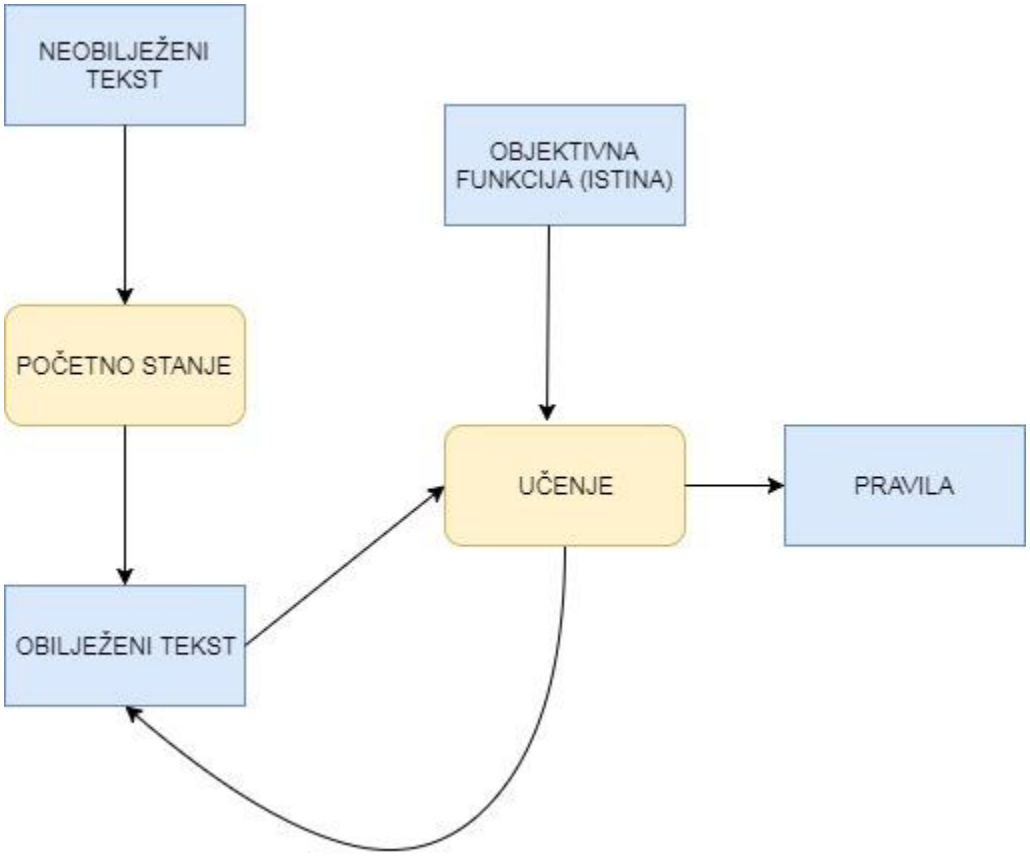
Slika prikazuje kako SMM uz pomoć Viterbi algoritma putem promjene stanja broji učestalost vjerojatnosti pojavljivanja neke riječi uz određenu oznaku. Tako je i označena regenica "The/DT book/NN is/VBZ on/IN the/DT table/NN" (Meyers, 2012). Ova vrsta obilježavanja ima postotak uspješnosti od 96%.

Osim obilježavanja pomoću Skrivenih Markovljevih modela, pristupi obilježavanju su i obilježavanje zasnovano na transformacijama (engl. transformation based tagging) i obilježavanje zasnovano na gramatikama ograničenja (engl. constraint grammar tagging).

2.3.2. Obilježavanje zasnovano na transformacijama

Obilježavanje zasnovano na transformacijama je jedna od uporaba učenja zasnovanog na transformacijama. Ono se sastoji od dvije faze – prva je faza učenja pravila koja se uglavnom obavlja samo jednom, nakon koje dolazi faza primjene pravila, koja se provodi više puta sve dok riječi nisu pravilno obilježene.

Rad faze učenja kod obilježivača zasnovanog na transformacijama može se opisati priloženim dijagramom (Ruef, 2003).



Slika 4-Rad obilježivača zasnovanog na transformacijama (faza učenja) (Ruef, 2003)

„Objektivna funkcija (istina)“ su podaci koje unosimo za fazu učenja, to je dio teksta kojeg unosimo, ali sa već ručno dodanim oznakama. Svrha toga je da se minimiziraju pogreške pri obilježavanju.

Obilježavanje pomoću transformacijskog učenja radi tako da se u početnom stanju riječi otprije poznate iz riječnika obilježe sa njihovom najučestalijom oznakom. Nakon toga se nepoznate riječi označe sa najvjerojatnijom oznakom, tj onom najučestalijom u korpusu, ovisno o tome počinju li velikim slovom ili ne (takve se automatski označuju kao osobna imenica). Sljedeći koraci su leksičko i kontekstualno označavanje, u leksičkom se nepoznate riječi označavaju odvojeno od rečenice, ovisno samo o njihovom morfološkom značenju i susjednoj riječi. Kontekstualno označavanje uzima u obzir sve riječi u rečenici, te nudi oznaku za nepoznatu riječ

u odnosu na sve riječi od kojih je rečenica sastavljena, riječ dobiva oznaku i značenje s obzirom na kontekst (Ruef, 2003).

Obilježivači koji rade na principu transformacijskog učenja vrlo lako uče nova pravila jer su skupovi pravila mali te su ti obilježivači u pravilu do deset puta brži od onih zasnovanih na SMM-u. Također, obuhvaća više konteksta od SMM-a, međutim, trajanje učenja je dugo ako se radi o velikim korpusima. Također, ne može odmah vratiti više od jednog rezultata, niti mjeriti uspješnost rezultata ali je rezultat jasan i jednostavan za analizu.

Postoji više implementacija obilježivača zasnovanog na transformacijskom učenju- prvi je bio Brillov algoritam iz 1993. Njegovo vrijeme učenja je bilo dugačko, što je ispravljeno u narednim algoritmima od Ramshawa i Marcusa (1994.) te Ngaijev i Florianov iz 2001., koji je znatno brži, nije ograničen samo na obilježavanje teksta te podržava višedimenzionalno učenje (Meyers, 2012).

2.3.3. Obilježavanje zasnovano na gramatikama

Obilježavanje zasnovano na gramatikama ograničenja je jedan od načina obilježavanja vrsta riječi, koji je značajan razvoj doživio u nedavnim godinama. Ta metoda nije zasnovana na pravilima koja automatski traže točne oznake već je većina pravila osmišljena tako da otklanjaju dvosmislenost, što u končnici vodi do ispravno označenog teksta. Označavanje se ovdje provodi kroz tri koraka- predprocesiranje, pretraživanje u leksikonu i primjenjivanje pravila (Eineborg M., Lindberg N., 1998).

U fazi predprocesiranja se traže idiomi, odnosno skupovi riječi čije značenje se znatno razlikuje od značenja svake riječi zasebno (npr. eng. "over the moon"- presretan). Idiomima se potom daje zasebna oznaka, kao da se radi o jednoj riječi. U narednom koraku, sve se riječi provjeravaju u leksikonu kako bi se otklonila morfološka dvosmislenost te se na kraju sav tekst označuje primjenom pravila, pomoću kojih se pregledavaju sve dvosmislenosti i otklanjaju netočne, dvosmislene pretpostavke.

