UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

DUŠAN BOŽIČ

Avtomatsko povzemanje besedil s pomočjo semantične analize

MAGISTRSKO DELO

MENTOR: prof. dr. Igor Kononenko

Ljubljana, 2016

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*

Večna pot 113
1000 Ljubljana, Slovenija
telefon: 01 47 98 100
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Številka: 156-MAG-RI/2016
Datum: 06. 04. 2016

Dušan BOŽIČ, univ. dipl. inž. rač. in inf.

L j u b l j a n a

Fakulteta za računalništvo in informatiko Univerze v Ljubljani izdaja naslednjo magistrsko nalogo

Naslov naloge: **Avtomatsko povzemanje besedil s pomočjo semantične analize**

Automatic text summarization using semantic analysis

Tematika naloge:

Avtomatsko povzemanje besedil je računalniški proces obravnave naravnega jezika, katerega cilj je pridobiti krajši povzetek ali izvleček besedila. Obravnavana tema je pomembna, ker nam omogoča vpogled in analiziranje naravnega jezika, ki ima za razliko od umetnih jezikov kompleksno in dvoumno gramatiko. Kljub temu, da so se prvi algoritmi povzemanja začeli razvijati že leta 1958, je razmah socialnih omrežij povzročil, da se dovršeni sintaktični pristopi nadgrajujejo s semantičnimi, ki zahtevajo obravnavo in upoštevanje pomena besedila. Tema magistrske naloge Semantična nadgradnja obstoječega sintaktičnega algoritma, ki deluje s pomočjo analize arhetipov.

V magistrski nalogi obravnavajte besedila, napisana v slovenskem jeziku. V fazi semantičnega predprocesiranja z algoritmom Latent Semantic Analysis pridobite stavke, ki so najboljši kandidati za izvleček. Po izvedeni dekompoziciji stavkov na posamezne besede izpeljite njihovo normalizacijo v nedoločniku in skladenjsko analizo. Označene besede povežite v semantične relacije na podlagi pomena besed v stavku v ontološki anotaciji (angl. Abstract Meaning Representation), zgenerirajte nove stavke s pomočjo lastnih pristopov in dobljeni izvleček primerjajte s povzetki, dobljenimi z analizo arhetipov.

M e n t o r :

prof. dr. Igor Kononenko

D e k a n :

prof. dr. Nikolaj Zimic

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Zahvaljujem se mentorju prof. dr. Igorju Kononenku za strokovne napotke in usmeritve pri izdelavi magistrske naloge.

Zahvaljujem se dr. Ercanu Canhasiju za nesebično pomoč pri interpretaciji njegove doktorske dizertacije in pristopa ekstraktnega povzemanja s pomočjo analize arhetipov.

Zahvaljujem se doc. dr. Tomažu Erjavcu za vpeljavo v svet slavističnih virov na področju raziskovanja slovenskega jezika in za posredovanje elektronskih virov, ki so predmet obravnave magistrskega dela.

Zahvaljujem se sodelavcu Srečku Mandlju za strokovno pomoč pri izdelavi lastne programske rešitve, ki je rezultati magistrske naloge.

Zahvaljujem se doc. dr. Darji Fišer za pomoč pri dostopu in razlagi slovenskih digitalni viri.

Zahvaljujem se staršem in prijateljem za podporo ter razumevanje tekom študija.

Zahvaljujem se ženi dr. Maji Šubelj za pomoč pri lektoriranju in prevajanju magistrske naloge. Zahvaljujem se ji za vse vzpodbudne besede, ki so odločno pripomogle k dokončanju teme, kjer ne manjka multidisciplinarnih izzivov.

Kazalo

Seznam slik

Seznam tabel

Slovar pojmov

Prazne besede (angl. *stop words*) – so besede, ki predstavljajo lepilo pri vzpostavitvi stavčne oblike, ki ustreza gramatiki, vendar ne vplivajo na pomen stavka. Te besede so v fazi predprocesiranja naravnih jezikov odstranjene, ker so zelo pogoste in motijo značilke povezane s frekvenco pojavitve besed [19].

Pogoste besede (angl. *common words*) – seznam pogostih besed ni dokončen in je rezultat analize zelo velikega števila besedil. V primeru *Oxford English* besedil je bilo upoštevanih preko milijarde besed. *The Reading Teacher's Book of Lists* trdi, da prvih 25 besed s seznama predstavlja eno tretjino vseh besed v vseh angleških besedilih, prvih 100 besed pa polovico vseh besed napisanega materiala [20].

Polisemija (angl. *polysemy*) – polisemija je asociacija z eno ali več besedami, ki imajo različen pomen. Polisem je beseda ali fraza z več pomeni. Nasprotno imenujemo natančno ujemanje besede s pomenom monosemija (angl *monosemy*). Monosemijo lahko najdemo v specializiranih slovarjih, ki se nanašajo na tehnične teme [22].

Izvlečki (angl. *extractive summary*) – izvlečki so krajši sestavki besedila, pri katerih so stavki izbrani iz izvornega besedila in kot taki tvorijo novo ter krajše besedilo, ki ohranja pomen in koncepte izvornega besedila.

Povzetki (angl. *abstractive summary*) – povzetki so krajši sestavki besedila, pri katerih se ohranja pomen in koncepte izvornega besedila, vendar le-ti ne ustrezajo izvornim stavkom.

Povezovalna beseda/fraza (angl. *cue word/phrase*) – beseda ali fraza, ki povezuje dele besedila in označuje določeno semantično relacijo v diskurzu besedila. Primeri takih besed so: mimogrede (angl. *incidentally*), na primer (angl. *for example*), kakorkoli (angl. *whatever*), vseeno (angl. *anyway*), mimogrede (angl. *by the way*), poleg tega (angl. *furthermore*), prvi (angl. *first*), drugi (angl. *second*), potem (angl. *then*), sedaj (angl. *now*), tako (angl. *thus*), poleg tega (angl. *moreover*), zato (angl. *therefore*), zato (angl. *hence*), nazadnje (angl. *lastly*), končno (angl. *finally*), po drugi strani (angl. *on the other hand*) [3].

Preprosti pristop (angl. *shallow approach* ali *surface level approach*) – najstarejši pristop, kjer se uporabljajo preprosti indikatorji, na podlagi katerih se odloča, kateri deli besedila so pomembni. Primeri takih indikatorjev so: frekvenca pojavitev besed, pojavitev pogostih besed, lokacija besed v stavku, pojavitev besed iz naslova, ipd.

Anafora (angl. *anaphoric expression*) – je izraz, katerega interpretacija je odvisna od drugega izraza v skupnem kontekstu [35].

Neredundantnost (angl. *non-redundancy*) – je lastnost, da so vsi podatki informativni tj., da se informativnost podatkov niti ne ponavlja niti je odveč.

Lematiziranje (angl. *lemmatization*) – je spreminjanje besedne vrste iz skladenjske oblike v nedoločnik na podlagi digitalnega vira.

Krnjenje (angl. *stemming*) – je alternativa lematiziranju, ker ne uporablja digitalnih virov in je hevristični proces odstranjevanja prefiksov ali postfiksov.

Slovar kratic

LSA (angl. *Latent Semantic Analysis*) – latentna semantična analiza;

SVD (angl. *Singular Value Decomposition*) – dekompozicija singularne vrednosti;

ATS (angl. *Automatic Text Summarization*) – avtomatsko povzemanje besedila;

AA (angl. *Archetypal Analysis*) – analiza arhetipov;

ROUGE (angl. *Recall-Oriented Understudy for Gisting Evaluation*) – algoritem ocenjevanja;

CRUD (angl. create, read, update and delete) – so štiri osnovne operacije za delo z bazo;

LBS (angl. local database simplex) – lokalna baza programa za povzemanje SimpleX;

CLOB (angl. character large object) – objekt, ki lahko sprejme do 2.147.483.647 znakov;

XML (angl. Extensible Markup Language) – jezik za opisovanje strukturiranih podatkov;

PDF (angl. Portable Document Format) – odprt standard za izmenjavo dokumentov;

MSD (angl. Master Standard Data) – oblikoskladenjska oznaka;

AMR (angl. Abstract Meaning Representation) – anotacija predstavitve pomena;

Povzetek

V magistrski nalogi smo uporabili metodo latentne semantične analize (LSA) za avtomatsko povzemanja besedila. Algoritem LSA analizira razmerja med besedami in dokumentom za izdelavo konceptov, ki opisujejo ta razmerja. V fazi predprocesiranja smo vse besede lematizirali s pomočjo slovenskega leksikona. V nalogi se povzemajo slovenski akademski prispevki, zato se uporabljajo slovenski digitalni viri. Rezultat analize LSA so odstavki, rangirani po pomembnosti. Najbolj obetavni odstavki so kandidati za povzetek. Za pravilno preslikavo lematiziranih odstavkov v izvorne smo v fazi predprocesiranja izvedli skladenjsko analizo izvornih besed. Pridobljeni izvleček smo spremenili v abstrakni povzetek s pomočjo semantične analize stavkov in leksikalnega veriženja. Pri tem smo uporabljali slovenski morfološki leksikon. Kvaliteto pridobljenih povzetkov smo ocenili s pomočjo algoritma ROUGE. Primerjali smo jih z izvlečki analize arhetipov in človeškimi povzetki. Za izvajanje povzemanja smo implementirali samostojen spletni program pod imenom SimpleX, ki se izvaja v strežniškem okolju s podporo podatkovne baze. Eksperimentalni rezultati kažejo, da predlagani semantični pristop omogoča povzemanja obsežnih dokumentov.

Ključne besede: povzemanje naravnih besedil, semantična analiza, leksikalno veriženje, analiza arhetipov, ekstraktivno povzemanje, semantično povzemanje, povzetki, izvlečki, SimpleX.

Abstract

In this thesis, we used a method of latent semantic analysis (LSA) for automatic multi-document summarization. LSA algorithm analyzes the relationships between words and document by producing a set of concepts that describe this relationship. In the preprocessing stage, all words were lemmatized based on Slovenian lexicon. Our work reiterated Slovenian academic contributions to science acquired from the Slovenian digital lexicons. The results of the LSA analysis are paragraphs ranked by relevance. The most promising paragraphs are candidates for the summary. For the proper mapping of the lemmatized paragraphs into the original in the phase of preprocessing, we performed syntactical analysis of the source text. The resulting extract was changed into the abstract summary, using semantic analysis of sentences and lexical chaining. For this purpose, we used Slovenian morphological lexicon. The quality of the obtained summaries was evaluated using the Rouge algorithm. We compared those summaries with abstracts from the analysis of archetypes and human summaries. To implement the summarization, we implemented a stand-alone web application named SimpleX, which was implemented in a server environment to support the database. Experimental results show that the proposed semantic approach helps to build a way towards the large collections of documents.

Keywords: natural text summarization, semantic analysis, lexical chaining, analysis of archetypes, extraction-based summarization, semantic-based summarization, summaries, abstracts, SimpleX.

POGLAVJE 1

1. Uvod

Naravni jezik je katerikoli jezik, ki se je spontano razvijal hkrati z razvojem človeka. Človek je skozi svojo zgodovino razvil različne manifestacije jezika za medsebojno komuniciranje. Poleg govorne komunikacije poznamo tudi pisano in simbolno. Z vzponom računalnikov se je pojavila potreba po primerni komunikaciji uporabnika z računalniki. Komunikacija naj bi bila čim bližje naravnem jeziku. Vsi naravni jeziki so dvoumni, ker vsebujejo dvoumne stavke. Gramatika, ki generira dvoumne stavke, je dvoumna gramatika in iz stališča računalniške izračunljivosti predstavlja časovno zahteven problem. Zato so se v začetku razvoja računalništva razvili približki naravnemu jeziku. Formalnim jezikom, ki jih je človek razvil s premislekom za potrebe na področju računalništva, rečemo umetni ali programski jeziki. Ti jeziki imajo nedvoumno gramatiko. Iz stališča računalnikov, ki so deterministični stroji, je prepoznava jezika nedvoumne gramatike rešljiv izziv. S pojavom kontroliranih jezikov smo se približali naravnemu jeziku. Pri kontroliranih jezikih uporabljamo stavke iz naravnega jezika, ki so nedvoumni in imajo posledično nedvoumno gramatiko. Uporabljamo jih tako na področjih človekovega delovanja, kjer ni prostora za dvoumnost, kot tudi v računalništvu za predstavitev znanja na uporabniku prijazen način. Razmah socialnih omrežij, razvoj komercialne oglaševalske industrije in nenazadnje boj proti terorizmu je postavil procesiranje govorne in pisane besede naravnega jezika z vsemi svojimi posebnostmi v ospredje svetovnega zanimanja.

1.1 Raziskovalne teme

Cilji magistrske naloge so bili:

- predstaviti področje raziskovanja naravnih jezikov in njihove lastnosti;
- raziskati področje računalniškega procesiranja naravnih jezikov, opisati pomembne metode in tehnike ekstraktnega in semantičnega povzemanja besedil;
- predstaviti lastno implementacijo programske rešitve za povzemanje naravnih jezikov s pomočjo semantične analize;
- oceniti rezultate povzemanja s knjižnico/oceno ROUGE;
- predstaviti možnost nadgradnje semantične analize s pomočjo tehnike analize predstavitve pomena (angl. *Abstract Meaning Representation*) kot primer nadaljevanja raziskav in dela na področju povzemanja besedil;
- preveriti naslednji hipotezi:
    - izvlečki na podlagi analize arhetipov so boljši/učinkovitejši od povzetkov s pomočjo semantične analize glede na oceno ROUGE;
    - človeški povzetki akademskih del so resnično povzetki celotnega dela in ne samo nekaj strani.

1.2 Prispevek naloge

Razvili smo lastno programsko rešitev povzemanja naravnih jezikov pod imenom SimpleX. SimpleX je spletni program. Za interakcijo smo razvili uporabniku prijazen grafični vmesnik. Rešitev smo objavili na aplikacijskem strežniku. Vse nastavitve in rezultati so shranjeni v podatkovni bazi, kar omogoča ponovljivost, preglednost, analizo in ocenjevanje razvitih tehnik povzemanja naravnih besedil. V fazi predprocesiranja se s pomočjo knjižnice CoreNLP, ki so jo razvili raziskovalci z Univerze v Stanfordu, izvede dekompozicija stavčnih oblik na besede, s pomočjo elektronskih slovarjev Filozofske Fakultete Univerze v Ljubljani pa se nato izvede njihova skladenjska analiza in normalizacija. V fazi procesiranja se s pomočjo knjižnic TML z Univerze v Sydneyju ekstraktivno povzame tiste odstavke, ki jih metoda LSA (angl. *Latent Semantic Analysis*) prepozna kot primerne kandidate za končni povzetek na podlagi pomena. Te odstavke se doda na seznam obetavnih kandidatov za semantično analizo. Kot prispevek znanosti smo razvili semantična pravila, uporabljena pri izdelavi končnega povzetka. Semantična pravila določajo, kako se s tehniko leksikalnega veriženja zamenjuje besede, ki določajo pomen stavka s sopomenkami, nadpomenkami ali podpomenkami. SimpleX nad seznamom obetavnih kandidatov za semantično analizo izvede tehniko leksikalnega veriženja in tako pridobi končni povzetek naravnega besedila s pomočjo semantične analize. Rezultate povzemanja s pomočjo semantične analize smo želeli primerjati z rezultati povzemanja ekstraktnega povzemanja, ki ga je razvil dr. Ercan Canhasi v okviru doktorske dizertacije. Njegova implementacija analize arhetipov v Matlabu je integrirana s programom SimpleX. Integracija omogoča, da se izvajanje proži iz grafičnega okolja SimpleX, rezultati se shranjujejo v podatkovno bazo in izdelani ekstraktni izvlečki primerjajo s semantičnimi povzetki s pomočjo knjižnice ROUGE. Sistemsko izdelani izvlečki in povzetki se primerjajo med sabo in tudi s človeškimi povzetki. Novi prispevek znanosti bi lahko bil preveritev teze, ali so človeški povzetki besedil povzetki pomena celotnega besedila

ali samo prvih nekaj strani besedila. SimpleX rezultate povzemanja prikaže tabelarično in vizualno s pomočjo grafov. Grafi prikazujejo trend ocen izboljšav tehnik povzemanja med lastnimi semantičnimi povzetki, ekstraktnimi povzetki na podlagi analize arhetipov in človeškimi povzetki. Uporaba in implementacija ontotološke anotacije (angl. *Abstract Meaning Representation*) je presegla okvir naloge, zato je ta tehnika teoretično opisana v zadnjem odseku. V zadnjem odseku naloge so predstavljene tudi možnosti za nadaljnje delo in raziskovanje.

1.3 Organizacije naloge

Magistrsko delo je organizirano na naslednji način:

- Poglavje 1: zajema uvod, predstavi raziskovalne teme in potencialen prispevek k znanosti.

- Poglavje 2: predstavi zgodovino raziskovanja človeških besedil in motivov za današnji obstoj dveh različnih pristopov: deterministični in statistični. Predstavljene so lastnosti jezikov, ki jih srečamo pri računalniškem povzemanju, upoštevaje pomen. Predstavljeni so različni vidiki povzemanja besedil, odvisno od tega, kakšen je namen povzetka, komu je namenjen, kakšno predznanje ima tako tisti, ki izdeluje povzetek, kot tudi uporabnik povzetka, ipd. Glede na način povzemanja ločimo dve vrsti povzetkov. Opisane so prednosti in slabosti obeh vrst povzetkov.

- Poglavje 3: predstavi zgodovino računalniškega procesiranja naravnih jezikov. Poleg osnovnih metod in tehnik so predstavljene tudi tiste, ki so bile uporabljene v lastnem sistemu povzemanju SimpleX (npr. analiza arhetipov, latentna semantična analiza, leksikalno veriženje). Poglavje je razdeljeno na ekstraktivno povzemanje, opis teorije strukture jezikov kot zgodnji primer razmišljanja o mehanizmu, ki bi opisal strukturo jezika, in relacije med deli besedila ter nazadnje na semantično povzemanje.

- Poglavje 4: predstavi lastno implementacijo sistema za povzemanje naravnih jezikov s pomočjo semantične analize. Opisana je vzpostavitev razvojnega in produkcijskega okolja, uporabniške maske sistema in vključene tehnike povzemanja s pomočjo tujih knjižnic (tj. Stanford CoreNlp, Sydney University TML), podpornih sistemov povzemanja (tj. dr. Ercan Canhasi z analizo arhetipov) in s pomočjo elektronskih slovarjev Filozofske Fakultete Univerze v Ljubljani. Opisan je zajem vhodnih besedil, faze predprocesiranja, stavčne dekompozicije, semantična analiza, izdelava povzetka in vizualizacija rezultatov.

- Poglavje 5: predstavi okolje za ocenjevanje povzetkov s pomočjo knjižnice ROUGE. Predstavi rezultate povzemanja in njihovo primerjavo med sistemskimi povzetki kot tudi človeškimi povzetki.

POGLAVJE 2

2. Raziskovanje naravnih besedil

Od leta 1960 do 1985 je struja t. i. racionalistov zagovarjala raziskovanje besedila na način, ki ne upošteva povezav med mentalno prezentacijo jezika in njeno manifestacijo v pisani ali govorni obliki. Tak pristop omogoča takojšen preskok k praktični dimenziji raziskovanja jezikov. Podlaga za takšno razmišljanje je bilo prepričanje, da večji del človeškega znanja naravnih jezikov ni izpeljan na podlagi zaznavanja s čutili, temveč na podlagi predefiniranega genetskega zapisa. Podpora takemu pristopu je t. i. problem oslabelega stimulusa, ki ga premore otrok v svojem zgodnjem razvoju. Raziskovalci so bili prepričani, da se otrok v zgodnji fazi razvoja ne more naučiti kaj tako kompleksnega, kot je naravni jezik skozi osiromašen oz. omejen pritok informacij preko čutil in njegovo interpretacijo, zato so zagovarjali tezo, da je obvladovanje naravnega jezika možno le na podlagi predefiniranih struktur, ki natančno opredeljujejo posamezen jezik. Te strukture so v naših možganih že vnaprej opredeljene. Prepričanje se je še posebej utrdilo zaradi vse splošnega sprejemanja argumentov jezikoslovca in prof. na Tehnološkem Inštitutu v Massachusettsu MIT (angl. *The Massachusetts Institute of Technology*) Noama Chomskyja. Prof. Chomsky je avtor hierarhije Chomskega in razvrstitve formalnih jezikov. Hierarhija Chomskega je zbirka štirih razredov formalnih jezikov. Vsi razredi formalnih jezikov ustrezajo določeni gramatiki, ki jih generira, in stroju, ki jih prepoznava. Na vsakem nivoju hierarhije so pravila bolj omejujoča. Pri tem postane jezik podrejen drugemu jeziku oz. je v relaciji glede na vsebovanost. Razredi jezikov od bolj omejujočih proti manj so regularni jeziki; kontekstno neodvisni jeziki; kontekstno odvisni jeziki in Turingovi jeziki oz. jeziki brez omejitev [7].