Postoji četiri vrste pravila a to su redom, lokalno-kontekstna pravila (engl. local-context rules), leksička pravila (engl. lexical rules), pravila prepreka (engl. barrier rules) i odabrana pravila (engl. select rules). Lokalno-kontekstna pravila otklanjaju pretpostavke o mogućoj oznaci ovisno o tome jesu li uvjeti slaganja sa ostalim riječima u rečenici zadovoljeni, tj slaže li se oznaka neke riječi sa kontekstom rečenice. Leksička pravila utječu na odbacivanje mogućih oznaka s obzirom na određene uvjete koje analizirana riječ mora zadovoljavati sama za sebe, bez utjecaja ostalih riječi odnosno konteksta. Leksička pravila mogu biti flektivna (engl. inflectional rules) (promjena roda ili broja i ostalih karakteristika riječi mijenjanjem morfema- 1. lice jednine „play“ (engl. igra) + sufiks „s“ -> 3. lice jednine „plays“ (engl. igra)) i derivacijska (engl. derivational rules) (pravila koja se odnose na stvaranje novih vrsta riječi iz prethodno postojećih riječi dodavanjem prefiksa ili sufiksa, npr. pridjev „playful“ (engl. zaigran) + sufiks „ness“ -> imenica „playfulness“ (engl. zaigranost)) Pravila prepreka se odnose na kontekst, odnosno dopuštaju da na analiziranu riječ utječe neka druga riječ u smislu konteksta, čak i ako te dvije riječi ne stoje jedna do druge u rečenici. Međutim te riječi koje ih odvajaju (prepreke ili barijere) moraju odgovarati određenim uvjetima. Odabrana pravila se koriste kada je ispravna oznaka riječi pronađena (tj ispravna analiza) a sve ostale analize se otklanjaju (Eineborg M., Lindberg N., 1998).

Ukratko, ovom metodom se riječi dodaje više mogućih oznaka, koje se kasnije jedna po jedna eliminiraju ukoliko ne odgovaraju skupu pravila sve dok na kraju ne preostane jedna i po mogućnosti ispravna oznaka.

Jedan od prvih obilježivača vrsta riječi koji rade po ovom principu je ENGCG, Sveučilišta u Helsinkiju. On dodaje sintaksne i morfološke oznake riječima te je prije dvadesetak godina uspješno označavao 93-97% testiranih riječi (Voutilainen, 2000).

Razlog zašto je obilježavanje vrsta riječi općenito toliko zahtjevno je u tome da mnoge riječi u engleskom bez obzira na isti redoslijed morfema pripadaju različitim vrstama riječi ovisno o kontekstu. Tako primjerice, riječi „heat“ i „warm“ mogu biti imenice – „vrućina“ i „toplina“, dok mogu biti i glagoli „to heat“ – podgrijati i „to warm“ – ugrijati. Zadatak obilježivača riječi je da prepozna ove dvosmislenosti i ispravno ih ukloni. Obilježavanje vrsta riječi je najkorisnije u dohvaćanju informacija (engl. Information Retreival), pretvaranju teksta u govor – ovisno o naglasku „object“ može značiti i imenica – „predmet“ i glagol „protiviti se“, tako da riječ mora

biti ispravno označena kako bi bila i ispravno naglašena. Olakšava i automatsko prevođenje teksta te je korisno i pri parsiranju s obzirom da se određivanjem jedinstvenih oznaka riječima smanjuje i broj mogućih parsiranja (Nau, 2010).

2.3.4. Kolekcije oznaka i primjer

Zajedno sa različitim pristupima obilježavanju vrsta riječi, postoji i više kolekcija oznaka koje mogu međusobno odudarati jedna od druge. Te zbirke oznaka se nazivaju i tagsetovima te je danas među najkorištenijima Penn Treebank tagset (Liberman, 2003) sastavljen na Sveučilištu Pennsylvania (Jurafsky & Martin, 2016).

Penn treebank je velikim dijelom zasnovan na Brownovom korpusu (W3-Corpora Project, 1998), koji je bio pionir na ovom području. Brownov korpus se međutim sastojao od 87 oznaka a Penn Treebank je napravljen sa ciljem da se taj broj sažme i da se smanji redundancija uzimajući u obzir leksičke i sintaktičke informacije (Jurafsky & Martin, 2016). Brownov korpus su sastavili W.N.Francis i H.Kuceran sa Sveučilišta Brown te se on sastoji od 500 tekstova koji su podijeljeni u 15 kategorija.

Table 2
The Penn Treebank POS tagset.

1. CC	Coordinating conjunction	25. TO	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. $	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP$	Possessive pronoun	43.)	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. '	Left open single quote
22. RBS	Adverb, superlative	46. "	Left open double quote
23. RP	Particle	47. '	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. "	Right close double quote

Slika 5– Pen Treebank tagset (wenhoujx.blogspot.hr, 2013.)

Primjer Part Of Speech Tagganja pomoću Stanfordovog korijenovatelja (Toutanova & Klein, 2012) zasnovanog na Pen Treebank setu oznaka:

Izvorni tekst:

> ‘’About once a week, Uncle Vernon looked over the top of his newspaper and shouted that Harry needed a haircut. Harry must have had more haircuts than the rest of the boys in his class put together, but it made no difference, his hair simply grew that way -- all over the place. Harry was frying eggs by the time Dudley arrived in the kitchen with his mother. Dudley looked a lot like Uncle Vernon. He had a large pink face, not much neck, small, watery blue eyes, and thick blond hair that lay smoothly on his thick, fat head. Aunt Petunia often said that Dudley looked like a baby angel -- Harry often said that Dudley looked like a pig in a wig.’’ (Rowling, 1997)

Tekst nakon označavanja:

About/IN once/RB a/DT week/NN ,/, Uncle/NNP Vernon/NNP looked/VBD over/RP the/DT top/NN of/IN his/PP$ newspaper/NN and/CC shouted/VBD that/IN Harry/NNP needed/VBD a/DT haircut/NN ./.

Harry/NNP must/MD have/VB had/VBN more/JJR haircuts/NNS than/IN the/DT rest/NN of/IN the/DT boys/NNS in/IN his/PRP$ class/NN put/VBN together/RB ,/, but/CC it/PRP made/VBD no/DT difference/NN ,/, his/PRP$ hair/NN simply/RB grew/VBD that/DT way/NN --/: all/DT over/IN the/DT place/NN ./.

Harry/NNP was/VBD frying/VBG eggs/NNS by/IN the/DT time/NN Dudley/NNP arrived/VBD in/IN the/DT kitchen/NN with/IN his/PRP$ mother/NN ./.

Dudley/NNP looked/VBD a/DT lot/NN like/IN Uncle/NNP Vernon/NNP ./.

He/PRP had/VBD a/DT large/JJ pink/JJ face/NN ,/, not/RB much/JJ neck/NN ,/, small/JJ ,/, watery/JJ blue/JJ eyes/NNS ,/, and/CC thick/JJ blond/JJ hair/NN that/WDT lay/VBD smoothly/RB on/IN his/PRP$ thick/JJ ,/, fat/JJ head/NN ./.

Aunt/NNP Petunia/NNP often/RB said/VBD that/IN Dudley/NNP looked/VBD like/IN a/DT baby/NN angel/NN --/: Harry/NNP often/RB said/VBD that/IN Dudley/NNP looked/VBD like/IN a/DT pig/NN in/IN a/DT wig/NN ./.

3. Praktična usporedba korijenovatelja i lematizatora te obilježivača vrsta riječi

3.1. Usporedba korijenovatelja i lematizatora

Kako bi usporedili funkcionalnost korijenovatelja, izvorni tekst je proveden kroz korijenovatelje i lematizator te je određena postotak njihove točnosti odnosno pogreške.

Kako bi se saznao postotak točnosti ovih korijenovatelja i lematizatora te korištenih alata biti će korištena formula:

$$\text{Postotak pogrešaka} = \left(\frac{\text{Broj pogrešaka}}{\text{Ukupan broj riječi}}\right) * 100 \qquad (5)$$

Uspoređena su dva korijenovatelja i lematizator, Porterov (McKenzie, 2010), Paice-Huskov korijenovatelj (Text Analysis Online, 2016) i NLTK Wordnet lematizator (Text analysis Online, 2016). Korištena su dva teksta, od kojih je prvi sastavljen od 266 nepovezanih riječi dok je drugi kratki tekst od 48 riječi. Izvorni tekstovi te njihove verzije nakon korijenovanja i lematizacije su u poglavlju „Prilozi“.

U sljedećoj tablici su prikazani rezultati prvoga teksta (koji se sastoji od 266 riječi), u Prilozima označenog brojem 1).

Tabela 1- Usporedba korijenovatelja 1

Korijenovatelj	Alat	Link	Broj pogrešaka	Postotak pogrešaka	Postotak ispravnosti
Porterov	JavaScript Porter korijenovatelj	(McKenzie, Porter json demo, 2010)	108	40.60%	59.39 %

Paice-Huskov	NLTK korijenovatelj	(Text Analysis Online, 2016)	104	39.09%	60.91%
Wordnet Lematizator	NLTK lematizator	(Text analysis Online, 2016)	63	23.68%	76.32%

U ovoj tablici su prikazani rezultati nakon korijenovanja teksta koji je u Prilozima označen brojem 2). Tekst se sastoji od 55 riječi.

Tabela 2 - Usporedba korijenovatelja 2

Korijenovatelj	Alat	Link	Broj pogrešaka	Postotak pogrešaka	Postotak ispravnosti
Porterov	JavaScript Porter korijenovatelj	(McKenzie, Porter json demo, 2010)	19	34.54%	65.46%
Paice-Huskov	NLTK korijenovatelj	(Text Analysis Online, 2016)	18	32.72%	67.28%
Wordnet Lematizator	NLTK lematizator	(Text analysis Online, 2016)	13	23.63%	76.37%

Po ovim rezultatima, vidimo da je Porterov korijenovatelj imao najmanji postotak uspješnosti dok je NLTK lematizator koji je koristio WordNetove korpuse imao najmanji postotak grešaka.

Porterov korijenovatelj mnoge riječi je skratio, ali ne na njihov ispravan korijen (često propuštajući dodati morfem „e“ nakon skraćivanja), kao kod riječi „dancing“, koja nije skraćena na „dance“ već na „danc“. Neke riječi su nepotrebno skraćene, jer su one same sebi korijen, kao „proportion“ koja je skraćena na „proport“ iako se ona ne bi trebala skraćivati. Isto tako, vidimo da riječi kojima je sufiks „less“ kao npr. „hairless“ i „landless“ nisu skraćene na „hair“ i „land“ kao što bi trebale biti, dok su riječi sa sufiksom „s“ koja označava množinu prikladno skraćene.

Glagoli koji su u prošlom vremenu su uspješno korijenovani, kao npr. „upturned“ na „upturn“, gdje je sufiks koji označava prošlo vrijeme uklonjen.

U drugome tekstu korijenovanom pomoću Portera, glagol u infinitivu „failing“ je skraćen ispravno, maknuto je „ing“ te je ostala riječ „fail“. Riječ „tanned“ je uspješno skraćena te je pravilom: (*d and not (*L or *S or *Z)[4] otklonjeno dvostruko slovo „n“, dok je riječ „agree“ skraćena, iako to nije bilo potrebno, kao i riječ „his“, gdje je korijen manji od jednog sloga i ne bi se trebalo na osnovu toga oduzimati –s iz „plurala“. Isto tako, riječi „replacement“ koja se skratila kao „replac“ trebalo je dodati „e“ da to bude morfološki ispravan korijen.Kao što je bilo i očekivano, otklonjeni su nastavci „ness“, „ance“ i „ement“ iz „calloussness“, „hesitance“ i „replacement“ međutim ti korijeni nisu uređeni da bi postali leksički ispravne riječi (osim „callous“).

Paice-Huskov orijenovatelj je bio sklon pretjeranom skraćivanju riječ a riječi je skratio puno više nego što je to učinio Porterov stemer. Primjećuje se da je npr. u riječi „accidental“ (engl. slučajno) koju je Porterov korijenovatelj skratio na „accident“ (engl. nezgoda, slučajnost), Paice-Huskov ju krati na „accid“ (engl. kiselina), riječ sa potpuno drugačijim značenjem. Isto kao i Porterov korijenovatelj, i Paice-Huskov pogrešno skraćuje riječ „upstairs“, zaključujući da je u množini.

U drugom tekstu, riječ „tan“ je ispravno skratio i maknuo dvostruko n, dok riječi „agree“, „his“ i „industry“ nisu skraćene kako i treba biti. Međutim „operations“ je umjesto da je samo uklonjeno –s u pluralu je skraćena na „op“, od „have“ je nepotrebno uklonjeno „e"“ a riječ „callousness“ koju je Porterov korijenovatelj uspješno skratio na „callous“, je skraćena samo na „cal“.

Bez obzira na ovo, vidljivo je iz rezultata da Paice-Huskov i Porterov korijenovatelj imaju prilično sličan postotak točnosti, uz neznatnu prednost od otprilike 2% za Paice-Huska.

[4] Pravila u poglavlju „Prilozi“

NLTK-ov lematizator koristi Wordnet, bazu podataka od Sveučilišta u Princetonu. Ta baza podataka sastoji veliki broj engleskih riječi, koje su međusobno grupirane u setove sinonima, gdje svaki set označuje jedan koncept (Princeton University, 2015).

Uzimajući to u obzir, nije začuđujuće da je lematizator najuspješniji u skraćivanju riječi na njihove korijene, odnosno leme. Lematizator je ispravno otklonio puno više sufiksa nego Porterov ili Paice-Huskov a i učinio je to na leksički ispravniji način. Međutim, u mnogo je slučajeva ipak propustio skratiti riječ na njegovu lemu. Neke od tih riječi su npr ‘’hairless’’ ili „admittance’’ koje nisu skraćene na njihove korijene „hair“ i „admit“ dok je uspješno riješen slučaj "-ing" sufiksa, koji je primjerice otklonjen kod riječi „applying“ u "apply".

U drugom tekstu neke riječi je propustio skratiti, kao „troubled“ koja je trebala postati „trouble“, „failing“ kojemu nije maknut „ing“, kao i „decisiveness“ i „sensibility“ koje nisu skraćene. „Tolerated“ je skraćen te mu je ostavljeno „e“ kao zadnje slovo, što je ispravno a maknuta je „ed“ oznaka prošlog vremena.