Poleg deterministočnega pristopa poznamo danes tudi statistični pristop. Namesto, da se sprašujemo, ali nek stavek ustreza določeni gramatiki ali ne, se vprašamo, kateri so tisti vzorci, ki so skupni pri uporabi jezika. Glavno orodje za identifikacijo teh vzorcev je štetje oz. statistika. S pomočjo statistik dobimo predstavo o tem, kako so zgrajeni statistični modeli jezika. Podporniki te ideje so t. i. empiriki, ki zagovarjajo, da otroški možgani v svojem zgodnjem razvoju uporabljajo splošne operacije za asociacije, prepoznavanje vzorcev in posploševanje. Ta pristop uporabijo nad zaznavanji s čutili, kar je osnova za učenje kompleksnih struktur naravnega jezika. Empirični pristop predlaga učenje kompleksnih in obsežnih struktur jezika na podlagi primernega splošnega modela jezika. V ta namen izvedemo izpeljavo vrednosti parametrov z uporabo statističnih metod, prepoznavo vzorcev in metodami strojnega učenja nad primeri uporabe naravnega jezika [7].

Metode strojnega učenja, s katerimi lahko raziskujemo, vključujejo klasično statistiko, odločitvena klasifikacijska in regresijska drevesa, naivni Bayesov klasifikator, verjetnostne mreže, umetne nevronske mreže, podporne vektorje in druge jedrne metode, metode k-najbližjih sosedov, induktivno logično programiranje, povezovalna pravila, hierarhično in verjetnostno razvrščanje, mehko razvrščanje, indukcijo nevronsko-mehkih pravil ter genetske algoritme [8].

2.1 Jezikovne strukture

Jezikovna znanost je znanost, ki raziskuje in razlaga jezikovne strukture in posebnosti ter njihove relacije. Jezikovne relacije vplivajo na kognitivne sposobnosti človeka, ki ustvarja in razume jezik za medsebojno komunikacijo. Za lažje razumevanje si je človek omislil jezikovna pravila, na podlagi katerih strukturiramo jezikovne izraze. Tak pristop obstaja že več kot 2000 let [7].

Jezike lahko ločimo na podlagi osnovnih gramatičnih struktur [9]:

- izolirani jeziki: uporaba nespremenljivih besed skupaj s strogimi pravili glede vrstnega reda besed in s tem ohranjanje razumljivosti gramatičnega pomena;
- aglutinacijski jeziki: uporaba regularnih predpon in pripon glavnih besed za izražanje podrobnosti. Te jezike se lahko hitro naučimo;
- infleksijski jeziki: poleg predpon in pripon, se pojavljajo spremenjene besede osnovnih besed. Teh jezikov se težje naučimo zaradi številnih izjem;
- amalgamatični jeziki: vsebujejo tudi kompleksne besede, ki določajo kontekst stavka. Teh jezikov se zelo težko naučimo, razen če niso materin jezik.

Drugi način klasifikacije je na podlagi uporabe vrstnega reda besed [9]:

- SOV (ang. *Subject-Object-Verb*) skupina jezikov. Ta struktura je najbolj pogosta pri jezikih kot so npr. indoevropski jeziki indije, armenski, madžarski, turški, korejski, japonski, brunejski, baskovski in avstralsko aboridžinski jeziki. Večina teh jezikov uporablja tako besede s predponami kot tudi besede s priponami. Besede s predponami in priponami izražajo prostorske in časovne relacije. V teh jezikih običajno pridevniku sledi samostalnik, izjemoma obratno.

- SVO (ang. *Subject-Verb-Object*) skupina jezikov. Ta skupina jezikov je druga največja skupina. Razdeljena je na dve skupini, ki uporabljata besede s priponami ali besede s predponami. Med jezike, ki uporabljajo predpone, uvrščamo npr. romanski, albanski, grški jeziki, jezike jugovzhodne Azije in germanske jezike. Pri večini teh jezikov samostalniku sledi pridevnik, razen pri germanskih, kjer je pridevnik pred samostalnikom. Druga skupina, ki uporablja zgolj pripone, so npr. kitajski, finski, estonski, nekateri južno ameriški indijski jeziki in nekateri afriški jeziki kot je npr. mandingo. Nekateri raziskovalci so mnenja, da kitajski jezik postaja SOV.

- VSO (ang. *Verb-Subject-Object*) skupina jezikov. Ta skupina jezikov vedno uporablja predpone. Primeri teh jezikov so npr. arabski, židovski, keltski jeziki, polinezijski in nekateri ameriško indijski jeziki kot so npr. britansko kolumbijski in azteški jeziki. Večina jih ima pridevnik za samostalnikom.

- VOS (ang. *Verb-Object-Subject*) skupina jezikov. Tej skupini jezikov pripadajo npr. severno ameriški jeziki in kanadsko indijski jeziki.

Poznamo tudi jezike, ki jih ne moremo strogo razvrstiti v naštete skupine, saj uporabljajo lastnosti več skupin. Primer je npr. angleščina, ruščina in latinščina, vendar zadnja dva jezika pogosto dovolita tudi spremenljiv vrstni red.

2.2 Povzemanje naravnega jezika

Povzetek naravnega besedila lahko opredelimo kot besedilo, ki vsebuje najpomembnejše informacije prvotnega besedila in je hkrati strnjeno v dolžino, ki je krajša za vsaj polovico od prvotne dolžine. Razumevanje besedila je tesno povezano s ključnimi informacijami, ki so bralcu za razumevanje besedila bistvenega pomena. Pri tem naletimo na problem, kaj je za koga ključna informacija znotraj besedila. Izkaže se, da imajo različni uporabniki različno mnenje o tem, kaj je v besedilu pomembno.

Zato poznamo več vidikov povzemanja besedila. Ločimo tri glavne vidike povzemanja: vhod, namen in izhod [1, 10].

2.2.1 Vidik vhoda povzemanja besedila

Značilnosti vhodnega besedila odločilno vplivajo na zajem besedila v sistem povzemanja. Nekaj vidikov vhoda v sistem povzemanja:

(I.) Struktura dokumenta: v dokumentu se poleg same vsebine pojavljajo tudi meta informacije kot so oznake oz. labele, ki označujejo poglavja, odseke, tabele, grafe, ipd. Če so ti podatki pravilno zajeti, se lahko uporabijo pri sami analizi dokumenta.

(II.) Dolžina dokumenta: sistemi povzemanja pričakujejo različne dolžine vhodnega besedila. V primeru obravnave članka so posamezni stavki obravnavani kot najmanjše pomenske enote. Pri daljših besedilih kot so npr. poročila ali knjige se išče pomen znotraj skupine stavkov ali odstavkov.

(III.) Število vhodnih besedil: Sistem povzemanja lahko sprejme enega ali več vhodnih besedil kot vhod v analiziranje. Nekateri sistemi obravnavajo samo posamezne dokumente, medtem ko znajo drugi obravnavati več dokumentov hkrati. V tem primeru si morajo obravnavani dokumenti deliti relacije o skupni temi. Sistem povzemanja pri izdelavi povzetka uporablja te relacije. Primer več dokumentnega sistema povzemanja je npr. SUMMONS (angl. *SUMMarizing Online NewS articles*) [10].

(IV.) Raznolikost naravnih jezikov: Na podlagi raznolikosti jezikov vhodnega besedila ločimo enojezikovne in večjezikovne sisteme povzemanja. Enojezikovni sistemi povzemanja sprejemajo vhodna besedila v izključno izbranem naravnem jeziku in v istem tudi zgenerirajo povzetek. Primer enojezikovnega sistema je npr. FarsiSUM (angl. *Persian Text Summarizer*) [14]. Večjezikovni sistemi povzemanja pa sprejemajo besedila v različnih jezikih. Jezik povzetka lahko določi uporabnik. Primer večjezikovnega sistema je npr. SUMMARIST [10].

2.2.2 Vidik namena povzemanja besedila

Ločimo tri vrste povzetkov glede na omejitev predznanja in s tem povezanim pomenom povzetka:

(I.) Žanrski sistemi povzemanja besedil: sprejmejo samo posebne tipe besedila, kjer je omejitev predznanja določena na podlagi predlog besedila. Predloge besedila so npr. časopisni članki, znanstveni članki, zgodbe in priročniki. Sistem si pri povzemanju pomaga z lastnostmi, ki so povezane s specifičnimi strukturami samih predlog besedila.

(II.) Neodvisni sistemi povzemanja besedil: le-ti ne poznajo predefiniranih omejitev in lahko sprejemajo različne tipe besedil.

(III.) Domenski sistemi povzemanja besedil: so specializirani za povzemanje besedil, katerih predmet je povezan z domeno. Domenski sistemi povzemanja imajo popolno predznanje o domeni, ki jo obravnavano besedilo opisuje. To predznanje pri teh sistemih uporabljamo kot podlago za povzemanje te vrste besedil. Primer domenskega sistema je npr. TRESTLE [10].

2.2.3 Vidik izhoda povzemanja besedila

(I.) Kontekst: na podlagi konteksta povzetka ločimo generične in poizvedbene povzetke. V generičnih povzetkih se sistem ne nanaša na temo dokumenta. Uporabnik nima nobenega predznanja o besedilu. Predpostavljamo, da bodo povzetke uporabljali različni uporabniki, zato imajo posledično vse informacije znotraj besedila enako stopnjo pomembnosti. V nasprotju mora uporabnik pri poizvedbenih povzetkih pred procesom povzemanja določiti interesno temo originalnega besedila v obliki poizvedbe. Uporabnik ima splošno predznanje o besedilu in išče predvsem specifične informacije, ki so običajno odgovor na vprašanje oz. poizvedbo. Uporabnika tako zanima specifična informacija v obliki poizvedbe. Sistem povzemanja povzame to informacijo iz besedila in jo nato predstavi v obliki povzetka [10].

(II.) Funkcija: glede na funkcijo povzetka ločimo indikativne in informativne povzetke. Indikativni sistemi povzemanja uporabniku predstavijo izključno glavno idejo besedila. Ti povzetki običajno niso daljši od 10% dolžine originalnega besedila. Osnovni namen indikativnih povzetkov je pridobiti zanimanje bralca za branje originalnega besedila. Po drugi strani informativni

sistemi povzemanja podajajo jedrnato informacijo o glavnem besedilu. Ti povzetki predstavljajo zamenjavo originalnega besedila. Dolžina teh povzetkov ne presega 30% dolžine originalnega besedila. Primer sistema, ki generira indikativno informativne povzetke, je npr. SumUM [10].

(III.) Vrsta: povzemanje besedila na način, kjer sestavek vsebuje stavke izključno iz obravnavanega besedila, je enostavnejše. Tak sestavek imenujemo izvleček. Ključne segmente besedila izberemo na podlagi statistične analize individualnih ali kombiniranih osnovnih značilk, kot so npr. frekvenca besed oz. fraz, lokacije besed in povezovalnih izrazov (angl. *cue words*), ki jih najdemo znotraj besedila [11]. Pri ekstraktni analizi je pomen vsebine omejen na značilke, kot sta frekvenca in lokacija besed. Taka obravnava besedila ne zahteva posebej globokega razumevanja samega besedila. Nasprotno pa t. i. povzetki vsebujejo preoblikovane stavke izvornega besedila. Z drugimi besedami poizkušajo izraziti razumevanje glavnega koncepta izvornega dokumenta. Pri tem uporabljamo jezikovne metode za raziskovanje in interpretiranje besedila s ciljem poiskati nove koncepte in izraze, ki besedilo najbolje opisujejo, z ustvarjanjem krajšega besedila, ki nosi najpomembnejše informacije izvornega besedila.

(IV.) Dolžina: dolžina ciljnega povzetka odločilno vpliva na informativni prispevek. Dolžina povzetka se določa v razmerju glede na dolžino izvornega dokumenta. To imenujemo kompresijsko razmerje med dolžino povzetka in dolžino izvornega besedila.

POGLAVJE 3

3. Računalniško povzemanje besedil

Zgodovina računalniškega povzemanja je stara 60 let. Področje avtomatskega povzemanja besedil se je začelo razvijati z raziskavo Hansa Petra Luhna, ki je na IBM 704 podatkovnem procesorju analiziral frekvenco pojavljanja besed, odstranjeval prazne besede (angl. *stop words*) in gručil podobne besede [27]. Publikacije citirajo Luhnov algoritem iz leta 1958 kot prvi algoritem, ki računalniško povzema naravne jezike na način preprostega pristopa (angl. *shallow/surface approach*) [1]. Preprosti pristop je pristop, kjer se uporablja preprosto statistično hevristiko za selekcijo najbolj pomembnih stavkov izvornega besedila. Luhn je predlagal kot kriterij izbire pomembnih stavkov frekvenco pojavitev besed. Tak način povzemanja je računsko in časovno nezahteven. Povzemanje, ki ne zahteva globljega poznavanja besedila in njegovega pomena, imenujemo ekstraktni način povzemanja [12]. Rezultat ekstraktnega povzemanja besedil je izvleček. Šestindvajset let kasneje je Kupiec z dovoljenjem Edmunsona razširil njegov pristop z naivnim Bayesovim klasifikatorjem. Conroy in O'Leary sta leta 2001 z novim klasifikatorjem *Hidden Markov Model* (HMM) dodala nov prispevek k statističnim pristopom. Model HMM je matematični model oz. ogrodje za predstavitev verjetnostnih porazdelitev čez zaporedje opazovanj [13]. Leta 2001 sta Nomoto in Matsumoto predlagala gručenje na podlagi k-najbližjih sosedov, ki je služilo kot orodje za raziskovanje raznolikosti besedila [14]. Leta 2009 je Binwahlan predlagal tehniko PSO za raziskovanje značilk besedila. Predlagana metoda PSO generira 43% podobnost ročno generiranim povzetkom v primerjavi s povzetki MS Word, ki imajo 37% podobnost v primerjavi z ročno generiranimi [15].

3.1 Ekstraktni pristop povzemanja besedil

Ekstraktivno povzemanje besedila kot strategija izbire in združevanja izbranih stavkov iz prvotnega besedila v izvleček ne omogoča svobode pri izdelavi izvlečka na podlagi razumevanja besedila, ker smo omejeni z vnaprej opredeljenimi stavki. Uteži pomembnosti računamo na podlagi statistične analize individualnih ali kombiniranih osnovnih značilk. Tak pristop ne zahteva razumevanje besedila in je enostaven za implementacijo. Opisanih je nekaj pristopov ekstraktnega povzemanja.

3.1.1 Preprosti pristop ekstraktnega povzemanja besedil

Luhnov algoritem iz leta 1958 predlaga uteževanje stavkov na podlagi statistike, ki beleži frekvenco pojavljanja besed (angl. *word frequency method*). Frekvenca besed služi kot mera za določanje pomembnosti besed. Luhn je sledil ideji, da avtor ponavlja besede medtem, ko razvija določeno temo besedila. Pomanjkljivost Luhnove metode je, da ne upošteva možnosti precenjevanja pomembnosti, ker temelji izključno na frekvenci pojavitve. Pomanjkljivost Luhnove metode je odpravil 11 let kasneje Edmundson [2]. Edmundson je pomembnost določil na podlagi lokacije stavka znotraj besedila, besed iz naslova povezovalnih besed (angl. *cue-words*) in izrazov, ki povezujejo dele stavkov in hkrati namigujejo na semantične relacije v besedilu [3, 4]:

- Povezovalna metoda (ang. *cue phrase method*): temelji na predpostavki, da pomembnost stavka opredelimo glede na prisotnost ali odsotnost določenih povezovalnih besed ali pragmatičnih izrazov (npr. angl. *incidentally, for example, anyway, by the way, furthemore, first, second, then, now, thus, moreover, hence, lastly, finally, in summary, and on the other hand*). Povezovalne besede ali izrazi

povezujejo rdečo nit ali idejo med posameznimi deli besedila in označujejo semantične relacije znotraj besedila. Primer fraze, ki nakazuje pozitivno relevantnost, je npr. ang. *the conclusion of this paper*, primer negativne pa je npr. ang. *for example*. Edmundson pri povezovalni metodi ni uporabljal izrazov, temveč izključno posamezne besede. Besede je razdelil v tri razrede: bonus-besede, nične besede in stigma-besede. Ko stavek vsebuje stigma-besedo, dobi negativno utežitev. Pozitivno utež dobi stavek, ki ne vsebuje stigma-besed in vsebuje eno ali več bonus-besed. Nične besede ne vplivajo na uteževanje. Teufel in Moens sta leta 1997 metodo razširila z uporabo izrazov namesto besed. Tako je metoda postala bolj fleksibilna. Podobno se je povečalo število razredov iz tri na pet. Ideja je bila, da se bolj utežuje stavke, ki vsebujejo tiste fraze, ki pripadajo bolj pomembnemu seznamu oz. razredu [11].

- Pozicijska metoda (ang. *location method*): ideja temelji na predpostavki, da je tendenca pomembnih vsebinskih stavkov na začetku ali na koncu besedila oz. odstavka. Edmundson je predlagal pozitivno uteževanje za vse tiste stavke, ki so se pojavili v prvem ali zadnjem odstavku. Podobno je menil, da je smiselno pozitivno utežiti tudi stavke pod določenimi naslovi, kot je npr. zaključek [11].

- Naslovna metoda: utež stavka se izračuna kot vsota vseh besed, ki se pojavljajo v naslovih ali podnaslovih besedila [11].

3.1.2 Pristop ekstraktnega povzemanja besedil na podlagi korpusa besedil

Vhodni dokumenti vsebujejo tudi primere pogostih besed (angl. *common words*), ki pa za konkretni dokument niso pomembni. V tem pogledu jim je potrebno zmanjšati pomen. V akademskem prispevku so predstavili, da je pomembnost besede obratno sorazmerna številu dokumentov v korpusu, ki vsebujejo besedo [5]. Pomembnost besede v korpusu besedil se računa po formuli: $tf_i * idf_i$, kjer je tf_i frekvenca besede i v dokumentu in idf_i inverzna frekvenca dokumenta [6]. Mera tf-idf se proporcionalno povečuje s številom pojavitve besede v dokumentu in zmanjšuje s frekvenco pojavitve v celotnem korpusu, kar pomaga prilagoditi pomen besede glede na dejstvo, da se določene besede pojavljajo bolj pogosto, kot pa je dejansko njihov vpliv na pomen [23].

3.1.3 Pristop ekstraktnega povzemanja besedil na podlagi grafa

Algoritmi na osnovi grafov se uporabljajo predvsem v raziskovanju struktur spleta in analizi citatov ter socialnih omrežij. Algoritem na osnovi grafa izračuna pomembnost vozlišča rekurzivno, upoštevaje celoten graf. Graf se skreira z dodajanje novih vozlišč za vsak stavek. Povezave med vozlišči so vzpostavljene na podlagi relacij med stavki. Povezave so vzpostavljene s pomočjo relacije podobnosti. Podobnost je definirana na podlagi deleža pokritosti stavkov med seboj. Pokritost je definirana na podlagi števila enakih besed, ki se pojavljajo v sosednjih stavkih. V fazi izdelave povzetka se požene algoritem za rangiranje pomembnih stavkov. Algoritem razvrsti stavke v obratnem vrstnem redu, kot imajo le-ti dodeljene vrednosti uteži. Najvišje uvrščene stavke nato algoritem vključi v povzetek [24, 25, 26].