Prema istraživanju provedenom od državnog sveučilišta u Montclaireu, (Wiese, Ho, & Hill, 2011) pokazano je da je Porterov stemer točan samo u 29% slučajeva, te je u tom postotku uspio producirati točne i potpune riječi. Bolji od njega bio je Paice-Huskov stemer sa 32%, dok su prednjačili Snowball, Kstem koji koristi Krovetzov stemer (53%) i Mstem (58%).

Prema testiranju provedenom za ovaj završni rad, prednjačio je WordNet- zasnovan lematizator sa prosječno 76.34% ispravno pronađenih lema, dok je na drugom mjestu Paice-Huskov korijenovatlj sa prosječno 64.09% ispravno korijenovanih riječi a Porterov korijenovatelj je ispravno skratio 62.42% unesenih riječi.

3.2. Usporedba obilježivača riječi

Većina obilježivača riječi danas, pa tako i onaj najšire korišten- Stanfordov (Stanford NLP Group, 2015), se koriste Penn Treebank setom oznaka (engl. tagova) kako bi označili vrste i funkcije riječi u rečenici. U ovom završnom radu biti će prikazane razlike između Stanfordovog obilježivača riječi, NlpDotNet-ovog obilježivača riječi (NlpDotNet, 2009) i Genie-a (Tsuruoka, 2006). Sva tri obilježivača koriste Penn Treebenk set oznaka. Sva tri obilježivača su stohastički obilježivači, zasnovani na SMM-a.

Kao i za usporedbu korijenovatelja i lematizatora, za usporedbu obilježivača vrsta riječi koristit ćemo se formulom:

$$\text{Postotak pogrešaka} = \left(\frac{\text{Broj pogrešaka}}{\text{Ukupan broj riječi}}\right) * 100 \quad (5)$$

Izvorni tekst i tekstovi analizirani obilježivačima vrsta riječi se može pronaću u poglavlju „Prilozi“. Ukupan broj riječi u izvornom tekstu je 123.

Tabela 3- Usporedba obilježivača vrsta riječi

Obilježivač vrsta riječi	Link	Broj pogrešaka	Postotak pogrešaka	Postotak ispravnosti
Stanfordov obilježivač	(Stanford NLP Group, 2015)	11	8.94%	91.06 %
NlpDotNet obilježivač	(NlpDotNet, 2009)	17	13.82%	86.18%
Genia obilježivač	(Tsuruoka, 2006)	16	13.00%	87.00%

Prema ovim rezultatima, vidimo da je Stanfordov obilježivač riječi imao najveći postotak ispravno obilježenih riječi, sa 11 pogrešaka od ukupno 123 riječi. Među pogrešno označenim riječima našla se riječ „that" koja je u prvom slučaju trebala biti determinator a označena je kao wh-determinator, a u drugom je označena kao prijedlog a u stvarnosti je prilog. Riječ „feeling" je prepoznata kao imenica, što ona može biti (engl. osjećaj) ali na ovom mjestu u rečenici je ona glagol, te je ista pogreška napravljena i kod riječi „caring", dok je glagol „aware" označen kao pridjev. Ostale greške su manje značajne npr. „better" nije označen kao prilog u komparativu, već samo kao prilog, „up" je označen kao čestica ali je s ovom funkcijom u rečenici prilog, a „lot" je označeno kao imenica a u stvarnosti je prilog.

NlpDotNet-ov obilježivač riječi je imao najmanji postotak ispravno obilježenih riječi, 86.18%. On je radio učestale pogreške na glagolima. Nije prepoznao „may" kao modalni glagol, „stressed" je označio kao glagol ali u krivome vremenu te nije prepoznavao modalne glagole kada su u njihovom skraćenom obliku npr. „they'll" (they will) je umjesto kao dvije riječi, odnosno osobnu zamjenicu i modalni glagol- prepoznao je to kao jednu riječ i to determinator. Isto je napravio i za „you're" i za „don't". Riječi „happens", „aware" i „see" nije uopće prepoznao kao glagole.

Genia obilježivač riječi je možda najzanimljiviji u ovome pregledu. Osim što se prikazuje i "jednostavniji" oblik parsiranja, odnosno "chunking" koji označava parsiranje bez rekurzivnih fraza. Osim toga, u tekstovima koji su vezani uz medicinu ili biomedicinu prepoznaje ključne riječi vezane uz to polje. Automatski prikazuje riječ u osnovnom obliku. Ovdje je imao 87% točnosti, te je isto imao problema s prepoznavanjem skraćenih modalnih glagola kao npr. „they'll" (they will), „isn't" (is not) i „don't" (do not). Ostale pogreške bile su označavanje imenice „burnout" kao priloga, priloga „lot" kao imenice, glagola „feel" kao imenice, itd.

Iako je Genia obilježivač riječi zbog raznih mogućnosti koje nudi najzanimljiviji, najispravnijim se pokazao Stanfordov obilježivač riječi koji je i jedan od najkorištenijih obilježivača inače.

4. Zaključak

U ovome radu obrađeni su algoritmi korijenovanja i lematizacije te pristupi obilježavanju vrsta riječi. To su ključni pojmovi u obradi prirodnog jezika i koriste se najviše kako bi grupiranjem riječi i pronalaženjem vrste riječi olakšali pretraživanje informacija. Uspoređena su dva korijenovatelja, Porterov i Paice-Huskov te WordNetov lematizator na temelju izvornih tekstova, kako bi mogli odrediti njihovu uspješnost u pronalaženju korijena riječi. Prema tome, WordNet lematizator je bio najuspješniji sa prosječno 76.34% ispravno pronađenih lema, dok je na drugom mjestu Paice-Huskov korijenovatelj sa prosječno 64.09% ispravno korijenovanih riječi a Porterov korijenovatelj je ispravno skratio 62.42% unesenih riječi. Uspoređena su i tri obilježivača riječi, Stanfordov, NlpDotNet obilježivač i Genia obilježivač. Stanfordov se pokazao najbolji sa 91.06% ispravno označenih riječi, sljedeći je bio Genia sa 87% i NlpDotNet sa 86.18%. Zaključujemo da su za pretraživanje informacija pogodniji lematizatori i korijenovatlelji koji koriste korpuse te da je Stanfordov obilježivač riječi uspješno rješenje za pronalaženje vrsta riječi.

5. Prilozi

5.1. Pravila Porterovog algoritma

Korak 1. a)

SSES -> SS

IES -> I

SS-> SS

S ->

Korak 1. b)

(m>0) EED->EE

(*v*) ED->

(*v*) ING ->

AT	-> ATE
BL	-> BLE
IZ	-> IZE

(*d and not (*L or *S or *Z)) -> single letter

Korak 2.

(m>0) ATIONAL	->	ATE
(m>0) TIONAL	->	TION

(m>0) ENCI	->	ENCE
(m>0) ANCI	->	ANCE
(m>0) IZER	->	IZE
(m>0) ABLI	->	ABLE
(m>0) ALLI	->	AL
(m>0) ENTLI	->	ENT
(m>0) ELI	->	E
(m>0) OUSLI	->	OUS
(m>0) IZATION	->	IZE
(m>0) ATION	->	ATE
(m>0) ATOR	->	ATE
(m>0) ALISM	->	AL
(m>0) IVENESS	->	IVE
(m>0) FULNESS	->	FUL
(m>0) OUSNESS	->	OUS
(m>0) ALITI	->	AL
(m>0) IVITI	->	IVE
(m>0) BILITI	->	BLE

Korak 3.