3.1.4 Pristop ekstraktnega povzemaja besedil z matrično faktorizacijo

Analiza z arhetipi je metoda nenadzorovanega strojnega učenja s pomočjo matrične faktorizacije. Pri splošnem povzemanju besedil predstavljenih z grafom bodo vrednosti

pozitivno in negativno najbolj pomembnih stavkov na robu množice podatkov. Za izračun teh ekstremnih vrednosti se uporablja metodologija analize arhetipov.

Stavki so modelirani kot mešanice arhetipnih stavkov. Pri izbiri arhetipov se omejimo na konveksne kombinacije izvirnih stavkov. Narava analize arhetipov zagotavlja pestre izvlečke, ker analiza arhetipov privzeto izbira različne arhetipne stavke [15].

3.1.5 Ekstraktivno povzemanje besedil z analizo arhetipov

Analizo arhetipov (angl. *archetypal analysis*) sta prva predlagala Cutler in Breiman leta 1994 v okviru ocenjevanja konveksne ovojnice množice podatkov [38]. Analiza arhetipov predstavlja opazovanje multivariantne množice podatkov kot konveksno kombinacijo nekaj ekstremno odstopajočih vrednosti (angl. *outliers*), ki ležijo na sami meji konveksne ovojnice (angl. *convex hull*). Ekstremno odstopajoče vrednosti imenujmo arhetipi (angl. *archetypes*). Primeri, ki se razlikujejo od večine, imajo velik vpliv na rešitev [43]. Bolj kot je oddaljen primer, bolj vpliva na končen rezultat. En sam tak ekstremno odstopajoči primer lahko drastično vpliva na rešitev oz. jo lahko v celoti spremeni [39]. Uporabnost analize arhetipov je na različnih področjih, kot so npr. ekonomija, astrofizika in prepoznavanje vzorcev [40-42].

Imamo *n x m* matriko *X*, ki predstavlja multivariantno množico podatkov z *n* primeri in *m* atributi [39]. Za dani *k* je potrebno v primeru problema arhetipov poiskati matriko *Z*, ki ima *k* *m*-dimenzionalnih arhetipov [39]. Imamo optimizacijski problem, kjer je potrebno poiskati dve matriki koeficientov α in β dimenzij *n x k*, ki minimizirata rezidualno vsoto kvadratov [39]:

$RSS = \|X - \alpha Z^T\|_2$,

kjer je $Z = X^T\beta$ predmet naslednjih dveh omejitev:

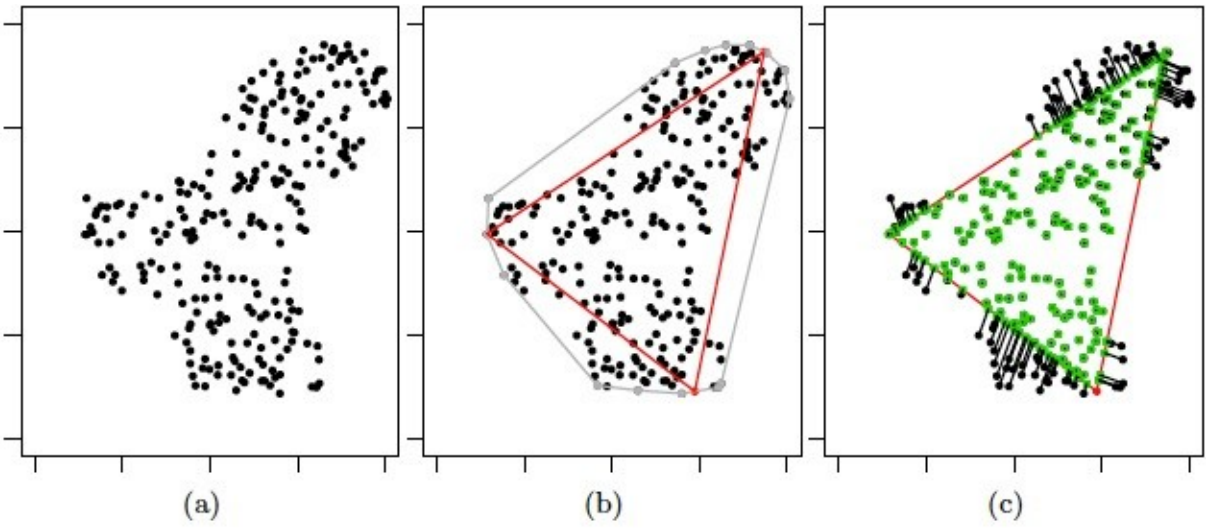
$\sum_{j=1}^{k} \alpha_{ij} = 1$, kjer je $\alpha_{ij} \geq 0$ in i=1, ..., n

in

$\sum_{i=1}^{n} \beta_{ji} = 1$, kjer je $\beta_{ji} \geq 0$ in j=1, ..., k

Omejitve pomenijo, da so aproksimirani podatki konveksne kombinacije arhetipov, tj. $X = \alpha Z^T$ in da so arhitipi konveksna kombinacija podatkov, tj. $Z = X^T \beta$. $\|\cdot\|_2$ označuje normo oz. velikost evklidske matrike."

Na **Sliki 1** so prikazani primeri arhetipov.



Slika 1: Primer analize arhetipov: 1(a): na sliki imamo 250 primerov in dva atributa x in y. Optimalna rešitev za dani primer je k=3 arhetipov; 1(b): slika kaže arhetipe, aproksimacijo njihove konveksne ovojnice (notranji trikotnik) in konveksno ovojnico podatkov (zunanji poligon); 1(c): slika kaže aproksimacijo podatkov skozi arhetipe in korespondenčne vrednosti α. Vsi podatki izven aproksimirane konveksne ovojnice so preslikano na njen rob.

3.1.6 Pristop ekstraktnega povzemanja besedil z algebraično redukcijo

Latentna semantična analiza (angl. *Latent Semantic Analysis*; LSA) je metoda, ki deluje na podlagi statističnih izračunov, ki pridobijo informacije o kontekstnem pomenu besed in

podobnosti stavkov. Je nenadzorovana metoda, ki izračuna vektorski prostor semantične predstavitve besedila [21].

LSA upošteva kontekst vhodnega dokumenta tako, da upošteva informacije, povezane s sosednostjo besed:

- katere besede se uporabljajo skupaj, in katere pogoste besede (angl. *common words*) se pojavljajo v različnih stavkih;
- če je število pogostih besed (angl. *common words*) med stavki visoko pomeni, da so stavki bolj semantično povezani.

LSA uporablja matematično tehniko imenovano dekompozicija singularne vrednosti (angl. *Singular Value Decomposition*, SVD).

SVD je matrična dekompozicija, ki:

- identificira vzorce v relacijah med besedami in stavki v nestrukturiranem besedilu;
- določa podobnost pomena besed in stavkov.

LSA vsebuje tri glavne korake:

- Kreiranje vhodne matrike: vhodni dokument je predstavljen kot matrika. Vsaka vrstica predstavlja besedo in vsaka kolono predstavlja stavek. Vrednost celice predstavlja pomembnost besede. Obstaja več načinov opredeljevanja vrednosti celic. Običajno se upošteva frekvenca pojavitev besed v stavkih;
- Dekompozicija singularne vrednosti (SVD): SVD je matematična metoda, ki se izvede nad vhodno matriko. SVD se uporablja za prepoznavo vzorcev v relacijah med besedami in stavki. SVD lahko kot matematično enačbo predstavimo kot mxn matriko (M). M je oblike :

$$M = U \sum V^T,$$

kjer je U matrika dimenzij m x n in predstavlja izvorne vrstice kot vektorje izbranih vrednosti. $\sum$ je pravokotna diagonalna matrika dimenzij n x n s pozitivnimi realnimi vrednostmi, ki predstavljajo skalarne vrednosti. V^T (konjugirano transponirana matrika V) je enotna matrika dimenzij n x n z realnimi ali kompleksnimi vrednostmi. V^T predstavlja izvorne kolone kot vektorje izbranih vrednosti.

- Izbira stavkov: po izvedbi SVD aktivnosti je rezultat uporabljen pri izbiri stavkov za potrebe generiranja povzetka. Obstaja več metod in algoritmov za izbiro stavkov.

LSA ima številne lastnosti, ki jih lahko uporabimo na široki paleti problemov. Prednosti so:

- LSA je algoritem, ki prepozna globalne trende in vzorce nad celotnim besedilom in besedami;
- LSA ima lastnost pridobivanja dokumentov na podlagi besed in obratno. LSA se uporablja pri preslikavi dokumentov in besed v enak konceptni prostor;
- konceptni prostor vsebuje manj dimenzij tam, kjer je največ informacij in manj šuma.

LSA ima določene omejitve, ki jih je potrebno upoštevati. Slabosti so:

- LSA ima težave s polisemijo (angl. *polysemy*). Polisemija pomeni, da imajo besede več pomenov glede na kontekst. Z drugimi besedami, ista beseda z različnimi pomeni ima isti koncept in bo LSA povzročala velike probleme;
- LSA je odvisen od SVD. SVD je časovno zahteven algoritem.

3.1.7 Algoritmi izbiranja stavkov v LSA

(V.) pristop Gong in Liu (2001):

Koraki Gong in Liu pristopa so predstavitev vhodnega dokumenta kot matrike in izračun SVD. SVD kreira V^T matriko. Vrstni red vrstic v kreirani V^T matriki predstavlja pomembnost koncepta. Vrednosti celic predstavljajo povezanost med stavkov in konceptom. Stavki, ki imajo močnejšo povezanost s konceptom imajo višjo vrednost celice. Slabosti pristopa so :

- če je veliko število stavkov kandidatov za povzetek, potem se bodo v povzetku pojavili taki, ki niso pomembni, da bi bili vključeni;
- v primeru, da ima koncept opisuje več pomembnih stavkov, potem se bo v povzetku pojavil samo eden izmed njih;
- nad vsemi koncepti je upoštevan enak nivo pomembnosti, ki se upošteva pri izbiri stavkov.

(VI.) pristop Steinberger in Jezek (2004):

Njun pristop se imenuje strategija dolžine (angl. *Lengthy Method*). Pristop si deli enaka prva dva koraka kot pri Gon in Liu pristopu. Razlikuje se v zadnjem koraku, kjer se izvede izbira stavkov. Dolžina vektorja, ki predstavlja stavke, se uporabi kot kriterij za izbiro stavka. Dolžina se izračuna kot:

$$Dolžina = \sqrt{\sum_{j=1}^{n} V_{ij} * \sum_{jj}} ,$$

dimenzija n novega prostora je podan kot parameter. Če so indeksi konceptov manjši ali enaki dani dimenziji, so ti koncepti uporabljeni pri izračunu dolžine. Za pridobitev najbolj pomembnih konceptov, se uporabi $\sum$ matrika kot parameter za množenje. Stavki z največjo dolžino so izbrani med kandidati za povzetek. Njun pristop se izogiba omejitvam Gong in Liu pristopa. Izboljšavi

sta naslednji: vsi izbrani stavki so v relaciji vsem pomembnim konceptom in iz vsakega koncepta je izbran več kot en stavek.

(VII.) pristop Murray, Renals in Carletta (2005):

Pristop Murray, Renals in Carletta želi omejiti slabosti Gon in Liu pristopa. Njihov pristop se začne izvajati v koraku izbire stavkov za povzetek. Pristop ima dve glavni funkciji, ki imata naslednji lastnosti: prvi je, da izbere več kot en stavek iz najbolj obetavnih konceptov in druga lastnost se nanaša na uporabo $\sum$ matrike pri določanju koliko stavkov se bo vzelo pri izbranem konceptu. Število stavkov se izračuna z izračunom deleža singularnih vrednosti, ki so v relaciji z obravnavanimi stavki, čez vsoto vseh singularnih vrednosti za vsak koncept posebej.

(VIII.) pristop Ozsoy (2010):

Ozsoy pristop je najbolj uporabljen pristop v trenutnih študijah LSA. Njegov pristop predlaga dve metodi za izbiro stavkov. To sta križna metoda in metoda teme.

a. Križna metoda:

Križna metoda (angl. *Cross Method*) je podaljšek Steinberger in Jezek pristopa. Križna metoda ima dodan korak predporocesiranja med SVD izračunom in izbiro stavkov. Za izbiro stavkov se uporabi VT matrika. Faza predprocesiranja odstrani stavke, ki niso pomembni za koncept. Za vsak koncept se izračuna povprečna vrednost. Vrednost celice stavka se postavi na nič, če je njena vrednost manjša ali enaka povprečni vrednosti.

Po fazi predprocesiranja se izračuna dolžina vektorja za vsak stavek. Stavki najdaljše dolžine so vključeni v povzetek.

b. Metoda teme

Metoda teme (angl. *Topic Method*) je podobna križni metodi. Deluje na osnovi prepoznave glavnega koncepta in podkoncepta. Iz rezultata SVD je izbran koncept , ki se imenuje tema vhodnega dokumenta. Faza predprocesiranja je dodana med izračunom SVD in korakom izbire stavkov. Po fazi predprocesiranja se kreira koncept X konceptne matrike, da se poišče glavna tema. Izbran je koncept, ki ima skupne stavke. Za te primere je določena nova vrednost celice matrike. Vrednost prispevka koncepta ustreza seštevku doprinosa uteži vseh stavkov, ki pripadajo istemu konceptu. Na enak način se izračuna vrednost uteži za vsak koncept posebej. Koncept z najvišjo vrednostjo uteži je izbran za glavno temo besedila. Na tak način dobimo seznam najbolj obetavnih konceptov in stavkov, ki jim pripadajo. Na koncu se po načinu Gong in Liu izbere po en stavek iz vsakega koncepta [48].

V **Tabeli 1** so zbrane vse opisane metode in njihove značilnosti.

Algoritem z LSA pristopom	Ime algoritma	Vhod	Izbrani stavki	Glavni koraki	Lastnosti izbire stavkov	Izhod
Gong in Liu pristop (2001)	Gong in Liu metoda	Posamezen dokument	En stavek / pomemben koncept	1. Kreiranje vhodne matrike 2. Izračun SVD 3. Izbira stavka	Glede na: 1. matriko V^T	Izvleček
Steinberger in Jezek pristop (2004)	Metoda dolžine	Posamezen dokument	Več kot en stavek / pomemben koncept	1. Kreiranje vhodne matrike 2. Izračun SVD 3. Izbira stavka	Glede na: 1. matriko V^T 2. dolžino vektorja stavka	Izvleček
Murray, Renals in Carletta pristop (2005)	Murray, Renals in Carletta metoda	Posamezen dokument	Več kot en stavek / pomemben koncept	1. Kreiranje vhodne matrike 2. Izračun SVD 3. Izbira stavka	Glede na: 1. matriko V^T 2. dolžino vektorja stavka	Izvleček
Ozsoy pristop (2010)	Križna metoda	Posamezen dokument	Več kot en stavek / pomemben koncept	1. Kreiranje vhodne matrike 2. Predprocesiranje 3. Izračun SVD 4. Izbira stavka	Glede na: 1. matriko V^T 2. povprečno vrednostjo vsakega stavka 3. skupna dolžina vsakega vektorja stavka	Izvleček
Ozsoy pristop (2010)	Metoda teme	Posamezen dokument	Več kot en stavek / pomemben koncept	1. Kreiranje vhodne matrike 2. Predprocesiranje 3. Izračun SVD 4. Izbira stavka	Glede na: 1. matriko V^T 2. kreiranje koncepta X konceptne matrike 3. vrednost uteži za vsak koncept 4. Iskanje glavnega koncpta in podkoncpetov	Izvleček

Tabela 1: Primerjava algoritmov, ki delujejo v okviru algoritma povzemanja besedil LSA.

3.2 Semantični pristop povzemanja besedil

V človeškem povzemanju so prepoznane naslednje aktivnosti:

- redukcija stavkov;
- kombiniranje stavkov;
- sintaktično transformiranje;

- leksikalno parafraziranje;
- posploševanje in specializacija;
- spreminjanje vrstnega reda.

Povzetki, ki bi upoštevali te korake, bi bili bolj podobni človeškim povzetkom.

Danes je med avtomatsko generiranimi povzetki in povzetki strokovnjakov še vedno velika razlika. Sistem ne uspe vedno zaznati pomembnih tem. Če izbrani pomembni stavki povzetka niso povezani oz. v kakršnikoli relaciji v izvornem besedilu, je rezultat nekoherenten povzetek. Hovy in Lin sta leta 1997 prva predlagala ne-ekstraktni pristop oz. semantični pristop k reševanju problema avtomatskega povzemanja besedil. Predstavila sta ga kot trojček aktivnosti: identifikacija, interpretacija in generiranje z uporabo metod pozicije stavkov, povezovalnih izrazov in identifikacije teme [11]. Kasnejši avtorji so predlagali tehniko kompresije stavkov, kreiranja novih stavkov od začetka in strategije kopiraj in prilepi (angl. *cut-and-paste strategy*) [28-33].

Opisanih je nekaj pristopov semantičnega povzemanja besedila.

3.2.1 Pristop semantičnega povzemanja besedil s koherenco

Princip koherence je prvi predlagal Marcu leta 1997. Koherenca je odraz hierarhične organiziranosti. Za vsak del koherenčnega besedila obstaja nek razlog, da ta del besedila obstaja in ni smiselno, da bi kak del lahko izpustili. Če se kohezija uporablja za segmentacijo besedila na nivoju odstavkov, se koherenco uporablja za segmentacijo besedila na nižjem nivoju, npr. pomembni stavki. Omenjeni pristop je nadaljeval Mani leta 1998. Teorijo koherence uporabljamo kot analitično orodje za analizo strukture besedila. Najbolj razširjen primer te teorije je teorija jezikovne strukture in (ang. *Rhetorical Structure Theory*; RST), ki obravnava strukture besedila s pomočjo relacij/odnosov med posameznimi deli besedila. RST

sta prva predstavila Mann in Thompson leta 1988 [16]. Z delom je nadaljeval Marcu leta 1997, ko je formaliziral teorijo dvoumnosti in vpeljal pristope RST v povzemanje besedil. Teufel in Moens sta leta 2002 uporabila RST pri raziskovanju povzemanja znanstvenih člankov.

3.2.2 Semantično povzemanje besedil s teorijo jezikovne strukture

Ekipa raziskovalcev (Bill Mann, Sandy Thompson in Christian Matthiessen) je leta 1983 opozorila, da ne obstaja teorija, ki bi ponujala dovolj podrobnosti za vodenje programske rešitve, ki bi analizirala vlogo posameznih delov besedila glede na celotno besedilo. Pojavila se je teorija jezikovne strukture (angl. *Rhetorical Structure Theory*), ki opisuje glavne vidike organizacije naravnega besedila in služi kot analitično orodje ali orodje za gradnjo besedil. Teorija predpostavlja različne možne strukture, gradnike in relacije med njimi, ki jih preučujemo. Motivacije za razvoj teorije so bile naslednje [18]:

- besedila niso samo nizi stavkov, temveč jih sestavljajo hierarhično organizirani stavki in skupine stavkov med katerimi obstajajo različne relacije;
- relacije lahko opišemo funkcionalno glede na namen pisatelja besedila in pisateljevih predpostavk o bralcu besedila, kar se odraža na organizaciji besedila in predstavitvi konceptov besedila;
- najbolj pogosta relacija v besedilu je atom-satelit, kjer je del besedila pomožen glede na drugega.

Teorija jezikovne strukture pozna tri vrste mehanizmov: besedne strukture. relacije in sheme. Imamo dva področja besedila. Predefinirana relacija definira vzorce pogojev, ki so lahko za dani področji besedila resnični ali neresnični. Če so resnični, velja, da relacija drži [18].

Sheme so predefinirani vzorci, ki določajo kako področja besedila tvorijo večja področja vse do celotnega besedila. Najbolj enostavni in najštevilčnejši vzorci vsebujejo posamezne relacije, ki držijo. Bolj kompleksne sheme vsebujejo pare relacij, ki si delijo skupni del. Ostale sheme opisujejo izjeme pri katerih relacije atom-satelit ne opisujejo dovolj dobro lokalnih struktur [18].

Najbolj pogost strukturni vzorec je, da sta dva dela besedila (ponavadi sosednja, čeprav obstajajo izjeme) v relaciji tako, da ima eden izmed njiju specifično vlogo relativno glede na drugega. Vzorčni primer je npr. primer, kjer trditvi sledi dokaz o tej trditvi. V tem primeru teorija jezikovne strukture predpostavlja relacijo dokaza med dvema deloma besedila. Teorija predpostavlja, da je trditev bolj bistvena za besedilo kot določen dokaz. Ta reakcija je predstavljena na način, da se trditev označi kot atomarni gradnik, dokaz pa kot satelit. Vrstni red relacij ni omejen. Obstaja bolj ali manj verjeten vrstni red relacij. Še nekaj primerov relacij je prikazanih v **Tabeli 2** [17].

Ime relacije	Atomarni gradnik	Satelit
Ozadje	Besedilo, katerega razumevanje je bilo olajšano.	Besedilo, ki olajša razumevanje.
Izdelava	Osnovna informacija.	Dodatna informacija.
Priprava	Besedilo, ki se predstavlja.	Besedilo, ki pripravi bralca, da pričakuje in interpretira besedilo, ki ga predstavlja.

Tabela 2: Primeri relacij RST v besedilu.

Če relacija nima določenega dela besedila, ki bi se nanašal na avtorjev namen, jo imenujemo večatomarna relacija.

Rezultat je prezentacija v obliki drevesa. Deli besedila so izbrani za povzetek. Stavki so uteženi glede na jezikovno vlogo v drevesu [34]. Uteži z vrednostjo 1 so dodeljeni satelitom

in uteži z vrednostjo 0 so dodeljeni atomom. Končna vrednost uteži stavka je določena kot vsota uteži od korena drevesa do stavka.

3.2.3 Pristop semantičnega povzemanja besedil s kohezijo

Iskanje povezanosti oz. princip kohezije besedila vključuje relacije med besedami oz. izrazi, ki so povezani s temi besedami. Te relacije določajo, kako močno so te besede ali izrazi, ki so v relaciji, povezani med seboj. Primer take relacije je anafora (angl. *anaphoric expression*). Anafore, ki se nanašajo na dogodke ali entitete za nazaj lahko povzamemo samo, če upoštevamo celoten kontekst.

Aproksimacijski algoritmi rešujejo problem kohezije. Eden takih algoritmov je leksikalno veriženje [36]. Morris in Hirst sta leta 1991 prva predstavila princip leksikalnega veriženja. Leksikalno veriženje je seznam povezanih izrazov, ki opisujejo določen predmet oz. temo. Implementacija tega pristopa v tistem času ni bila izvedljiva zaradi odsotnosti digitalnih virov. Omenjeni princip sta leta 1997 izboljšala Barzilay in Elhadad in leta 2002 Silbe in McKoy.

Leksikalno veriženje besedila

Tehnika leksikalnega veriženja besedila izdela prezentacijo besedila v obliki sosednjih struktur tako, da si pomaga s kohezijo med besedami, ki so med seboj povezane. Leksikalno veriženje ustvarjamo s pomočjo grupiranja oz. veriženja množice besed, med katerimi veljajo semantične relacije (npr. ponavljanje, sinonimi, antonimi, hipernimi, holonimi, ipd.) in elektronskim slovarjem kot je npr. WordNet. Verige sestavljajo besede, ki so v opisanih

relacijah. Uteži so določene na podlagi števila in tipa relacij v verigi. Stavki pri katerih je prisotna koncentracija najmočnejših verig so izbrani za povzetek [6].

Možne semantične relacije so:

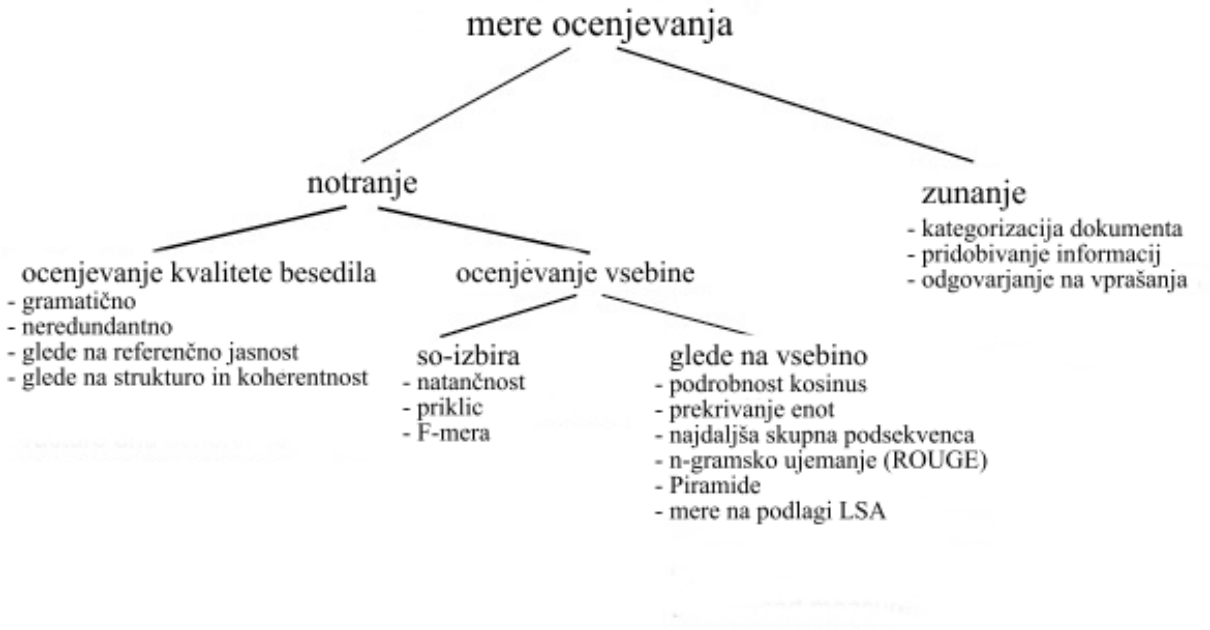
- Identični termini: termini, ki imajo unikatno izgovorjavo, črkovanje in pomen;
- Homofoni: termini, ki imajo isto izgovorjavo in različen pomen ter črkovanje (npr. v ang. poznamo besedo rose, ki lahko pomeni roža ali pa preteklik od glagola rise). Poznamo različne tipe homofonov:
    - Homografi: homofoni, ki so si podobni v črkovanju toda različni v pomenu (npr. ang. *hail*, ki lahko pomeni toča ali pa nekaj kar se zgodi pogosto);
    - Homonimi: homofoni, ki imajo enako izgovorjavo toda različen pomen (npr. ang. *sight/site*);
    - Heterografi: homofoni, ki imajo različno črkovanje, vendar enako izgovorjavo (npr. ang. *write/right*);
    - Oronimi: homofoni, ki jih sestavlja več besed ali fraz in podobno zvenijo (npr. ang. *ice scream* ali *I scream*);
    - Pseudo-homofoni: pseudo besede, ki so fonetično identične, pri čemer pa eden izmed parov besed ni prava beseda (npr. ang. *groan* in termin, ki ni beseda *grone*).
- Sinonimi: termin z enakim ali podobnim pomenom kot druga beseda ali fraza (npr. vesel in zadovoljen). Sinonimi imajo lastnost simetrije in refleksije.
- Antonimi: termin, ki ima nasprotni pomen kot druga beseda (npr. bogat in reven). Antonimi imajo lastnost simetrije;
- Hiponimi: termin, ki označuje specifičnost skupine, označene z besedo ali frazo (npr. pes je hiponim živali). Hiponimi imajo lastnost tranzitivnosti;

- Hipernimi: termin, ki opisuje generičnost hiponimov (npr. glasbeni inštrument je hipernim kitari). Hipernimi imajo lastnost tranzitivnosti;
- Meronimi: termin, ki je del nečesa večjega (npr. kolesa so meronim avtomobilu). Meronimi imajo lastnost tranzitivnosti;
- Holonimi: termin, ki opisuje generičnost meronimov (npr. beseda je holonim črki). Imajo lastnost tranzitivnosti;
- Troponimi: termini, ki opisujejo enega od načinov, kako doseči njihov troponim (npr. utopiti in zastrupiti sta troponima ubiti),
- Koordinacijski termini: vsak pripadajoči primer si deli hipernim s koordinacijskim terminom (npr. moški in ženska sta koordinacijska termina, ker si delita hipernim človek. Podobno velja za mačko in psa, ki si lahko delita celo več hipernimov: žival, sesalec, hišna žival, ipd.).

3.3 Mere za ocenjevanje metod povzemanja besedil

Opisane metode ocenjevanja kvalitete besedila so povzete po [37]. Glavni primer ocenjevanja kvalitete povzetka je ocenjevanje vsebine (angl. *content evaluation*). Ocenjevanje vsebine se izvaja v primerjavi s človeškim povzetkom. Za ocenjevanje izvlečkov se uporablja mera za ocenjevanje, imenovana so-izbira (ang. *co-selection*). Mera so-izbira vrne toliko stavkov v izvlečku, kolikor jih ustreza človeškemu povzetku. Naslednja je mera imenovana mera vsebine (angl. *content-based measures*). Mera vsebine primerja posamezne besede stavka namesto cele stavke. Prednost mere vsebine je, da lahko primerja tako človeške kot sistemske povzetke s človeškimi, ki vsebujejo nove stavke. Zadnje so metode, ki ocenjujejo povzetke, za točno določene namene. Te metode se imenujejo metode za posebne naloge (angl. *task-based methods*) [6].

Slika 2 prikazuje taksonomijo mer ocenjevanja povzetkov in izvlečkov [6].



Slika 2: Taksonomija mer ocenjevanja povzetkov in izvlečkov.

3.3.1 Kvaliteta besedila

Poznamo nekaj vidikov lingvistične kvalitete besedila [6]:

- Gramatično: gramatično (angl. *grammatically*) pomeni, da naj besedilo naj ne bi vsebovalo napačnih besednih oblik ali drugih gramatičnih napak, kot so npr. napačna ločila, ipd.;
- Neredundantnost: neredundantnost (angl. *non-redundancy*) pomeni, da besedilo ne bi smelo vsebovati redundantne informacije;
- Referenčna jasnost: referenčna jasnost (angl. *referential clarity*) pomeni, da morajo biti samostalniki in zaimki morajo biti jasno opredeljeni z entitetami v besedilu. To pomeni, da mora imeti beseda *on* jasno referenco v besedilu na določeno osebo;
- Koherenca in struktura: koherenca in struktura pomenita, da morajo biti stavki koherentni (vsebinsko ne smejo zanikati sami sebe oz. biti sebi v nasprotju) in imeti dobro strukturo.

3.3.2 Ocenjevanje kvalitete besedila z mero so-izbire

Glavne metrike ocenjevanja pri meri so-izbire (angl. *co-selection measures*) so natančnost (angl. *precision*), priklic (angl. *recall*) in F-mera (angl. *F-score*) [6]. Pri meri so-izbire pomeni metrika *natančnost* (P) število stavkov, ki se pojavljajo v idealnem povzetku (tj. človeškem povzetku strokovnjaka) in v sistemskem povzetku, deljeno s številom vseh stavkov v sistemskem povzetku. Pri tej meri pomeni priklic (R) delež stavkov, ki se pojavijo v obeh sistemih proti vsem stavkom v človeškem povzetku. F-mera (ang. *F-score*) je kompozitna mera, ki združuje meri natančnost in priklic. F-mera se izračuna kot harmonično povprečje natančnosti in priklica [6]:

$F = \frac{2 * P * R}{P + R}$."

3.3.3 Ocenjevanje kvalitete besedila z mero so-izbire glede na vsebino

Mera so-izbire upošteva stavke, za katere velja natančno ujemanje [6]. Slednje ne upošteva možnosti, da dva stavka vsebujeta enako informacijo, čeprav sta napisana drugače. Čeprav mere so-izbire ne uspejo oceniti takih primerov, pa jih mere glede na vsebino (angl. *content-based measures*) pa jih lahko. Poznamo naslednje mere glede na vsebino:

- podobnost kosinus: osnovni primer mere glede na vsebino je podobnost kosinus (angl. *cosine similarity*):

$$\cos(X, Y) = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i (X_i)^2} \cdot \sqrt{\sum_i (y_i)^2}},$$

kjer sta X in Y reprezentaciji sistemskega povzetka in njegove reference;

- prekrivanje enote: prekrivanje enote (angl. *unit overlap*) se računa na način:

$$prekrivanje(X, Y) = \frac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|},$$

kjer sta X in Y prezentaciji na osnovi množic besed ali lem. $\|X\|$ je velikost množice X;

- najdaljša skupna podsekvenca: najdaljša skupna podsekvenca (angl. *longest common subsequence*; LCS) se računa na način:

$$LCS(X, Y) = \frac{dolžina(X) + dolžina(Y) - razdalja(X, Y)}{2},$$

kjer sta X in Y prezentaciji glede na sekvence besed ali lem. LCS(X, Y) je dolžina najdaljše skupne podsekvence med X in Y. Dolžina(X) je dolžina niza X. Podobno velja za dolžino (Y). Razdalja (X,Y) je razdalja med X in Y;

- n-gramsko ujemanje (ROUGE): družina ROUGE mer, ki je bila predstavljena leta 2003, deluje glede na podobnost n-gramov [44]. N-gramski model je tip verjetnostnega jezikovnega modela, s katerim predvidevamo, s kakšno verjetnostjo se bo pojavil naslednji primer v taki sekvenci. Primer je npr., da opazujemo jezikovni prostor. V tem prostoru je enota beseda. Potem je 1-gramska sekvenca ravno beseda ali unigram, 2-gramska sekvenca sta dve besedi ali bigram, itd. N-gramski model modelira sekvence naravnega jezika s pomočjo statistične lastnosti n-gramov. Predpostavlja se, da je vsaka beseda odvisna samo od zadnjih n-1 besed. Predpostavka

je pomembna, ker je lastnost jezika, da se besede povezujejo [62]. N-gramsko ujemanje (angl. *N-gram co-occurrence statistics*) izračunamo kot:

$$ROUGE - n = \frac{\sum_{C \in RSS} \sum_{gram_n \in C} Števec_{ujemanj}(gram_n)}{\sum_{C \in RSS} \sum_{gram_n \in C} Števec(gram_n)},$$

kjer je RSS množica referenčnih povzetkov, $Števec_{ujemanj}(gram_n)$ je maksimalno število n-gramov, ki se pojavljajo v sistemskem in referenčnem povzetku in $Števec(gram_n)$ je število n-gramov v referenčnem povzetku. V družini ROUGE obstajajo še druge mere kot so: ROUGE-L (najdaljša skupna podsekvenca), ROUGE-SU4, idr.;

- Piramide: metoda Piramide je nova polavtomatska metoda ocenjevanja besedila [45]. Osnovna ideja je identifikacija vsebinskih enot povzetka (angl. *summarization content units*; SCUs), ki se uporabljajo za primerjavo informacij povzetkov. Vsebinske enote povzetka izhajajo iz anotacij korpusa povzetkov in niso večje od stavka. Anotacija se začne z identifikacijo podobnih stavkov in se nadaljuje s podrobnejšo analizo. Analiza lahko vodi do prepoznave bolj tesno povezanih delov besedila. Na vrhu piramide so vsebinske enote povzetka, ki se pojavljajo v večini povzetkov in tako imajo tako večjo težo. Nižje kot se pojavijo vsebinske enote povzetkov, manjšo imajo težo, ker se pojavijo v manj povzetkih. Vsebinske enote povzetkov v sistemskih povzetkih se primerjajo s tistimi v zgrajeni piramidi. Na ta način se oceni, koliko informacije se dejansko ujema med sistemskimi povzetki (angl. *peer summary*) in ročnimi povzetki (angl. *manual summary*) [45].

3.3.4 Zunanje mere ocenjevanja kvalitete besedila

Zunanje mere ocenjevanja kvalitete besedila imenujemo tudi namenske mere ocenjevanja (angl. *task-based measures*) [6]. Zunanje mere ocenjevanja ne analizirajo stavkov povzetka. Poizkusijo meriti možnost uporabe povzetka za določene naloge. V literaturi lahko najdemo različne mere namenskega ocenjevanja. Predstavljene so naslednje tri metode [6]:

- kategorizacija dokumenta: kvaliteta avtomatskega povzemanja se meri po ustreznosti zamenjave polnega dokumenta za kategorizacijo. Ocenjevanje ugotavlja, ali je povzetek učinkovit pri zajemanju kakršnekoli informacije dokumenta za pravilno kategorizacijo dokumenta (angl. *document categorization*) [6];

- pridobivanje informacij: relevantnost korelacije (angl. *relevance correlation*) je primer pridobivanja informacij (angl. *information retrieval*; IR) [46]. Relevantnost korelacije ocenjuje relativno zmanjševanje pridobivanja informacij, ko se premikamo od izvornega dokumenta proti povzetku [6];

- odgovarjanje na vprašanja: princip delovanja zunanjega ocenjevanja pod nazivom odgovarjanje na vprašanja (angl. *question answering*) je bilo predstavljeno v [47]. Predmet raziskave avtorjev znanstvenega prispevka je bila vaja z vprašanjem in več možnimi odgovori. Med odgovori je bil pravilen natanko en. Avtorji so merili, koliko vprašanj je imelo pravilne odgovore pod različnimi pogoji [6].

POGLAVJE 4

4. Opis knjižnic in uporabljenih orodij

Pri razvoju samostojne spletne aplikacije za povzemanje slovenskih akademskih besedil smo uporabili v tem poglavju opisana orodja, knjižnice in slovarje. Vsa orodja, knjižnice in slovarji so prosto dostopni, razen Matlab orodja, za katerega smo pridobili študentsko licenco. Za uporabo metodologije analize arhetipov na način, kot je opisan v dizertaciji, smo pridobili dovoljenje avtorja Ercana Canhasija.

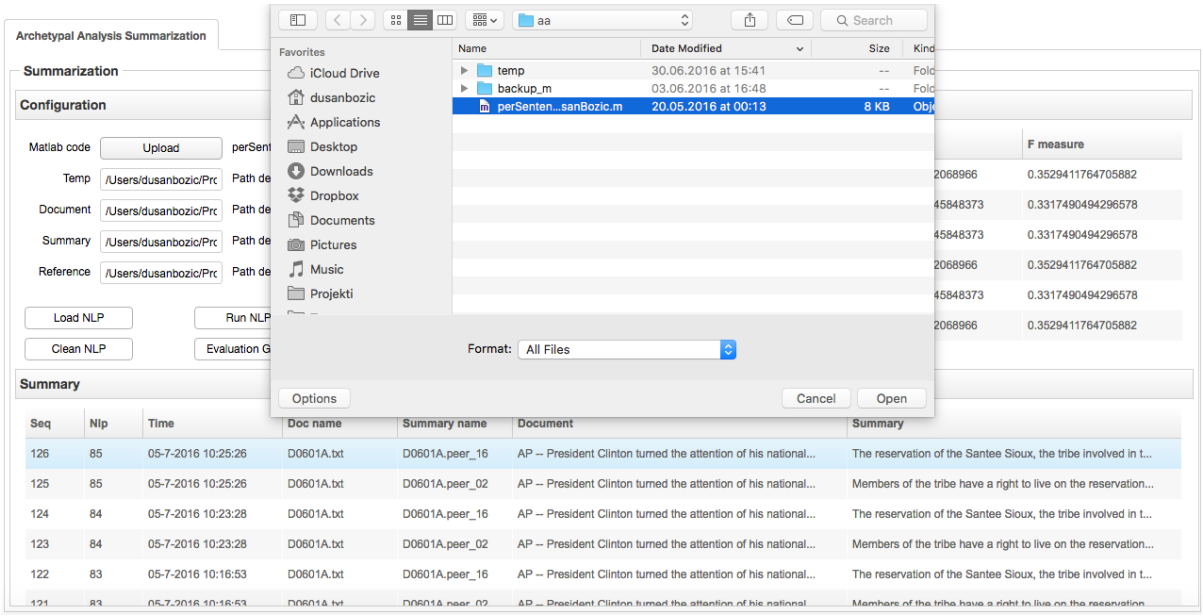
4.1 Java EE

Za implementacijo je bil izbran programski jezik Java. Razlogov za izbiro Jave je bilo več. Programska skupnost, ki uporablja Java programski jezik, je široka in aktivna. Iskanje odgovorov na vprašanja, povezana z implementacijo, je hitro in učinkovito. Knjižnic, napisanih v Javi, je zelo veliko. Izbrana je bila Java EE 6, ker izbrani aplikacijski strežnik WebLogic 12c podpira verzijo Java EE 6.