(m>0) ICATE	->	IC
(m>0) ATIVE	->	
(m>0) ALIZE	->	AL
(m>0) ICITI	->	IC
(m>0) ICAL	->	IC
(m>0) FUL	->	
(m>0) NESS	->	

Korak 4.

(m>1) AL
(m>1) ANCE
(m>1) ENCE
(m>1) ER
(m>1) IC
(m>1) ABLE
(m>1) IBLE
(m>1) ANT
(m>1) EMENT
(m>1) MENT
(m>1) ENT
(m>1 and (*S or *T)) ION
(m>1) OU
(m>1) ISM
(m>1) ATE
(m>1) ITI
(m>1) OUS
(m>1) IVE
(m>1) IZE

Korak 5. a)

(m>1) E

(m=1 and not *o) E

Korak 5. b)

(m > 1 and *d and *L) -> single letter

(Porter, Tartarus, 1980)

5.2. Lovinsov korijenovatelj

B dio, uvjeti:

A No restrictions on stem

B Minimum stem length = 3

C Minimum stem length = 4

D Minimum stem length = 5

E Do not remove ending after *e*

F Minimum stem length = 3 and do not remove ending after *e*

G Minimum stem length = 3 and remove ending only after *f*

H Remove ending only after *t* or *ll*

I Do not remove ending after *o* or *e*

J Do not remove ending after *a* or *e*

K Minimum stem length = 3 and remove ending only after *l, i* or *u*e*

L Do not remove ending after *u, x* or *s*, unless *s* follows *o*

M Do not remove ending after *a, c, e* or *m*

N Minimum stem length = 4 after *s***, elsewhere = 3

O Remove ending only after *l* or *i*

P Do not remove ending after *c*

Q Minimum stem length = 3 and do not remove ending after *l* or *n*

R Remove ending only after *n* or *r*

S Remove ending only after *dr* or *t*, unless *t* follows *t*

T Remove ending only after *s* or *t*, unless *t* follows *o*

U Remove ending only after *l, m, n* or *r*

V Remove ending only after *c*

W Do not remove ending after *s* or *u*

X Remove ending only after *l, i* or *u*e*

Y Remove ending only after *in*

Z Do not remove ending after *f*

AA Remove ending only after *d, f, ph, th, l, er, or, es* or *t*

BB Minimum stem length = 3 and do not remove ending after *met* or *ryst*

CC Remove ending only after *l*

(Porter, Tartarus, 1980)

5.3. Usporedba korijenovatelja i lematizatora

5.3.1. Izvorni tekstovi

1)

abandon abandoned abase abash abate abated abatement abatements abates abbess abbey abbeys abbominable abbot abbots abbreviated abed abel aberga abergavenny abet abetting abhominable abhor abhorr abhorred abhorring abhors abhorson abide abides abilities ability abject abjectly abjects abjur abjure able abler aboard abode aboded abodements aboding abominable abominably abominations abortive abortives abound abounding about accessible accidence accident accidental accidentally accidents accite accited accites acclamations accommodate accommodated accommodation accommodations accommodo accompanied accompany accompanying admission admit admits admittance admitted appetite appetites applaud applauded applauding applause applauses apple apples appletart appliance appliances applications applied

applies apply applying blubbering blue bluecaps bluest blunt blunted blunter bluntest blunting bluntly bluntness blunts blur blurr blurs blush crowded crowding crowds crowflowers crowing crowkeeper crown crowned crowner crownet crownets crowning crowns crows crudy damsel damsons dan danc dance dancer dances dancing dandle dandy dane dang danger dangerous dangerously dangers dangling daniel danish dank dankish danskers daphne dappled dapples dar dardan dardanian dardanius dare forget forgiven forgotten formless fox gaze gossip grief guardian gazer gaze gazeth gaze gear geck gees hairless hair hangmen hangman idea imperfect land landless landlord landmen laugh lean libel prophetess proportion prosper push mediterranean mediterraneum medlar medlar meed meed meek meekli meek meet meeter meetest meet quench quenchless quern quest questant rag ragged rage reckless razor upspring upstairs upstart upturned upward upwards urchin urchinfield urchins urg urge urged unthought unthread unthrift unthrifts unthrifty untie untied until untimber untimely volumnia volumnius voluntaries voluntary voluptuously voluptuousness vomissement vomit vomits vor vore vortnight wrathfully wraths wreak wreakful wreaks wreath wreathed wreathen wreaths your yours yourself yourselves youth youthful youths youtli zanies zany zeal zealous zeals

(Porter, Tatarus)

2)

A tanned troubled young man agreed to save the failing industry. His decisiveness and sensibility helped the workers. He expertly fixed the electrical issues and controlled the operations. They will have difficulty finding his replacement. Regardless, effective work will not cease. Callousness and hesitance will not be tolerated.

5.3.2 Analiza tekstova korijenovateljima

Pogrešno korijenovane riječi će biti označene žutom bojom.

5.3.2.1 Porterov korijenovatelj

1)

abandon abandon abas abash abat abat abat abat abat abbess abbei abbei abbomin abbot abbot abbrevi ab abel aberga abergavenni abet abet abhomin abhor abhorr abhor abhor abhor abhorson abid abid abil abil abject abjectli abject abjur abjur abl abler aboard abod abod abod abod abomin abomin abomin abort abort abound abound about access accid accid accident accident accid accit accit accit accit acclam accommod accommod accommod accommod accommodo accompani accompani accompani admiss admit admit admitt admit appetit appetit applaud applaud applaud applaus applaus appl appl appletart applianc applianc applic appli appli appli appli blubber blue bluecap bluest blunt blunt blunter bluntest blunt bluntli blunt blunt blur blurr blur blush crowd crowd crowd crowflow crow crowkeep crown crown crowner crownet crownet crown crown crow crudi damsel damson dan danc danc dancer danc danc dandl dandi dane dang danger danger danger danger dangl daniel danish dank dankish dansker daphn dappl dappl dar dardan dardanian dardaniu dare forget forgiven forgotten formless fox gaze gossip grief guardian gazer gaze gazeth gaze gear geck gee hairless hair hangmen hangman idea imperfect land landless landlord landmen laugh lean libel prophetess proport prosper push mediterranean mediterraneum medlar medlar meed meed meek meekli meek meet meeter meetest meet quench quenchless quern quest questant rag rag rage reckless razor upspr upstair upstart upturn upward upward urchin urchinfield urchin urg urg urg unthought unthread unthrift unthrift unthrifti unti unti until untimb untim volumnia volumniu voluntari voluntari voluptu voluptu vomiss vomit vomit vor vore vortnight wrathfulli wrath wreak wreak wreak wreath wreath wreathen wreath your your yourself yourselv youth youth youth youtli zani zani zeal zealou zeal

2)

A tan troubl young man agre to save the fail industri Hi decis and sensibl help the worker He expertli fix the electr issu and control the oper Thei will have difficulti find hi replac Regardless effect work will not ceas Callous and hesit will not be toler