4.2 ZK

Za implementacijo uporabniškega vmesnika je bilo izbrano ogrodje ZK, ker ponuja poleg licenčne verzije tudi prosto dostopno in odprtokodno različico. Ogrodje omogoča implementacijo spletnih in mobilnih aplikacij. Aplikacije napisane v ZK delujejo kot namizne aplikacije, čeprav so spletne aplikacija. V ozadju se izvajajo Ajax klici med spletno stranjo in aplikacijskih strežnikom, tako, da uporabnik doživlja aplikacijo kot namizno. Sintaksa ZK ogrodja je enostavna in omogoča hitro implementacijo grafičnega vmesnika.

Uvoz algoritma za analizo arhetipov v program Simplex prikazuje **Slika 3**.



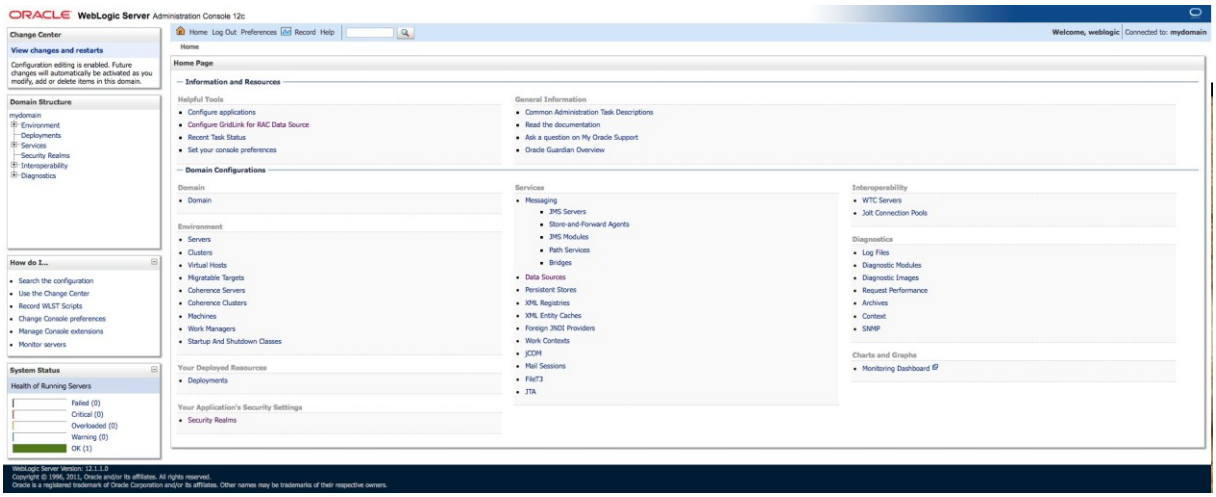
Slika 3: Primer uporabniškega vmesnika aplikacije SimpleX.

4.3 WebLogic strežnik

Aplikacija se izvaja na aplikacijskem strežniku Oracle WebLogic 12c. Aplikacijski strežnik je bil izbran, ker je robusten, zanesljiv in za osebne namene prosto dostopen. Strežnik ima skonfiguriran dostop do baze, ki ga aplikacije uporablja. Zaradi velike porabe pomnilnika smo strežniku dodelili 8GB pomnilnika. Nastavitev je shranjena v skreirani domeni v datoteki setDomainEnv.sh:

```
if [ "${USER_MEM_ARGS}" != "" ] ; then
MEM_ARGS="-Xms2048m -Xmx8192m -XX:PermSize=1024m -XX:MaxPermSize=3072m"
#"${USER_MEM_ARGS}"
export MEM_ARGS
fi
```

WebLogic strežnik omogoča konfiguracijo strežnika preko spletnega vmesnika (**Slika 4).**

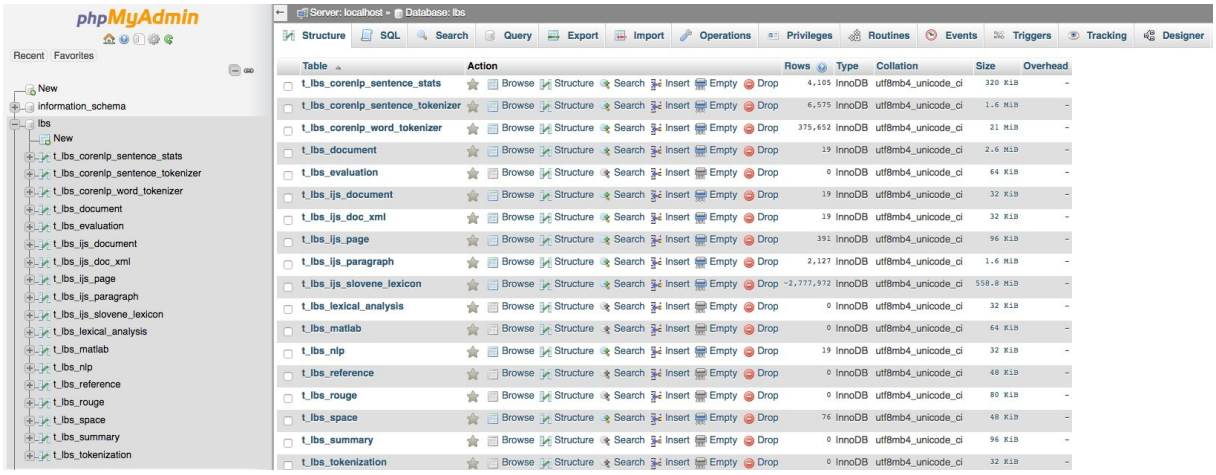


Slika 4: Administratorska konzola aplikacijskega strežnika WebLogic.

4.4 WAMP strežnik

Za shranjevanje podatkov smo postavili in skonfigurirali WAMP strežnik. WAMP strežnik je Apache strežnik, ki podpira MySQL bazo in PHP skriptni jezik. Dostop do MySQL baze je mogoč preko konzole ali brskalnika. Za upravljanje z bazo smo uporabljali spletni vmesnik, ki je dostopen preko spletnega naslova: http://localhost/phpmyadmin/.

Apache strežnik omogoča delo z MySQL bazo preko spletnega vmesnika (**Slika 5).**



Slika 5: Spletni uporabniški vmesnik za administracijo MySQL baze.

4.5 Matlab

Za izvajanje povzemanja besedil na podlagi analize arhetipov smo integrirali Matlab okolje s programom SimpleX preko Matlab interpreterja, ki ga podpira študentska licenca. Dr. Ercan Canhasi je v svoji dizertaciji implementiral metodologijo analize arhetipov v Matlab skriptnem jeziku. Njegovo implementacijo smo vključili v program Simplex za primerjavo ekstraktnega pristopa na podlagi analize arhetipov in semantičnega pristopa na podlagi latentne semantične analize.

4.6 CoreNLP

CoreNLP je ogrodje za procesiranje naravnih jezikov Univerze Stanford. Ogrodje podpira množico analitičnih orodij:

- za določitev besedne oblike, kot je npr. samostalnik, glagol, pridevnik, ipd. (angl. *part-of-speech*);
- za označevanje sekvenc besed, ki predstavljajo entiteto, kot so npr. imena oseb, podjetij, zdravil, ipd. (angl. *named entity recognizer*);
- prevajalnik za gramatično strukturo stavkov (angl. *parser*);
- povzemanje relacij med besedami (angl. *open information extraction*);
- ...

Nekatera orodja je bilo možno uporabiti nad slovenskimi besedili, kot je npr. prepoznava stavkov in besed. Večina orodij pa vsebinsko povezana z angleškim jezikom.

4.7 TML

TML je knjižnica za raziskovanje besedil Univerze v Sydney-ju. Knjižnica vsebuje algoritem za latentno semantično analizo besedil. Knjižnica zahteva integracijo z bazo MySQL, ker pri svojem delovanju uporablja podatkovni model. Ob zagonu algoritma se prebere seznam praznih besed (angl. *stop-words*), ki so ob analizi izločene. Seznam praznih besed smo pripravili na podlagi slovenskega leksikona. Na seznamu so vse besede tipa: členek, zaimek, veznik, števnik, predlog, okrajšave in medmet.

4.8 Log4J

Log4J je Java-knjižnica, ki podpira krožno logiranje izvajanje programa. Podpira ločeno beleženje informacij kot napak. Aplikacija SimpleX uporablja logiranje vseh metod - zaradi lažjega spremljanja izvajanja, kot tudi reševanja napak.

4.9 Analiza Arhetipov

Rezultate povzetkov s pomočjo semantične analize smo primerjali z rezultati ekstraktnih izvlečkov, izdelanih s pomočjo analize arhetipov. Algoritem je povzet po doktorski nalogi dr. Ercana Canhasija z naslovom "*Graph based models for multi-document summarization*" [49].

Izmed vseh različic implementacije analize arhetipov, ki so na voljo v nalogi dr. Ercana Canhasija, smo izbrali tisto, ki je shranjena pod imenom »perSentece2006.m«. Implementacija je napisana v jeziku Matlab. Dodali smo Matlab-logiranje za sprotno obveščanje o izvajanju v sistemski konzoli Matlab. Uporabili smo pomemben parameter Archetype Number (AN) in njegove vrednosti NOC (ang. *number of components*) s tistimi, ki so se v doktorski nalogi izkazale za najbolj obetavne. AN je 2 in pripadajoči vrednosti NOC sta 2 in 16.

4.10 JROUGE

Za ocenjevanje smo uporabili Java implementacijo ROUGE [50]. Bistvene značilnosti knjižnice JROUGE so naslednje:

- JROUGE uporablja samo metriko N-gram (ROUGE-N) v primerjavi z izvornim ROUGE, ki pozna naslednje metrike: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S in ROUGE-SU;
- za prepoznavo numeričnih in alfa numeričnih znakov uporablja unicode regularne izraze;

- JROUGE v primerjavi z izvornim ROUGE ne zaokrožuje rezultatov na petem mestu za decimalko;
- za razdelitev besedila na stavke uporablja orodja Stanford NLP;
- JROUGE ima izpostavljene vmesnike za dostop do struktur, ki jih vrača knjižnica JROUGE.

4.11 Digitalni viri

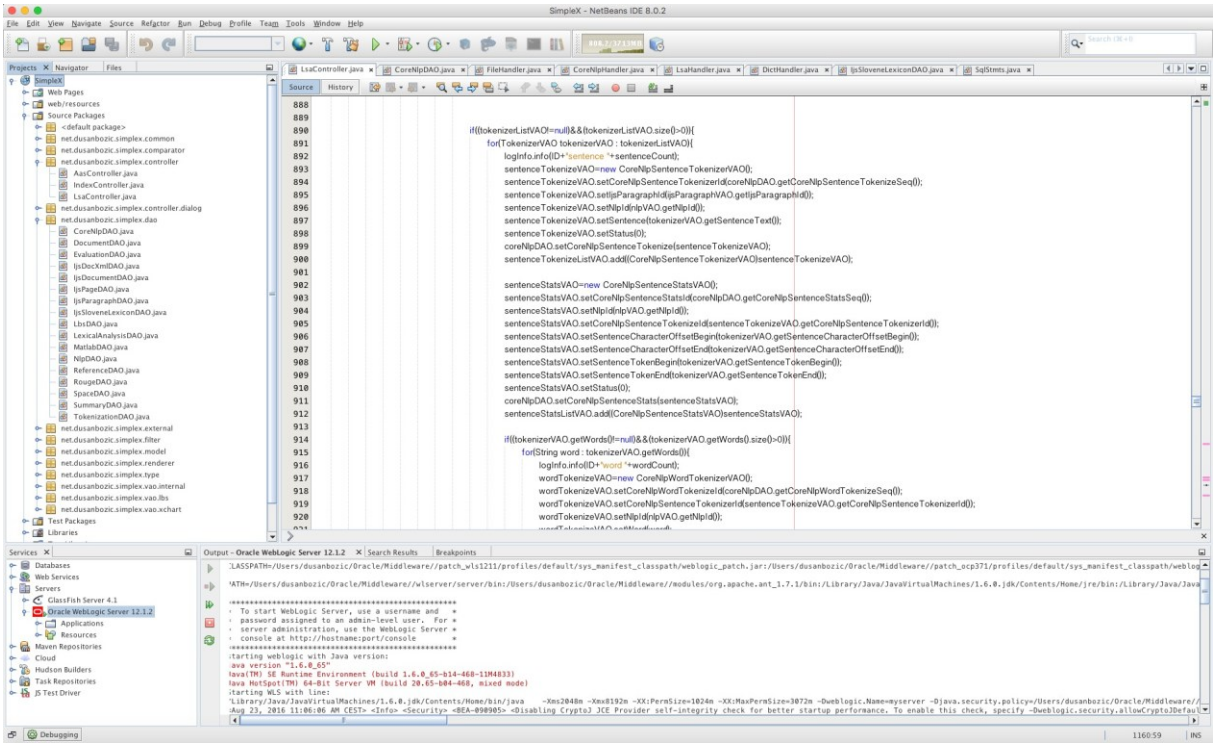
Iz gramatičnih razlogov naravni jeziki poznajo različne besedne in skladenjske vrste. Za prepoznavo skladenjskih oblik, lematiziranje in semantično analizo smo uporabili slovenski leksikon [51], slovenski morfološki leksikon [52], šifrant besednih [53] in skladenjskih oblik ter semantični slovenski leksikon sloWNet [54].

4.12 NetBeans

Razvoj programa za povzemanje besedil se je izvajal v razvojnem okolju NetBeans. V razvojno okolje je bila dodana prosto dostopna knjižnica za delo z ZK ogrodjem. Zaradi obsežnega projekta smo razvojnemu okolju dodelili spominske vire v velikosti 4GB. Konfiguracijska vrstica je bila zavedena v konfiguracijski datoteki okolja netbeans.conf:

netbeans_default_options="-J-client -J-Xss512m -J-Xms2048m -J-Xmx4096m -J-XX:PermSize=2048m -J-Dapple.laf.useScreenMenuBar=true -J-Dapple.awt.graphics.UseQuartz=true -J-Dsun.java2d.noddraw=true -J-Dsun.java2d.dpiaware=true -J-Dsun.zip.disableMemoryMapping=true"

Iz NetBeans okolja smo aplikacijo zaganjali, razhroščevali in testirali, kar je razvidno na **Sliki 6**.



Slika 6: Razvojno okolje NetBeans.

POGLAVJE 5

5. SimpleX

Procesiranje naravnih jezikov je tesno povezano z interakcijo uporabnika z računalnikom preko uporabniškega vmesnika. Vmesnik je lahko ukazna konzola ali prijazen grafičen vmesnik. Na računalniku se izvaja program, ki naj bi z različnimi pristopi izpeljal povzemanje ter sproti obveščal o dogajanju z namenom, da lahko uporabnik sproti sprejema odločitve o izboljšanju samega algoritma, kot tudi njegovih nastavitev.

Danes se rešitve na področju procesiranja naravnih jezikov implementira v programske pakete. Na ta način se kompleksne rešitve kot tudi module, povezuje med seboj. Poleg tega nove rešitve koristijo iste module, ki skrbijo za CRUD operacije in interakcijo s človekom ter integracijo z zunanjimi sistemi. Za izdelavo magistrske naloge smo implementirali samostojen spletni program SimpleX.

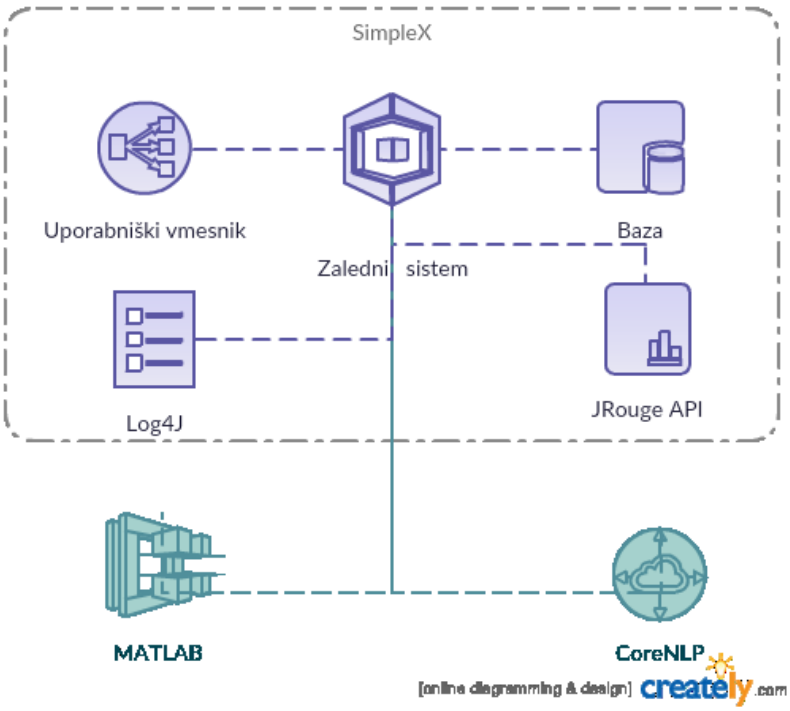
Seznam na internetu najdenih študijskih projektov univerz, komercialnih paketov in brezplačnih spletnih rešitev, ki smo ga uporabili v magistrski nalogi, je: CoreNLP (Stanford University), Summarist (University of Soutern California), SweSum (Royal Institute of Technology - Sweden), FociSum; MultiGen; CPS (Columbia University), Trestle (The Sheffiled University), Websumm (Mitre), InTEXT (InTEXT SYSTEMS), IslandInTEXT (Island soft), Vizie (CSIRO), MS Word Auto Summarization (Microsoft) GreatSummary, Smmry, FreeSummarizer, Summify in Autosummarizer [55-60].

V nadaljevanju je predstavljen program Simplex in njegovi rezultati povzemanja.

5.1 Zahteve za delovanje in arhitektura sistema

Program potrebuje za nemoteno in tekoče delovanje računalnik z dovolj procesorske moči in spomina. Za pogosto izvajanje mora imeti računalnik dovolj dobro hlajenje, sicer pride do pregrevanja in odpovedi. V fazi razvoja se je celotno okolje (aplikacijski strežnik WebLogic, spletni strežnik Apache, MySQL baza, NetBeans razvojno okolje in sam program SimpleX, ki z vsemi knjižnicami zasede 580MB) izvajalo na procesorju 2.8Ghz Intel Core i7, ki je imel na voljo 16GB spomina in SSD disk. Iteracija zagona povzemanja povprečnega akademskega dela na taki konfiguraciji računalnika potrebuje več kot minuto.

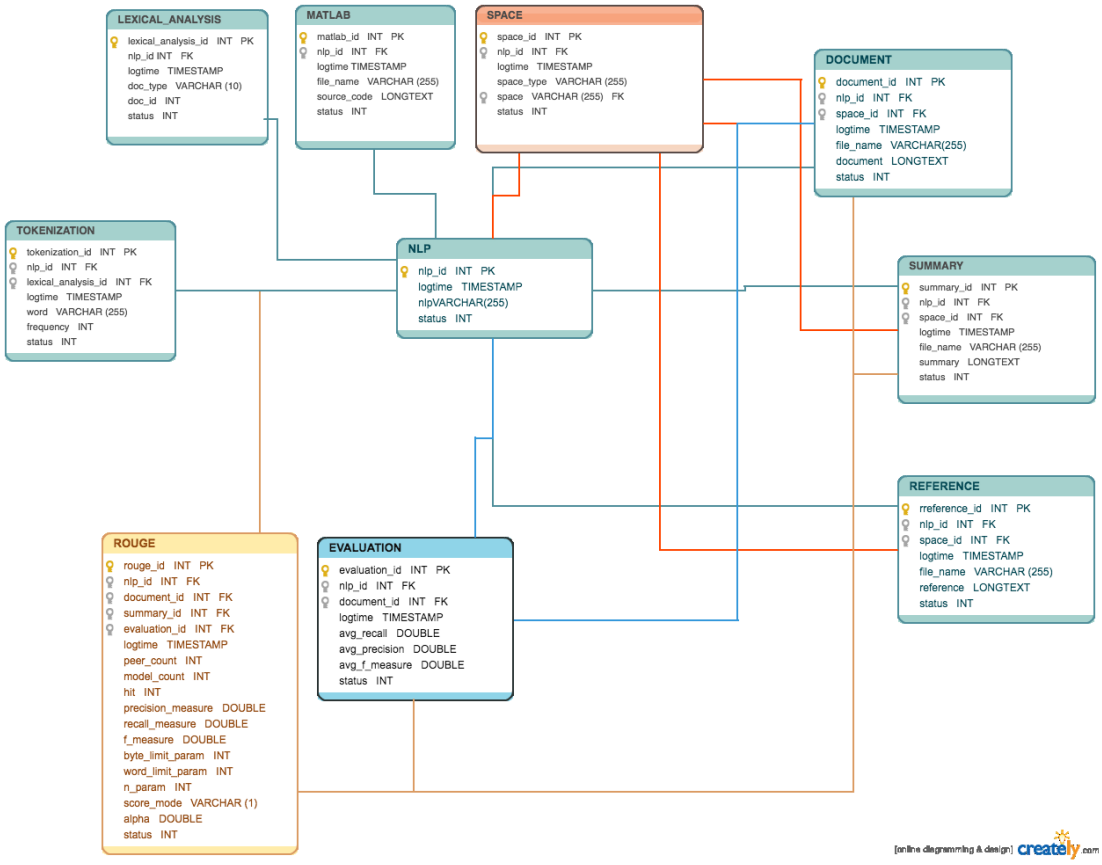
Program za svoje delovanje potrebuje aplikacijski strežnik, ki prepozna spletni Java paket war. Za shranjevanje podatkov potrebuje bazo, kamor se shranjujejo vsi podatki povezani z zajetjem dokumentov, njihovo obdelavo in končnim povzemanjem. Za analizo arhetipov strežnik potrebuje inštalirano instanco programa Matlab. Za uporabo programa potrebujemo spletni brskalnik. Aplikacijo SimpleX opisuje arhitektura na **Sliki 7**.



Slika 7: Arhitekturni opis programa za povzemanje naravnih besedil SimpleX.

Tabele podatkovne baze so vsebinsko razdeljene v tri sklope. Prvi sklop tabel je povezan s samim izvajanjem programa, drugi sklop je povezan z analizo arhetipov, tretji sklop pa je povezan z latentno semantično analizo. Vse tabele imajo lastno sekvenco, datum kreiranja zapisa in status, katerega pomen je odvisen od uporabe. Vsa polja tabel, ki shranjujejo tekstovne nize, imajo določen format sortiranja (angl. *collation*) na utf8mb4_unicode_ci, ki prepoznava tudi šumnike. Vsa tekstovna polja shranjujejo besedila v formatu UTF-8, ki podpira tudi šumnike.

Prvi in drugi sklop tabel ter relacij med njimi je grafično prikazan na **Sliki 8**.



Slika 8: Izsek podatkovnega relacijskega modela baze LBS.

Seznam tabel baze *LBS* po sklopih in opis atributov je:

- tabele, ki so povezane z izvajanjem programa:
    - *t_lbs_nlp*: tabela je osrednja tabela programa, ker vsebuje sekvenco izvajanja, ki je zapisana v vseh ostalih tabelah in poljem *nlp*, ki označuje vrsto povzemanja. Program povzema besedilo s pomočjo analize arhetipov (*AA*) in latentne semantične analize (*LSA*),

    - *t_lbs_space*: tabela shranjuje podatke o poteh do map na datotečnem sistemu v polju *space*. Shranjuje več tipov poti do: izvornih dokumentov, delovne mape, izhodne mape in mape, kjer so primerjalni povzetki (običajno človeški). Glede na tip poti so v polju *space_type* zavedene naslednje vrednosti: *sourcespace*, *workspace*, *peerspace* in *modelspace*;

    - *t_lbs_reference*: tabela shranjuje referenčni dokument in ime datoteke dokumenta. Referenčni dokument se uporabi v fazi ocenjevanja, kjer se primerja avtomatski povzetek z referenčnim;

    - *t_lbs_document*: tabela shranjuje podatke o vhodnih dokumentih v polju tipa CLOB. Program pričakuje dokument v XML obliki;

    - *t_lbs_summary*: tabela shranjuje podatke povzetke oz. izvlečke v polju tipa CLOB in ime datoteke, kjer je rezultat povzemanja shranjen;

- *t_lbs_evaluation*: tabela shranjuje vse tri metrike ocene: natančnost (angl. *precision*), priklic (angl. *recall*) in F-mera (angl. *F-score*);

- *t_lbs_rouge:* tabela shranjuje vse podatke, ki jih vrne knjižnica JROUGE in so povezani z ocenjevanje dokumenta. Poleg metrik ocen so shranjeni parametri s katerimi je bil klican JROUGE;

- tabele, ki so neposredno povezane z analizo arhetipov:
  - *t_lbs_matlab*: tabela shranjuje uvožen algoritem za analizo arhetipov, ki se v procesu analize interpretira preko Matlab interpreterja. Interpreter interpretira vsako vrstico posebej na Matlab instanci, kar je časovno gledano zahteven postopek;

- tabele, ki so neposredno povezane z latentno semantično analizo:
  - *t_lbs_ijs_doc_xml*: šifrantska tabela, ki shranjuje meta podatke o uvozu vhodnih dokumentov za latentno semantično analizo;

  - *t_lbs_ijs_document*: tabela shranjuje vse meta podatke vsebinsko povezane s samim vhodnim dokumentom. Omenjena polja so: cobiss_id, title, author, supervisor, year, publisher_abbr, type (tip akademskega dela), pdf_url (internet povezava do dokumenta), ipd.;

  - *t_lbs_ijs_page*: tabela shranjuje številko zaporedne strani in internetno pot do te strani;

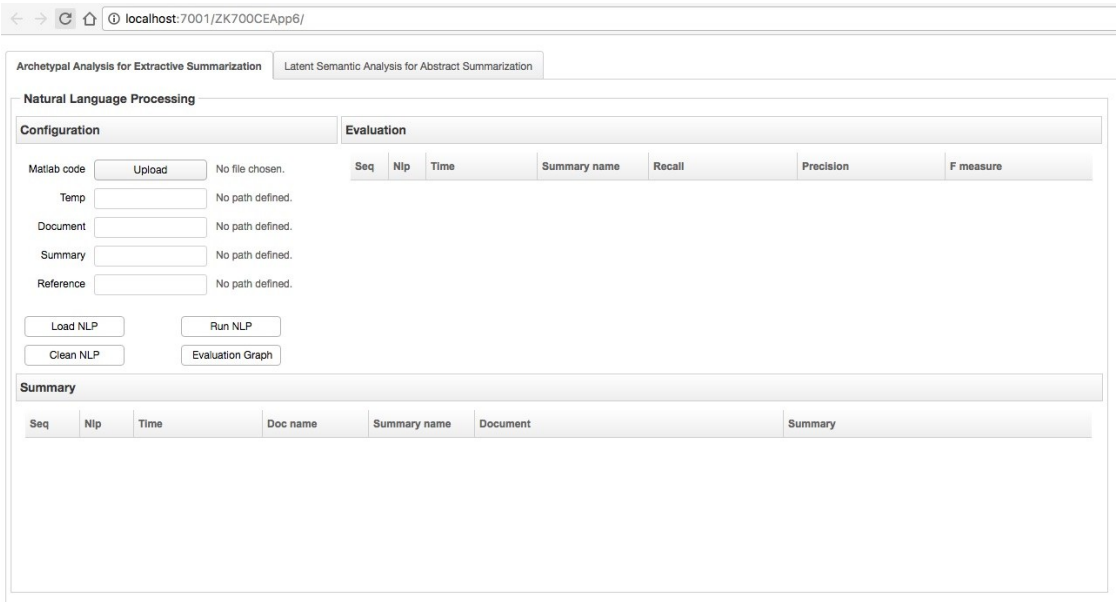
- *t_lbs_ijs_paragraph*: tabela shranjuje odstavke besedila in kateri strani le-ti pripadajo;

- *t_lbs_corenlp_sentence_tokenizer*: tabela shranjuje posamezne stavke in podatke kateremu odstavku pripadajo;

- *t_lbs_ijs_slovene_lexicon*: tabela shranjuje relevantne podatke vseh slovarjev slovenskega jezika. Tabela je podlaga za izvajanje besedne, skladenjske in semantične analize.

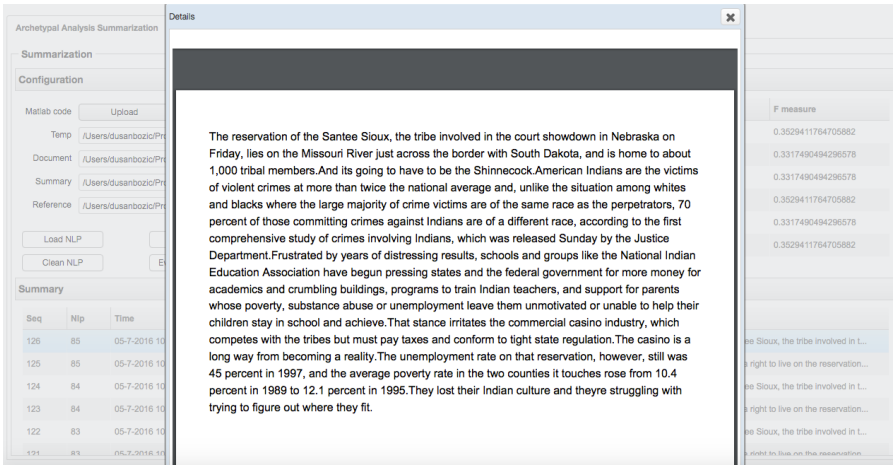
5.2 Uporabniški vmesnik

Uporabniški vmesnik se izvaja v spletnem brskalniku. ZK-ogrodje zagotavlja, neodvisno od vrste brskalnika, enak način prikazovanja in delovanja. V fazi razvoja smo uporabljali brskalnik Chrome (verzija 52 – 64 bit). Aplikacija je razdeljena z listi. Implementirana sta lista *Archetypal Analysis for Extractive Summarization* in *Latent Semantic Analysis for Abstract Summarization*, kar je razvidno na **Sliki 9**.



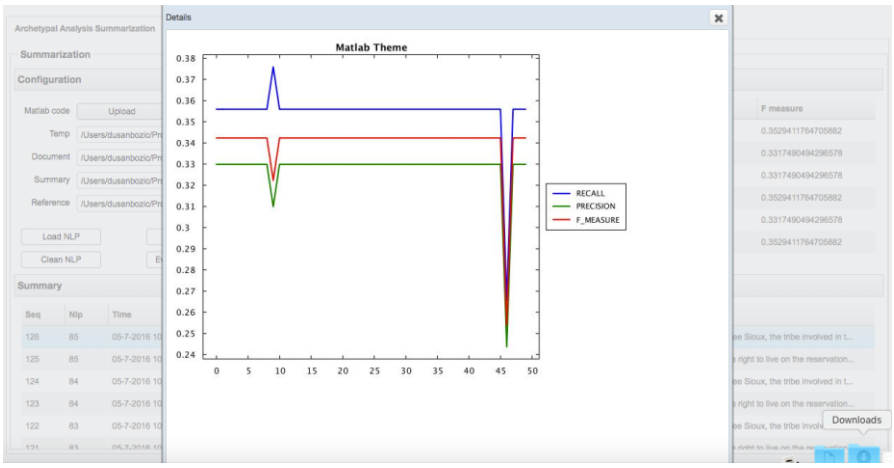
Slika 9: Lista uporabniške maske za različna načina povzemanja.

Prvi list uporabniškega vmesnika, ki se nanaša na povzemanje s pomočjo arhetipov, je implementiran do konca. Omogočen je vnos poti do datotek, zajem algoritma za povzemanje v datoteki Matlab, zagon povzemanja, spremljanje povzetkov in njihovih ocen. Povzetki se prikažejo v obliki PDF, če uporabnik izbere povzetek, kar je razvidno na **Sliki 10**.



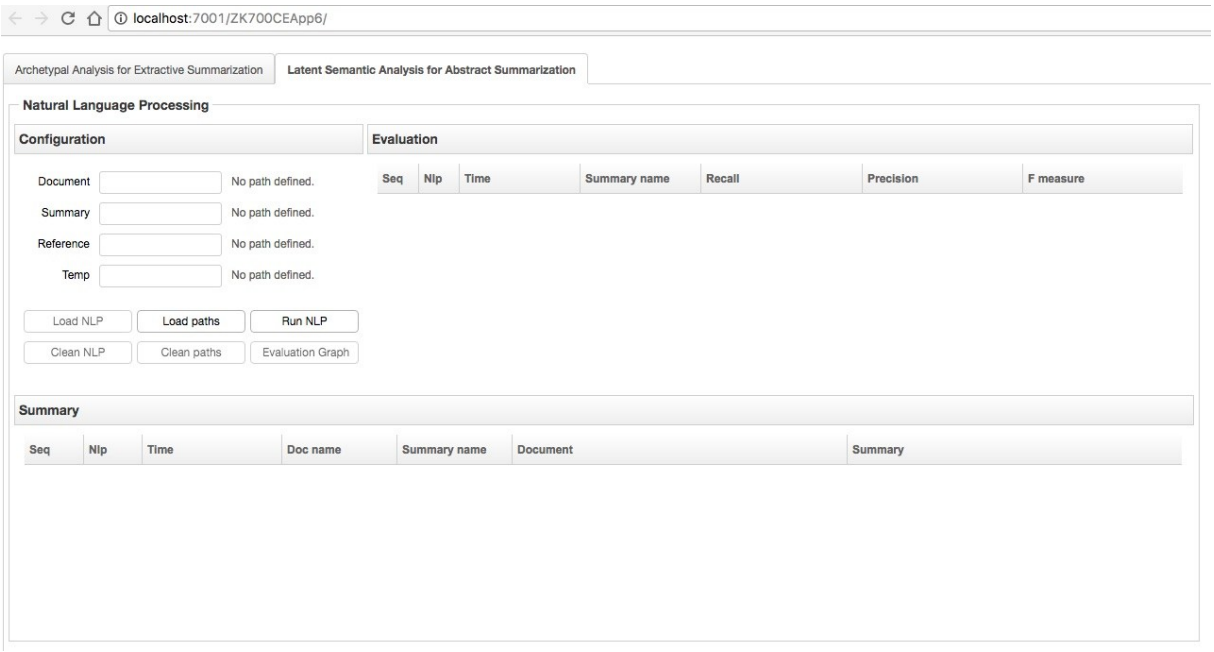
Slika 10: Generiranje dokumenta PDF, ki vsebuje izbrani povzetek.

Uporabniška maska odpira grafični prikaz trenda ocen ob pritisku na gumb *Evaluation Graph*. Rezultati se prikazujejo od starejših proti novejšim (od leve proti desni), kar je razvidno na **Sliki 11**.



Slika 11: Grafični prikaz trenda ocen povzemanja besedil s pomočjo analize arhetipov.

Drugi list je osiromašen glede funkcionalne podpore uporabniške maske. Večina omenjenih funkcionalnosti ni podprtih. Razlog je poleg pomanjkanje časa tudi dejstvo, da je dopolnitev funkcionalnosti enostavna, ker so zgledi za implementacijo funkcionalnosti uporabniške maske implementirani na prvem listu. Večina gumbov je onemogočenih. Ne deluje prikaz povzetkov, grafa in ocen. Delujeta gumba za vzpostavitev poti in gumb za zagon semantičnega povzemanja, kar je razvidno iz **Slike 12**.



Slika 12: Osiromašen drugi list uporabniškega vmesnika za povzemanje besedil s pomočjo semantične analize.

5.3 Najbolj pogoste besede

Tabela *t_lbs_ijs_slovene_lexicon* je bila za namen semantične analize razširjena s poljem *common_word*. To so besede, ki predstavljajo najbolj pogosto pojavitev glede na frekvenco pojavitev. Statistika teh besed je vzeta iz pridobljenega morfološkega slovarja in predstavlja 60% vseh pojavitev. Frekvenca pojavitev je bila izračunana na podlagi korpusa Gigafida. V okviru naloge je bil izdelan seznam najbolj pogostih 63 besed. Te besede so iz besednih vrst,

ki vplivajo na pomen stavka. To so samostalnik, pridevnik, glagol in prislov. Izdelava seznama je bila zahtevna, ker strukturirano datoteko, ki vsebuje frekvence pojavitev besed, sestavlja 29.818.139. vrstic. Najbolj pogoste besede smo uporabili pri semantični analizi. Priložena je **Tabela 3**, ki prikazuje 63 najbolj pogostih besed, njihovo frekvenco pojavitve in oblikoskladenjsko oznako (MSD).

BESEDA	POJAVITEV	MSD	BESEDA	POJAVITEV	MSD
niso	996.561	Gp-stm-d	foto	508.197	Somei
zdaj	993.420	Rsn	časa	506.856	Somer
veliko	987.131	Rsn	letih	501.181	Sosmm
let	868.744	Sosmr	tolarjev	499.929	Sommr
potem	855.740	Rsn	Ljubljana	490.453	Slzei
vse	841.736	Rsn	času	488.152	Somem
bili	839.160	Gp-d-mn	malo	485.538	Rsn
gre	824.251	Ggvste	mogoče	484.788	Ppnsei
danes	802.523	Rsn	ljudi	479.013	Sommr
najbolj	780.477	Rss	mora	475.932	Ggnste
bomo	729.525	Gp-ppm-n	sam	474.588	Ppnmein
treba	725.616	Rsn	strani	467.626	Sozem
res	686.730	Rsn	pravi	461.828	Ggvste
dobro	679.640	Rsn	precej	456.818	Rsn
leto	642.441	Soset	torej	448.998	Rsn
dan	636.602	Sometn	pomeni	446.971	Ggvste
kaj	628.294	Rsn	odstotkov	446.429	Sommr
zato	624.556	Rsn	glede	445.194	Rsn
imajo	607.457	Ggnstm-n	tam	444.952	Rsn
skupaj	604.838	Rsn	bom	438.556	Gp-ppe-n
Slovenije	604.485	Slzer	zakaj	436.699	Rsn
ste	594.230	Gp-sdm-n	delo	435.961	Soset
dela	583.625	Soser	spet	434.540	Rsn
manj	577.440	Rsr	tokrat	434.530	Rsn
boste	571.511	Gp-pdm-n	svetu	428.675	Somem
del	558.741	Somei	ljudje	420.352	Sommi
Sloveniji	557.546	Slzem	imeli	419.717	Ggnd-mm
letos	550.393	Rsn	Slovenija	416.221	Slzei
predsednik	530.214	Somei	imel	412.866	Ggnd-em
biti	528.243	Gp-n	povsem	402.756	Rsn
evrov	513.803	Sommr	dovolj	400.809	Rsn
nato	510.041	Rsn			

Tabela 3: Seznam 63 najbolj pogostih besed, skupaj s frekvenco, na podlagi korpusa Gigafida in pripadajoče oblikoskladenjske oznake.