5.3.2.2. Paice-Huskov korijenovatelj

1)

abandon abandon abas abash ab ab ab ab ab abbess abbey abbey abbomin abbot abbot abbrevy ab abel aberg abergavenny abet abet abhomin abh abhor abhor abhor abh abhorson abid abid abl abl abject abject abject abs abs abl abl aboard abod abod abod abod abomin abomin abomin abort abort abound abound about access accid accid accid accid accid accit accit accit acclam accommod accommod accommod accommod accommodo accompany accompany accompany admit admit admit admit admit appetit appetit applaud applaud applaud applaus applaus appl appl appletart apply apply apply apply apply apply apply blub blu bluecap bluest blunt blunt blunt bluntest blunt blunt blunt blunt blur blur blur blush crowd crowd crowd crowflow crow crowkeep crown crown crown crownet crownet crown crown crow crudy damsel damson dan dant dant dant dant dant dandl dandy dan dang dang dang dang dang dangl daniel dan dank dank dansk daphn dappl dappl dar dard dard dardani dar forget forg forgot formless fox gaz gossip grief guard gaz gaz gaze gaz gear geck gee hairless hair hangm hangm ide imperfect land landless landlord landm laugh lean libel prophetess proport prosp push mediter mediterrane medl medl mee mee meek meekl meek meet meet meetest meet quench quenchless quern quest quest rag rag rag reckless raz upspr upstair upstart upturn upward upward urchin urchinfield urchin urg urg urg unthought unthread unthrift unthrift unthrifty unty unty until untimb untim volumn volumni volunt volunt voluptu voluptu vomiss vomit vomit vor vor vortnight wrath wrath wreak wreak

wreak wrea wreath wreath wreath yo yo yourself yourselv you youth youth youtl zany zany zeal zeal zeal

2)

a tan troubl young man agree to sav the fail industry . his decid and sens help the work . he expert fix the elect issu and control the op . they wil hav difficul find his replac . regardless , effect work wil not ceas . cal and hesit wil not be tol .

5.3.2.3. NLTK WordNet lematizator

1)

abandon abandon abase abash abate abate abatement abatement abate abbess abbey abbey abbominable abbot abbot abbreviate abed abel aberga abergavenny abet abet abhominable abhor abhorr abhor abhor abhors abhorson abide abides ability ability abject abjectly abjects abjur abjure able abler aboard abode aboded abodements aboding abominable abominably abomination abortive abortives abound abound about accessible accidence accident accidental accidentally accidents accite accited accites acclamation accommodate accommodate accommodation accommodation accommodo accompany accompany accompany admission admit admit admittance admit appetite appetite applaud applaud applauding applause applauses apple apple appletart appliance appliance application apply applies apply apply blubber blue bluecaps blue blunt blunt blunt blunt blunt bluntly bluntness blunts blur blurr blur blush crowd crowd crowd crowflowers crow crowkeeper crown crown crowner crownet crownets crown crown crow crudy damsel damson dan danc dance dancer dance dance dandle dandy dane dang danger dangerous dangerously danger dangle daniel danish dank dankish danskers daphne dappled dapple dar dardan dardanian dardanius dare forget forgiven forgotten formless fox gaze gossip grief

guardian gazer gaze gazeth gaze gear geck gee hairless hair hangman hangman idea imperfect land landless landlord landman laugh lean libel prophetess proportion prosper push mediterranean mediterraneum medlar medlar meed meed meek meekli meek meet meeter meet meet quench quenchless quern quest questant rag rag rage reckless razor upspring upstairs upstart upturned upward upwards urchin urchinfield urchin urg urge urge unthought unthread unthrift unthrifts unthrifty untie untie until untimber untimely volumnia volumnius voluntary voluntary voluptuously voluptuousness vomissement vomit vomit vor vore vortnight wrathfully wraths wreak wreakful wreaks wreath wreathe wreathen wreath your yours yourself yourselves youth youthful youth youtli zany zany zeal zealous zeal

2)

A tan troubled young man agree to save the failing industry . His decisiveness and sensibility help the worker . He expertly fix the electrical issue and control the operation . They will have difficulty find his replacement . Regardless , effective work will not cease . Callousness and hesitance will not be tolerate .

5.4. Usporedba obilježivača vrsta riječi

5.4.1. Izvorni tekst

Burnout may be the result of unrelenting stress, but it isn't the same as too much stress. Stress, by and large, involves too much: too many pressures that demand too much of you physically and psychologically. Stressed people can still imagine, though, that if they can just get everything under control, they'll feel better.

Burnout, on the other hand, is about not enough. Being burned out means feeling empty, devoid of motivation, and beyond caring. People experiencing burnout often don't see any hope of positive change in their situations. If excessive stress is like drowning in responsibilities, burnout is being all dried up. And while you're usually aware of being under a lot of stress, you don't always notice burnout when it happens (Helpguide.org, 2017).

5.4.2. Analiza pomoću obilježivača riječi

Pogrešno obilježene riječi biti će označene žutom bojom.

Table 2
The Penn Treebank POS tagset.

1. CC	Coordinating conjunction	25. TO	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. $	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP$	Possessive pronoun	43.)	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. ‘	Left open single quote
22. RBS	Adverb, superlative	46. “	Left open double quote
23. RP	Particle	47. ’	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. ”	Right close double quote

Slika 6- Penn Treebank set oznaka (wenhoujx.blogspot.hr, 2013.)

5.4.2.1. Stanfordov obilježivač riječi

Burnout_NN may_MD be_VB the_DT result_NN of_IN unrelenting_JJ stress_NN ,_, but_CC it_PRP is_VBZ n't_RB the_DT same_JJ as_IN too_RB much_JJ stress_NN ._. Stress_NN ,_, by_IN and_CC large_JJ ,_, involves_VBZ too_RB much_JJ :_: too_RB many_JJ pressures_NNS that_WDT demand_VBP too_RB much_JJ of_IN you_PRP physically_RB and_CC psychologically_RB ._. Stressed_VBN people_NNS can_MD still_RB imagine_VB ,_, though_RB ,_, that_IN if_IN they_PRP can_MD just_RB get_VB everything_NN under_IN

control_NN ,_, they_PRP 'll_MD feel_VB better_RB ._. Burnout_NN ,_, on_IN the_DT other_JJ hand_NN ,_, is_VBZ about_IN not_RB enough_RB ._. Being_VBG burned_VBN out_RP means_VBZ feeling_NN empty_JJ ,_, devoid_JJ of_IN motivation_NN ,_, and_CC beyond_IN caring_NN ._. People_NNS experiencing_VBG burnout_NN often_RB do_VBP n't_RB see_VB any_DT hope_NN of_IN positive_JJ change_NN in_IN their_PRP$ situations_NNS ._. If_IN excessive_JJ stress_NN is_VBZ like_IN drowning_VBG in_IN responsibilities_NNS ,_, burnout_NN is_VBZ being_VBG all_DT dried_VBD up_RP ._. And_CC while_IN you_PRP 're_VBP usually_RB aware_JJ of_IN being_VBG under_IN a_DT lot_NN of_IN stress_NN ,_, you_PRP do_VBP n't_RB always_RB notice_VB burnout_NN when_WRB it_PRP happens_VBZ ._.

5.4.2.2. NLPDotNet obilježivač riječi

Burnout/JJ may/NN be/VB the/DT result/NN of/IN unrelenting/JJ stress/NN ,/, but/CC it/PRP isn’t/VBP the/DT same/JJ as/IN too/RB much/DT stress/NN ./.