Šifrant oblikoskladenjskih oznak je dosegljiv na spletnem naslovu [53]. Oblikoskladenjska specifikacija JOS vsebuje 1902 MSD šifri.

5.4 Zajem podatkov

Vhodni dokumenti, ki so predmet obravnave magistrske naloge, so akademski prispevki. Dostop do digitaliziranih dokumentov, zapisanih v strukturirani obliki XML, smo dobili s strani Inštituta Jožef Štefan.

Zajem vhodnih dokumentov se v obeh primerih zajema iz datotečnega sistema. Pri analizi arhetipov moramo vhodni strukturiran dokument spremeniti v nestrukturirano obliko na način, da je vsak stavek v svoji vrstici. Podobno velja za referenčni dokument. Pri analizi s pomočjo latentne semantične analize se pričakuje strukturiran vhodni dokument v obliki XML na način, kot ga imajo shranjenega na Inštitutu Jožef Štefan. Vsi dokumenti, ki so bili obravnavani, so že imeli v okviru besedila človeške povzetke. Pri analizi arhetipov je potrebno ročno poskrbeti, da se človeški povzetek ne nahaja v obravnavanem besedilu. Pri semantičnem povzemanju program, na podlagi ključne besede, ki je zavedena v konfiguracijski datoteki, sam prepozna začetek besedila, ki ne vključuje človeškega povzetka.

Pri testiranju uspešnosti obeh načinov povzemanj je bilo vključenih 100 akademskih prispevkov.

5.5 Predobdelava podatkov

Povzemanje s pomočjo analize arhetipov izvaja predobdelavo podatkov s pomočjo knjižnice Text to Matrix Generator (TMG) [61]. V nastavitvah klica funkcije tmg se določi odstranjevanje ločil, raba delimiterjev, sprememba v nedoločnik (angl. *stemming*), izvedba normalizacije in uteževanja, izločitev kratkih oz. dolgih terminov, ipd.

Povzemanje s pomočjo latentne semantične analize izvede korake semantične predobdelave podatkov za algoritem LSA. Prvi korak je dekompozicija besedila na odstavke, stavke in besede (angl. *tokenization*). Rezultat so osnovne enote (besede). V naslednjem koraku jim določimo besedno vrsto in skladenjsko obliko s pomočjo slovenskega slovarja. V tretjem koraku odstranimo ločila iz besedila: pika, vejica, podpičje, dvopičje, klicaj, vprašaj, okrogli oklepaj in zaklepaj ter oglati oklepaj in zaklepaj. V četrtem koraku pripravimo repozitorij praznih besed (angl. *stop-words*), ki ne vplivajo na pomen. V slovenskem slovarju so v polju *stop_word* označene vse besede, ki ne vplivajo na pomen stavka. To so naslednje besedne vrste: členek, medmet, okrajšava, predlog, številnik, veznik in zaiimek. Od skupno 2.786.365 besed je takih 25.008. Repozitorij shranimo v pripravljeno mapo knjižnice TML, ki te besede odstrani v predfazi zagona algoritma LSA. V zadnjem koraku se izvede lematiziranje. V fazi lematiziranja se vsaki besedi opredeli njen nedoločnik.

Med raziskovanjem anotiranih besed z oznako MSD so se pokazale težave, ki vplivajo na izvajanje latentne semantične analize, kot tudi na izdelavo semantičnega povzetka iz izvlečka. Operaciji lematiziranja in določitve oblikoskladenjske oznake nista bijektivni, kar pomeni, da preslikavi nista obratno enolični. Obstajajo primeri, kjer isti besedi pripada več nedoločnikov. Primer take besede je *dojeti*, ki ima nedoločnik *dojet* in *dojeti*. Obstajajo primeri, kjer ima ista beseda dve različni oblikoskladenjski oznaki in hkrati isti nedoločnik. Primer take besede je *cviknili*, ki ji pripada nedoločnik *cvikniti* in hkrati dve različni oznaki MSD: prva opisuje besedo kot glagol / moški spol / množina in druga opisuje besedo kot glagol / ženski spol / dvojina. Za kvalitetno izdelavo semantičnih pravil smo predpostavljali, da je število besed, ki imajo dvoumno anotacijo MSD in lematizacijo, zanemarljivo malo, vendar je naša analiza pokazala, da temu ni tako. Za pravilno opredelitev zapisa MSD in nedoločnika bi bila potrebna kompleksna semantična analiza, kar presega okvir naloge. Dvoumno anotirane

primere program posebej označi v tabeli *t_lbs_corenlp_word_tokenizer* v polju status. Če je status 1, imamo opravka z dvoumjem, sicer ne. V nalogi smo upoštevali samo nedvoumne MSD oznake. V sosednjem polju *ijs_slovene_lexicon_id* je zaveden ključ (sekvenca) do nedoločnika in oznake MSD, ki je bila dodeljena besedi. Sekvenca *ijs_slovene_lexicon_id* je zavedena kot tuji ključ tabele *t_lbs_ijs_slovene_lexicon*, kjer so zavedene vse šifre MSD, oblikoskladenjski zapisi in nedoločniki.

5.6 Analiza

5.6.1 Analiza arhetipov

Algoritem analize arhetipov v prvem koraku iteracije izvede matrično faktoriranje s pomočjo knjižnice Text to Matrix Generator (TMG) [61]. Kot rezulat faktoriranja dobimo: matriko indeksov terminov dokumenta (ang. *term-document matrices*), slovar besed besedila, strukturo update_struct, ki opisuje celotno stanje izvedbe TMG. Ostale vrednosti odgovora TMG so vezane na statistiko dokumenta. Nad zgrajeno matriko indeksov se požene funkcija vsm (angl. *vector space model*) knjižnice TMG. Funkcija vsm v iteraciji po vseh besedah dokumenta izdela model prostorskih vektorjev pri podani matriki indeksov terminov dokumenta in matriki poizvedb ter vrne matriko sortiranih koeficientov podobnosti (SC) in matriko korespondenčnih indeksov dokumenta (DOC_INDS) za besedo dokumenta. V gnezdeni iteraciji se na podlagi SC in DOC_INDS izdela matrika podobnosti na nivoju stavka. Ta informacija se shrani v matriki sentenceSimMatrix. Zadnja iteracija programa je povezana z iskanjem osnovne konveksne ovojnice v množici arhetipov stavkov. Iteracija se izvede tolikokrat, kolikor imamo določeno število arhetipov (AN). V našem primeru imamo dve: 2 in 16. Znotraj iteracije se pripravi vektor stavkov izvornega dokumenta, ki se matrično zmnoži z matriko podobnosti stavkov. Rezultat shranimo v isto matriko podobnosti stavkov. Funkcija PCHA na vhodu sprejme matriko podobnosti stavkov, število arhetipa, prag in druge

nastavitve. PCHA reši optimizacijski problem iskanja dominantne konveksne ovojnice (ang. *principal convex hull*); tj. K-1 dimenzionalne konveksne ovojnice, ki najbolje opišejo podatke.

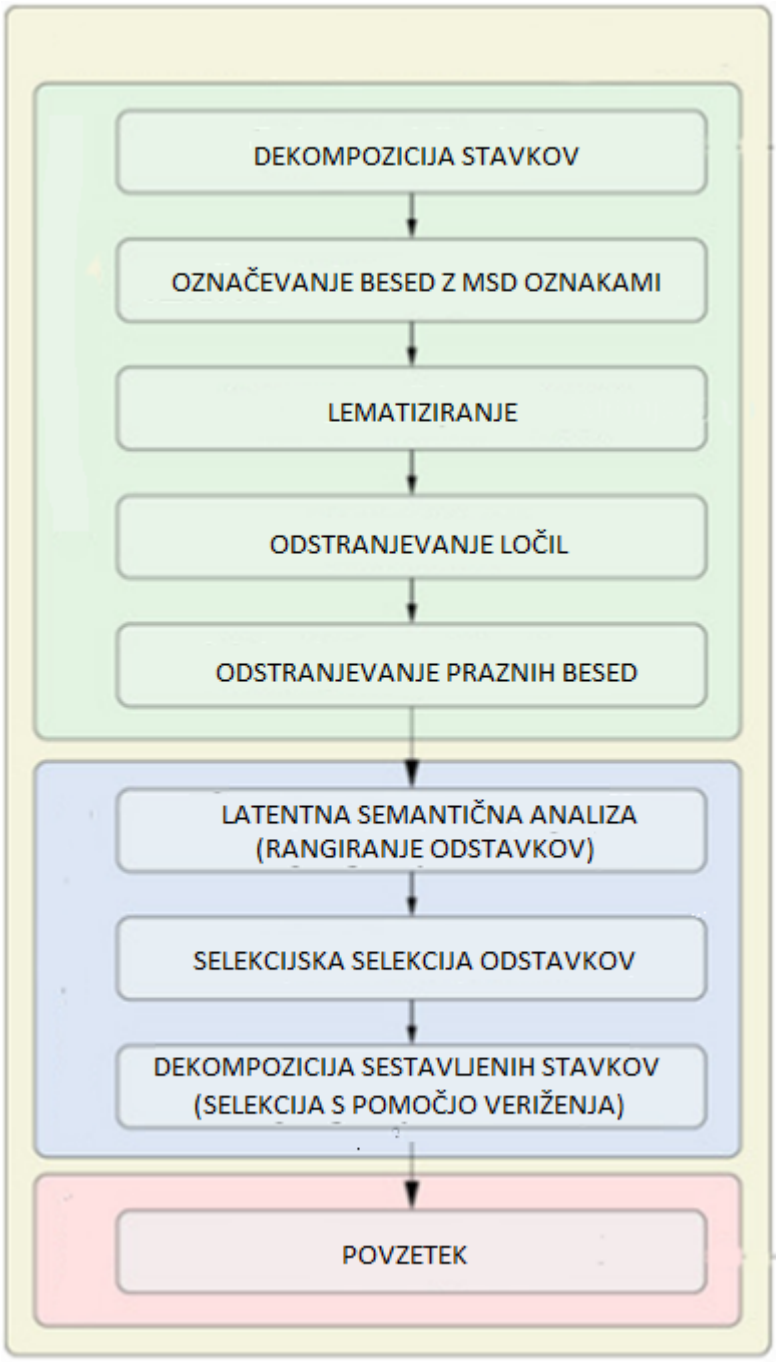
Kandidati z največjimi vrednostmi značilk arhetipov so tako izbrani. Stavki, ki so krajši od 30 znakov, so izločeni. Ko izvleček doseže velikost 600 besed, se proces povzemanja ustavi.

5.6.2 Semantična analiza

Povzemanje s pomočjo latentne semantične analize smo pognali nad delno predprocesiranimi stavki. V predfazi analize knjižnica TML izvede dokončno pripravo podatkov z odstranjevanjem praznih besed na podlagi pripravljenega repozitorija. Repozitorij praznih besed knjižnica TML pričakuje v mapi /opt/tml/stopwords/. V fazi analize algoritem LSA algebraično in statistično povzame pomen besed in podobnost stavkov glede na uporabo besed v kontekstu. Knjižnica TML izvede analizo na način, da primerja podobnost besedila, ki je shranjeno v ločenih datotekah. V ta namen smo v ločene datoteke shranili posamezne odstavke predpripravljenega besedila in celotno predpripravljeno besedilo. Na izhodu nam knjižnica TML vrne rangirane odstavke glede na pomembnost v kontekstu celotnega besedila. Rangirane odstavke smo izbrali po kriteriju položaja odstavka v celotnem besedilu in rangu. Celotno besedilo smo razdelili na tretjine. Iz vsake tretjine smo vzeli dva najbolj obetavna odstavka, ki po svoji dolžini ne smeta biti krajša od 100 besed. Tako pridobljene odstavke smo združili v prototip končnega povzetka, ki je po dolžini daljši od 600 besed. Izbrane odstavke spremenimo v izvorne stavke na podlagi predhodno vzpostavljene povezave med izvornimi in lematiziranimi stavki. Za končni povzetek izvedemo semantično analizo v naslednjih korakih. V prvem koraku iz izbranih odstavkov za vsak stavek preverimo prisotnost najbolj pogostih besed (angl. *common words*). Stavki, ki imajo prisotne pogoste besede, nakazujejo na semantično povezanost in jih ohranimo. Ohranimo jih tako, da jih označimo. V drugem koraku poiščemo sestavljene stavke, ki so povezani z vezniki "in", "ali"

in "ter". Preverimo, če so vsi osnovni stavki obravnavanega sestavljenega stavka oblike osebek-povedek. Samostalnik opredelimo kot osebek, če je v imenovalniku in ima isti spol in število kot glagol, ki mu sledi. Pri tej analizi se omejimo samo na nedvoumne MSD oznake. V nasprotnem primeru razbijanje tako sestavljenih stavkov preskočimo. V primeru uspešnega razbitja sestavljenega stavka s pomočjo veriženja določimo sopomenke razbitega stavka in stavkov v soseščini. Če najdemo ujemanje, stavek označimo in ga obdržimo. Kriterij za izdelavo končnega povzetka je največja dovoljena dolžina. Pri selekciji stavkov za končni povzetek imajo označeni stavki prednost pred ostalimi znotraj izbranih odstavkov. Pomeni, da jih vključimo v končni povzetek pred zadnjim ciklom izbire preostalih stavkov. Preostanek vsebine povzetka pridobimo tako, da začnemo z izbiro stavkov od začetka vsakega odstavka, dokler ne dosežemo maksimalne dovoljene dolžine povzetka. Tako pridobljeni povzetek je rezultat povzemanja s pomočjo semantične analize. Ideja predstavljenega semantičnega povzemanja je, da z obravnavo odstavkov zajamemo širši kontekst velikega besedila. Z uporabo hevristik na nivoju besed (npr. pogoste besede) in stavkov (razbijanje) želimo vplivati na samo strukturo odstavka in njegovo dodano vrednost k boljšemu povzetku.

Na **Sliki 13** so zapisani glavni koraki procesa semantične analize.



Slika 13: Glavni koraki procesa semantične analize (v zelenem področju so koraki predprocesiranja, v modrem koraki semantične analize in na koncu rezultat povzemanja).

5.7 Rezultati

Pri ocenjevanju povzetkov smo uporabljali mero vsebine (angl. *content-based*) na podlagi prekrivanja enot ali n-gramov (besed). Mero vsebine smo ovrednostili s pomočjo metrik natančnost (angl. *Precision*), priklic (angl. *Recall*) in F-mera (angl. *F-measure*), ki v našem primeru določajo stopnjo ujemanja med sistemskim in človeškim povzetkom na nivoju besed oziroma n-gramov. Velikost prekrivanja določa velikost n-gramov oz. število besed. V našem primeru smo uporabljali n-grame velikosti 1, 2 in 3. Omenjene metrike smo dobili s pomočjo programskega paketa Rouge, kjer smo uporabili tehniko Rouge-N. Metoda Rouge-N primerja N-grame dveh povzetkov in šteje število ujemanj n-gramov. Izračuna se na podlagi formule :

$$ROUGE - N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count_{match}(Ngram)}{\sum_{S \in Summ_{ref}} \sum_{Ngram \in S} Count(Ngram)}$$

pri čemer N pomeni dolžino n-grama, $Count_{match}(Ngram)$ je maksimalno število prekrivajočih n-gramov v sistemskem in referenčnem povzetku. $Count(Ngram)$ je število N-gramov v referenčnem povzetku [49].

Za empirično ocenjevanje našega pristopa povzemanja besedil smo uporabljali korpus besedil KAS proto z Inštituta Jožef Štefan. Korpus vsebuje preko 50.000 dokumentov.

Iz korpusa KAS proto smo izbrali 100 dokumentov. Tako pripravljene dokumente smo poslali v analizo arhetipov in semantično analizo z namenom izdelati povzetke v velikosti med 250 in 600 besed. Ekstraktne povzetke s pomočjo analize arhetipov smo primerjali s človeškimi

povzetki. Podobno smo primerjali semantične povzetke s človeškimi povzetki. Preverili smo medsebojno podobnost obeh sistemskih povzetkov z njuno medsebojno primerjavo.

5.7.1 Izvorni dokumenti

Nazivi izvornih dokumentov so sestavljeni iz zaporedne številke, naziva dokumenta in identifikatorja izvajanja, ki se navezuje na zapise v bazi. Vsi izbrani dokumenti vsebujejo povzetek avtorja besedila. V Prilogi 1 so našteti nazivi vseh izvornih dokumentov, njihovi avtorji in fakultete, ki so dokument izdale.

5.7.2 Analiza rezultatov

V prilogah od 2 do 10 so prikazane ocene natančnosti, priklica in F-mere povzemanja z arhetipi, semantičnim in človeškim povzemanjem po vseh izvornih dokumentih s pomočjo tehnike Rouge-N po n-gramih dolžine 1, 2 in 3. V nadaljevanju so tabelarično prikazane srednje vrednosti in standardni odklon za F-mero. Izvedli smo statistiko parni T-test za ugotavljanje značilnih razlik med povzemanjem z arhetipi in semantičnim povzemanjem glede na vse tri dolžine n-gramov. Dodatno smo uporabili parni T-test za ugotavljanje značilnih razlik istega načina povzemanja pri uporabi različnih dolžin n-gramov. Porazdelitve so prikazane s pomočjo histogramov. Statistične metode smo izvedli s programom Stata verzija 12. Primerjava povzemanj je prikazana na **Sliki 16**.

V **Tabeli 4** so prikazane povprečne vrednosti in standardni odkloni F-mere pri uporabi Rouge-1 za ocenjevanje podobnosti povzetka z arhetipi in človeškimi povzetki.

Rouge-1	povpr. vr. (F-mera)	st.dev. (F-mera)
AA/človeški	0,1918	0,0868
LSA/človeški	0,1636	0,0715
AA/LSA	0,2754	0,0915

Tabela 4: Primerjava analize arhetipov in ročnih povzetkov z uporabo tehnike Rouge-1.

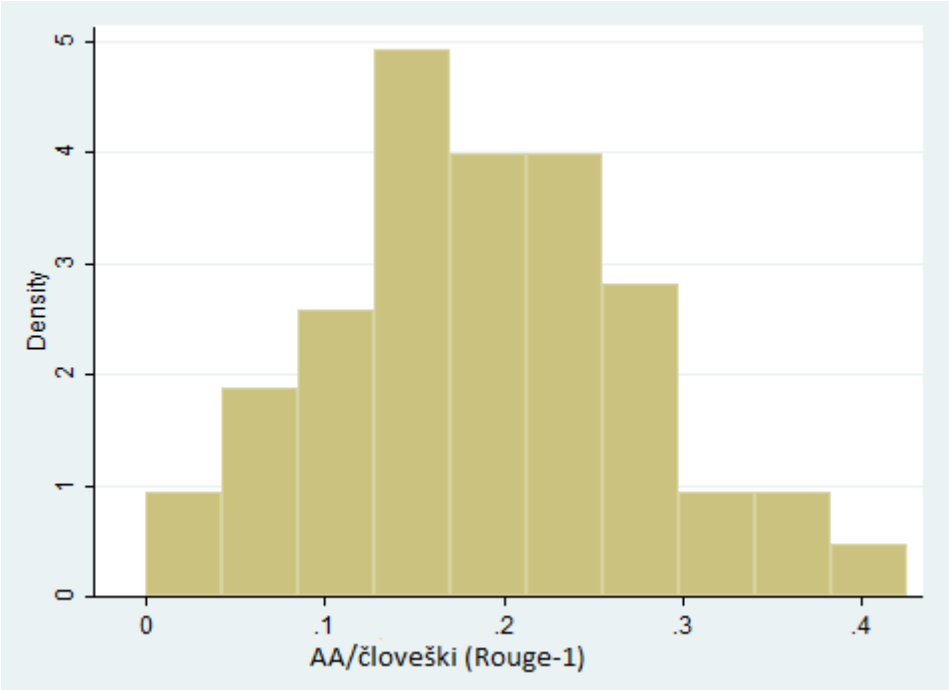
Pri 95% intervalu zaupanja smo izvedli parni T-test, s katerim smo preverili ničelno hipotezo, da ni statistične razlike med ocenami povzemanja AA/človeški in ocenami povzemanja LSA/človeški z uporabo Rouge-1.