Stress/NN ,/, by/IN and/CC large/JJ ,/, involves/VBZ too/RB much/JJ :/: too/RB many/JJ pressures/NNS that/WDT demand/VBP too/RB much/JJ of/IN you/PRP physically/RB and/CC psychologically/RB ./.

Stressed/VBD people/NNS can/MD still/RB imagine/VB ,/, though/RB ,/, that/DT if/NN they/PRP can/MD just/RB get/VB everything/NN under/IN control/NN ,/, they’ll/DT feel/NN better/JJR ./.

Burnout/NN ,/, on/IN the/DT other/JJ hand/NN ,/, is/VBZ about/IN not/RB enough/RB ./.

Being/VBG burned/VBN out/RP means/VBZ feeling/VBG empty/JJ ,/, devoid/JJ of/IN motivation/NN ,/, and/CC beyond/IN caring/VBG ./.

People/NNS experiencing/VBG burnout/NN often/RB don’t/.

see/UH any/DT hope/NN of/IN positive/JJ change/NN in/IN their/PRP$ situations/NNS ./.

If/IN excessive/JJ stress/NN is/VBZ like/IN drowning/VBG in/IN responsibilities/NNS ,/, burnout/NN is/VBZ being/VBG all/DT dried/JJ up/RP ./.

And/CC while/IN you're/NN usually/RB aware/JJ of/IN being/VBG under/IN a/DT lot/NN of/IN stress/NN ,/, you/PRP don't/.

always/RB notice/VB burnout/NN when/WRB it/PRP happens./

5.4.2.3. Genia obilježivač riječi

Word	**Base Form**	**Part-Of-Speech**	**Chunk**	**Named Entity**
Burnout	Burnout	NN	B-NP	O
May	may	MD	B-VP	O
Be	be	VB	I-VP	O
The	the	DT	B-NP	O
Result	result	NN	I-NP	O
Of	of	IN	B-PP	O
unrelenting	unrelenting	JJ	B-NP	O
Stress	stress	NN	I-NP	O
,	,	,	O	O
But	but	CC	O	O
It	it	PRP	B-NP	O

isn’t	isn’t	VBD	B-VP	O
The	the	DT	B-NP	O
Same	same	JJ	I-NP	O
As	as	IN	B-PP	O
Too	too	RB	B-NP	O
Much	much	JJ	I-NP	O
Stress	stress	NN	I-NP	O
.	.	.	O	O
Stress	Stress	NNP	B-NP	O
,	,	,	O	O
By	by	IN	B-ADVP	O
And	and	CC	I-ADVP	O
Large	large	JJ	I-ADVP	O
,	,	,	O	O
Involves	involve	VBZ	B-VP	O
Too	too	RB	B-ADVP	O
Much	much	RB	I-ADVP	O
:	:	:	O	O
Too	too	RB	B-NP	O
Many	many	JJ	I-NP	O

Pressures	pressure	NNS	I-NP	O
That	that	WDT	B-NP	O
Demand	demand	VBP	B-VP	O
Too	too	RB	B-NP	O
Much	much	JJ	I-NP	O
Of	of	IN	B-PP	O
You	you	PRP	B-NP	O
Physically	physically	RB	B-ADVP	O
And	and	CC	I-ADVP	O
psychologically	psychologically	RB	I-ADVP	O
.	.	.	O	O
Stressed	Stress	VBN	B-NP	O
People	people	NNS	I-NP	O
Can	can	MD	B-VP	O
Still	still	RB	I-VP	O
Imagine	imagine	VB	I-VP	O
,	,	,	O	O
Though	though	RB	B-ADVP	O
,	,	,	O	O
That	that	IN	B-SBAR	O

If	if	IN	B-SBAR	O
They	they	PRP	B-NP	O
Can	can	MD	B-VP	O
Just	just	RB	I-VP	O
Get	get	VB	I-VP	O
Everything	everything	NN	B-NP	O
Under	under	IN	B-PP	O
Control	control	NN	B-NP	O
,	,	,	O	O
they’ll	they’ll	DT	B-NP	O
Feel	feel	NN	I-NP	O
Better	well	RBR	B-ADJP	O
.	.	.	O	O
Burnout	Burnout	RB	B-ADVP	O
,	,	,	O	O
On	on	IN	B-PP	O
The	the	DT	B-NP	O
Other	other	JJ	I-NP	O
Hand	hand	NN	I-NP	O
,	,	,	O	O
Is	be	VBZ	B-VP	O
About	about	IN	B-PP	O

Not	not	RB	B-ADVP	O
Enough	enough	RB	I-ADVP	O
.	.	.	O	O
Being	Be	VBG	B-VP	O
Burned	burn	VBN	I-VP	O
Out	out	RP	B-PRT	O
Means	mean	VBZ	B-VP	O
Feeling	feel	VBG	I-VP	O
Empty	empty	JJ	B-NP	O
,	,	,	I-NP	O
Devoid	devoid	JJ	I-NP	O
Of	of	IN	B-PP	O
Motivation	motivation	NN	B-NP	O
,	,	,	O	O
And	and	CC	O	O
Beyond	beyond	IN	B-PP	O
Caring	care	VBG	B-NP	O
.	.	.	O	O
People	People	NNS	B-NP	O
experiencing	experience	VBG	B-VP	O
Burnout	burnout	NN	B-NP	O
Often	often	RB	B-ADVP	O

don’t	don’t	VBP	B-VP	O
See	see	VB	I-VP	O
Any	any	DT	B-NP	O
Hope	hope	NN	I-NP	O
Of	of	IN	B-PP	O
Positive	positive	JJ	B-NP	O
Change	change	NN	I-NP	O
In	in	IN	B-PP	O
Their	their	PRP$	B-NP	O
Situations	situation	NNS	I-NP	O
.	.	.	O	O
If	If	IN	B-SBAR	O
Excessive	excessive	JJ	B-NP	O
Stress	stress	NN	I-NP	O
Is	be	VBZ	B-VP	O
Like	like	IN	B-PP	O
Drowning	drown	VBG	B-VP	O
In	in	IN	B-PP	O
responsibilities	responsibility	NNS	B-NP	O
,	,	,	O	O
Burnout	burnout	NN	B-NP	O
Is	be	VBZ	B-VP	O
Being	be	VBG	I-VP	O
All	all	DT	O	O

Dried	dry	VBD	B-VP	O
Up	up	RP	B-PRT	O
.	.	.	O	O
And	And	CC	O	O
While	while	IN	B-SBAR	O
you’re	you’re	PRP	B-NP	O
Usually	usually	RB	B-ADVP	O
Aware	aware	JJ	B-ADJP	O
Of	of	IN	B-PP	O
Being	be	VBG	B-VP	O
Under	under	IN	B-PP	O
A	a	DT	B-NP	O
Lot	lot	NN	I-NP	O
Of	of	IN	B-PP	O
Stress	stress	NN	B-NP	O
,	,	,	O	O
You	you	PRP	B-NP	O
don’t	don’t	VBP	B-VP	O
Always	always	RB	B-ADVP	O
Notice	notice	RB	B-ADVP	O

Burnout	burnout	RB	I-ADVP	O
When	when	WRB	B-ADVP	O
It	it	PRP	B-NP	O
Happens	happen	VBZ	B-VP	O

Popis izvora

Text analysis Online. (2016). *NLTK Wordnet Lemmatizer*. Abgerufen am 2017 von http://textanalysisonline.com/nltk-wordnet-lemmatizer

Algorithmia. (11. August 2016). Abgerufen am 5. July 2017 von https://blog.algorithmia.com/introduction-natural-language-processing-nlp/

ACL WEB. (kein Datum). Abgerufen am June 2017 von http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)

Aylett, J. (kein Datum). *Tatarus*. Von http://snowball.tartarus.org/algorithms/lovins/festschrift.html abgerufen

Aylett, J. (kein Datum). *Tatarus*. Abgerufen am 5. June 2017 von http://snowball.tartarus.org/algorithms/porter/stemmer.html

Dennis, S. d. (kein Datum). *Ohio State University.* Von http://mall.psy.ohio-state.edu/DSTO2.pdf abgerufen

Dresen, T. U. (2006). *Computational-logic.org*. Abgerufen am July 2017 von http://www.computational-logic.org/iccl/master/lectures/summer06/nlp/part-of-speech-tagging.pdf

Eineborg M., Lindberg N. (1998). Induction of Constraint Grammar-rules using Progol.

Elastic. (kein Datum). Abgerufen am 15th. January 2017 von https://www.elastic.co/guide/en/elasticsearch/guide/current/stemming.html

Help guide. (kein Datum). Abgerufen am 2017 von https://www.helpguide.org/articles/stress/burnout-prevention-and-recovery.htm

Jivani, A. (2016). A comparatve stuy of stemming algorithms. *ICCOINS2016.*

Jurafsky, D., & Martin, J. (2016). *Speech and Language Processing.*

Liberman, M. (2003). *Linguistics 001*. Von University of Pennsylvania: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html abgerufen

Manning, C. D. (2008). Abgerufen am 15th. January 2017 von Stanford University: http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Martinčić-Ipšić, S. (2007). *RASPOZNAVANJE I SINTEZA HRVATSKOGA GOVORA KONTEKSTNO OVISNIM SKRIVENIM MARKOVLJEVIM MODELIMA.*

McKenzie, C. (kein Datum). Von http://9ol.es/porter_js_demo.html abgerufen

McKenzie, C. (kein Datum). Abgerufen am 5. June 2017 von http://9ol.es/porter_js_demo.html

McKenzie, C. (2010). *Porter json demo*. Von http://9ol.es/porter_js_demo.html abgerufen

Meyers, A. (2012). *New York University*. Abgerufen am July 2017 von http://cs.nyu.edu/courses/spring12/CSCI-GA.2590-001/lecture4.pdf

Moral, C. (2014). *Escuela Técnica Superior de Ingenieros Informáticos.* Von http://www.informationr.net/ir/19-1/paper605.html#.WIXfclMrLIV abgerufen

Nau, D. S. (2010). *University of Maryland.* Abgerufen am June 2017 von https://www.cs.umd.edu/~nau/cmsc421/part-of-speech-tagging.pdf

NlpDotNet. (2009). *NLP dot net*. Abgerufen am June 2017 von http://nlpdotnet.com/services/Tagger.aspx

Northwood, C. (2009). *cjn.org*. Abgerufen am 2017 von http://www.pling.org.uk/cs/com6791.html

NTLK. (2017). *Text-processing*. Von http://text-processing.com/demo/stem/ abgerufen

Pandžić, I. (9. October 2015). Oblikovanje korijenovatelja za hrvatski jezik. *Časopis instituta za hrvatski jezik i jezikoslovlje*, S. 301-327.

Porter, M. (1980). *Snowball Tatarus*. Abgerufen am 15. april 2017 von http://snowball.tartarus.org/algorithms/porter/stemmer.html

Porter, M. (1980). *Tartarus*. Abgerufen am 15. April 2017 von http://snowball.tartarus.org/algorithms/porter/stemmer.html

Porter, M. (2001). Lovins revisited. *Charting a New Course: Progress in Natural Language Processing and Information Retrieval: A Festschrift for Professor Karen Sparck Jones,*.

Porter, M. (Jan 2006). *Tartarus*. Abgerufen am 2017 von https://tartarus.org/martin/PorterStemmer/voc.txt

Porter, M. (kein Datum). *Tatarus*. Von tatarus.org: https://tartarus.org/martin/PorterStemmer/voc.txt abgerufen

Porter, M. (kein Datum). *Tatarus*. Abgerufen am June 2017 von https://tartarus.org/martin/PorterStemmer/voc.txt

Princeton University. (2015). *Wordnet, Princeton*. Abgerufen am 2017 von https://wordnet.princeton.edu/

Robin. (2009). *World of Computing*. Abgerufen am 2017 von http://language.worldofcomputing.net/pos-tagging/rule-based-pos-tagging.html

Rowling, J. (1997). *Harry Potter and the Philosopher's Stone.* London: Bloomsbery Publishing Plc.

Ruef, B. (2003). *University of Zurich.* Abgerufen am 2017 von Transformation based learning and Part-of-Spech Tagging of Old English: https://files.ifi.uzh.ch/cl/gschneid/KorpusSeminar/

Stanford NLP Group. (2015). *University of Stanford*. Abgerufen am 2017 von https://nlp.stanford.edu/software/tagger.shtml

Šnajder, J. (2011). *Postupci morfoloske normalizacije u pretraživanju i klasifikaciji teksta.* Fakultet elektrotehnike i racunarstva Sveucilista u Zagrebu.

Text analysis online. (kein Datum). Abgerufen am 4. June 2017 von http://textanalysisonline.com/nltk-wordnet-lemmatizer

Text analysis online. (kein Datum). Abgerufen am 5. June 2017 von http://textanalysisonline.com/nltk-lancaster-stemmer

Text analysis online. (kein Datum). Abgerufen am June 2017 von http://textanalysisonline.com/nltk-wordnet-lemmatizer

Text Analysis Online. (2016). *NLTK Lancaster Stemmer*. Abgerufen am 4. June 2017 von http://textanalysisonline.com/nltk-lancaster-stemmer

Toutanova, K., & Klein, D. (2012). *The Stanford Natural Proessing Group*. Abgerufen am 2017 von http://nlp.stanford.edu:8080/parser/index.jsp

Tsuruoka, Y. (2006). *University of Manchester*. Abgerufen am June 2017 von http://nactem7.mib.man.ac.uk/geniatagger/a.cgi

Voutilainen, A. (23. August 2000). *ENGCG-intro*. Abgerufen am 1. August 2017 von Lingsoft: http://www2.lingsoft.fi/doc/engcg/intro/

W3-Corpora Project. (1998). *The Brown Corpus.* Abgerufen am 2017 von University of Essex: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

wenhoujx.blogspot.hr. (3rd. June 2013.). *wenhoujx.blogspot.hr.* Abgerufen am 20th. January 2017. von http://wenhoujx.blogspot.hr/2013/06/penn-treebank-pos-tags.html

Wiese, A., Ho, V., & Hill, E. (2011). *A comparison of Stemmers on Source Code Identifiersfor Software Search.* Montclaire: Montclaire State University.

Wiese, A., Ho, V., & Hill, E. (kein Datum). *A comparison of Stemmers on Source Code Identifiersfor Software Search.* Montclair.

Wordnet, Princeton. (kein Datum). Abgerufen am 2. June 2017 von https://wordnet.princeton.edu/