Spremenljivke	Opazovanja	Sr.vr.	Std. napak	Std. dev.	[95% stopnji zaup.]
AA/človeški(Rouge-1)	100	0.1917	0.0087	0.08733	[0.1744 - 0.2090]
LSA/človeški(Rouge-1)	100	0.1636	0.0071	0 .07193	[0.1493 - 0 .1779]
Razlike	100	0 .0281	0.0070	0.07043	[0.0141 - 0.0421]

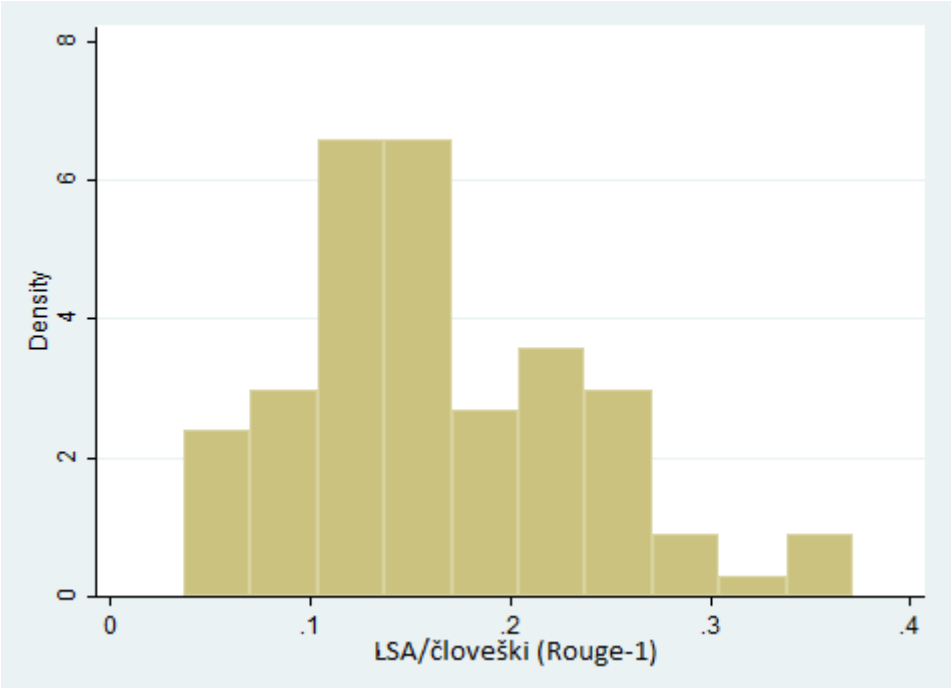
Rezultat testa: Pr(|T| > |t|) = 0,0001.

Pri 95% intervalu zaupanja smo izvedli parni T-test. Na podlagi P-vrednosti (P<0,05) lahko sklepamo, da sta povzemanji AA/človeški in LSA/človeški statistično različni pri Rouge-1. Povprečna vrednost podobnosti povzemanja z arhetipi s človeškimi je večja od semantične analize. Lahko sklepamo, da je povzemanje z analizo arhetipov pri ocenjevanju z Rouge-1 statistično značilno boljše od semantičnega povzemanja.

Na **Sliki 14** je prikazan histogram porazdelitve vrednosti ocene podobnosti povzemanja z arhetipi in človeškim povzemanjem z uporabo Rouge-1. Podobno je na **Sliki 15** prikazan histogram porazdelitve vrednosti ocene podobnosti semantičnega in človeškega povzemanja z uporabo Rouge-1. Histograma sta si po obliki porazdelitve podobna.

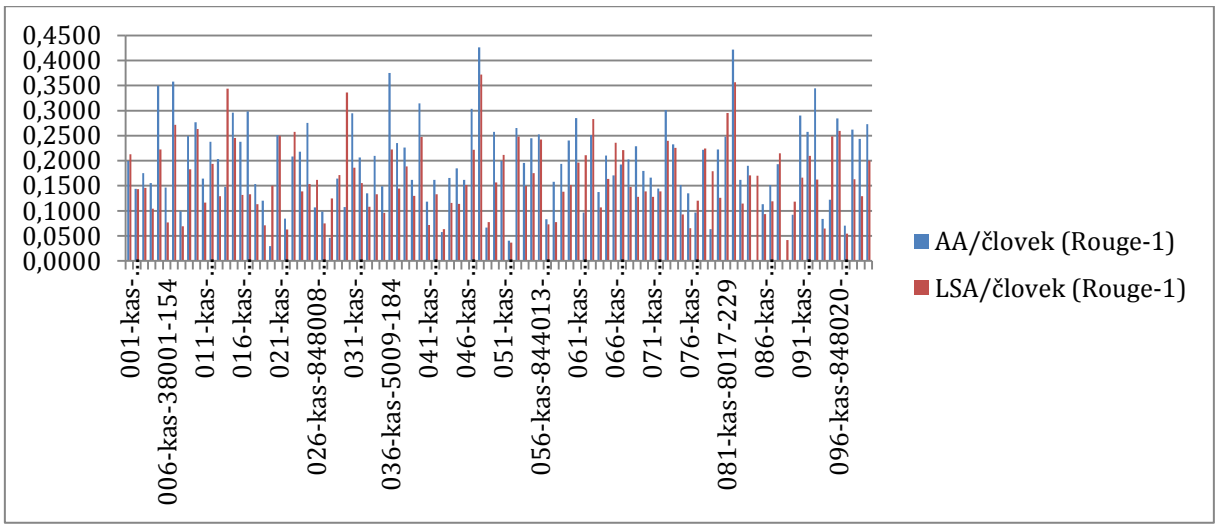


Slika 14: Histogram ocen podobnosti povzemanja z arhetipi in človeškim povzemanjem z uporabo Rouge-1. Na abcisni osi grafa so zapisane ocene podobnosti (F-mera) med povzemanji z arhetipi in človeškimi povzetki. Ocene so grupirane v 10 skupin. Na ordinatni osi je zavedeno število primerov ocen podobnosti (F-mera), ki pripadajo eni od 10 skupin. Graf je izdelan na podlagi **Priloge 2**.



Slika 15: Histogram ocen podobnosti povzemanja semantičnega in človeškega povzemanja z uporabo Rouge-1. Na abcisni osi grafa so zapisane ocene podobnosti (F-mera) med semantičnim povzemanjem in človeškimi povzetki. Ocene so grupirane v 10 skupin. Na ordinatni osi je zavedeno število primerov ocen podobnosti (F-mera), ki pripadajo eni od 10 skupin. Graf je izdelan na podlagi **Priloge 5**.

Izračun parnega T-testa lahko vizualno preverimo na **Sliki 16**. Vidimo, da se v večini primerov povzemanje z analizo arhetipov izkaže boljše v primerjavi s semantičnim povzemanjem.



Slika 16: Primerjava semantične analize in analize arhetipov z ročnimi povzetki. Na abcisni osi grafa so zapisani vsi izvorni dokumenti v takem vrstnem redu kot so v **Prilogi 2** in **Prilogi 5**. Na ordinatni osi so zapisane ocene podobnosti (F-mera) med povzemanji, ki so zavedeni v legendi grafa.

V **Tabeli 5** so prikazane povprečne vrednosti in standardni odkloni F-mere pri uporabi Rouge-2 za ocenjevanje podobnosti povzetka z arhetipi in človeškimi povzetki.

Rouge-2	povpr. vr. (F-mera)	st.dev. (F-mera)
AA/človeški	0,0387	0,0380
LSA/človeški	0,0366	0,0334
AA/LSA	0,0961	0,0793

Tabela 5: Primerjava analize arhetipov in ročnih povzetkov z uporabo Rouge-2.

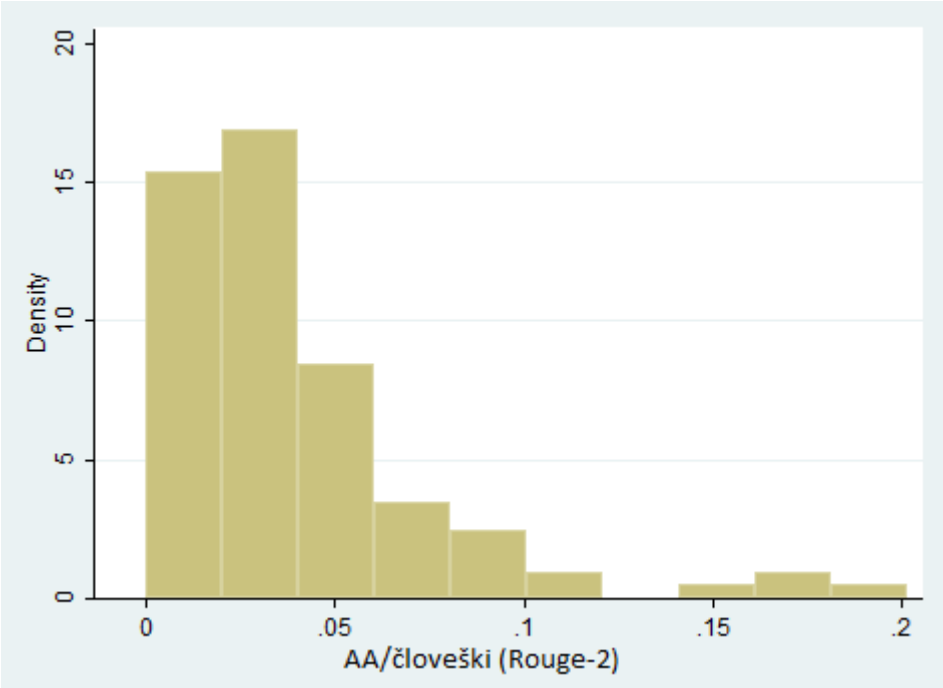
Pri 95% intervalu zaupanja smo izvedli parni T-test, kjer smo preverili ničelno hipotezo, da ni statistične razlike med ocenami povzemanja AA/človeški in povzemanja LSA/človeški z uporabo Rouge-2.

Spremenljivke	Opazovanja	Sr.vr.	Std. napak	Std. dev.	[95% stopnji zaup.]
AA/človeški(Rouge-1)	100	0.0387	0.0038	0.03828	[0.0311 - 0.0463]
LSA/človeški(Rouge-1)	100	0.0365	0.0033	0 .03358	[0.0299 - 0 .0432]
Razlike	100	0 .0021	0.0039	0.03904	[0.0056 - 0.0098]

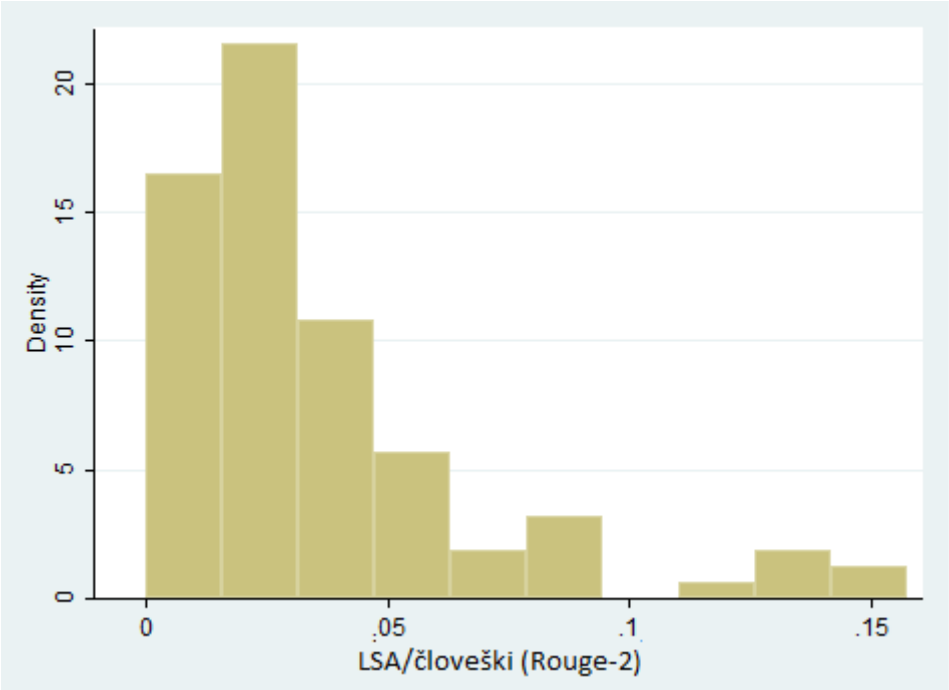
Rezultat testa: Pr(|T| > |t|) = 0,5846.

Ker je P-vrednost>0,05 ne moremo zavrniti ničelne hipoteze. Iz tega sledi, da sta povzemanji AA/človeški in LSA/človeški nista statistično značilno različni pri Rouge-2. Sklepamo, da se kvaliteta obeh načinov povzemanja približa, ko zaostrimo pogoje ocenjevanja podobnosti v primerjavi s človeškimi povzetki.

Na **Sliki 17** je prikazan histogram porazdelitve vrednosti ocene podobnosti povzemanja z arhetipi in človeškim povzemanjem z uporabo Rouge-2. Podobno je na Sliki 18 prikazan histogram porazdelitve vrednosti ocene podobnosti semantičnega in človeškega povzemanja z uporabo Rouge-2. Histograma sta si po obliki podobna.



Slika 17: Histogram ocen podobnosti povzemanja z arhetipi in človeškim povzemanjem z uporabo Rouge-2. Na abcisni osi grafa so zapisane ocene podobnosti (F-mera) med povzemanji z arhetipi in človeškimi povzetki. Ocene so grupirane v 10 skupin. Na ordinatni osi je zavedeno število primerov ocen podobnosti (F-mera), ki pripadajo eni od 10 skupin. Graf je izdelan na podlagi **Priloge 3**.



Slika 18: Histogram ocen podobnosti povzemanja semantičnega in človeškega povzemanja z uporabo Rouge-2. Na abcisni osi grafa so zapisane ocene podobnosti (F-mera) med semantičnim povzemanjem in človeškimi povzetki. Ocene so grupirane v 10 skupin. Na ordinatni osi je zavedeno število primerov ocen podobnosti (F-mera), ki pripadajo eni od 10 skupin. Graf je izdelan na podlagi **Priloge 6**.

V **Tabeli 6** so prikazane povprečne vrednosti in standardni odkloni F-mere pri uporabi Rouge-3 za ocenjevanje podobnosti povzetka z arhetipi in človeškimi povzetki.

Rouge-3	povpr. vr. (F-mera)	st.dev. (F-mera)
AA/človeški	0,0130	0,0194
LSA/človeški	0,0135	0,0237
AA/LSA	0,0618	0,0727

Tabela 6: Primerjava analize arhetipov in ročnih povzetkov z uporabo Rouge-3.

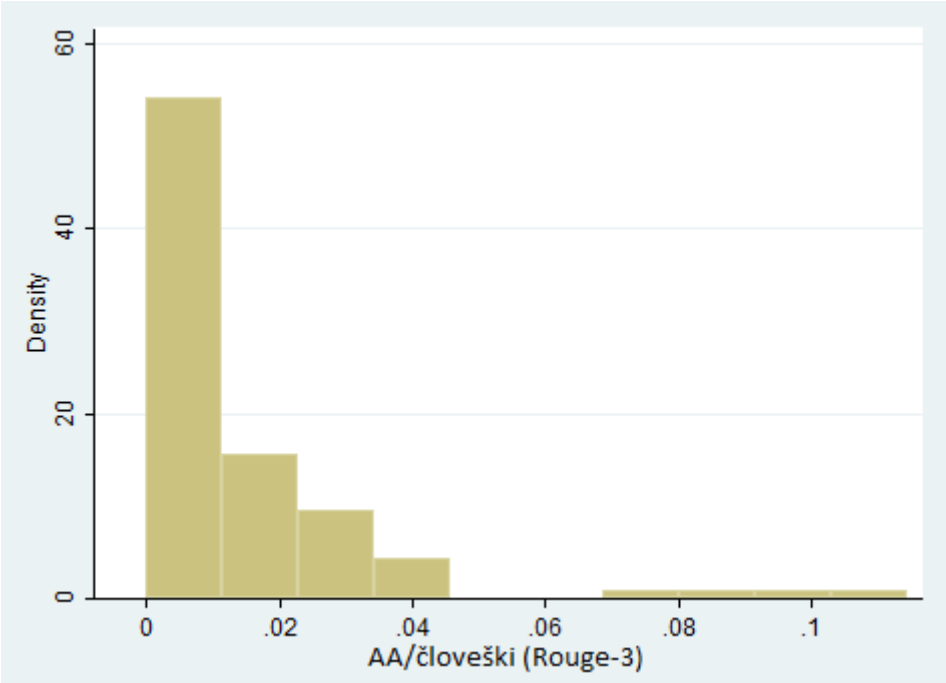
Pri 95% intervalu zaupanja smo izvedli parni T-test, kjer smo preverili ničelno hipotezo, da ni statistične razlike med ocenami povzemanja AA/človeški in povzemanja LSA/človeški povzemanja z uporabo Rouge-3.

Spremenljivke	Opazovanja	Sr.vr.	Std. napak	Std. dev.	[95% stopnji zaup.]
AA/človeški(Rouge-1)	100	0.0129	0.0019	0.01955	[0.0090 - 0.0168]
LSA/človešk(Rouge-1)	100	0.0135	0.0023	0 .02382	[0.0087 - 0 .0182]
Razlike	100	0 .0005	0.0026	0.0260	[-0.0057 - 0.0046]

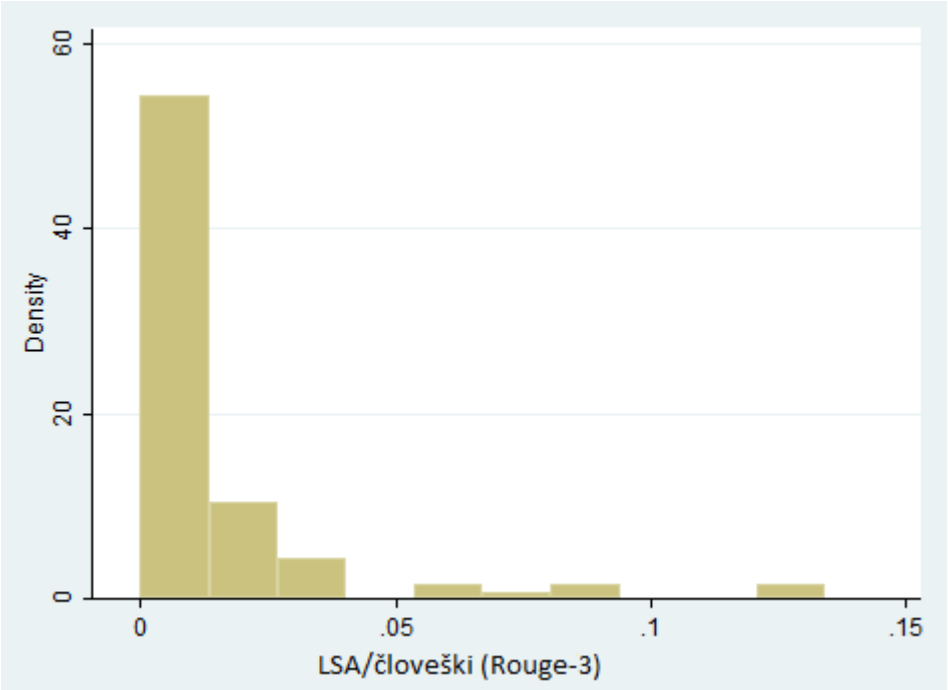
Rezultat testa: Pr(|T| > |t|) = 0,8357.

Ker je P-vrednost>0,05 ne moremo zavrniti ničelne hipoteze. Iz tega sledi, da povzemanji AA/človeški in LSA/človeški statistično nista različni pri Rouge-3. Kot v prejšnjem primeru sledi, da se kvaliteta povzemanja obeh približa, ko zaostrimo pogoje ocenjevanja podobnosti v primerjavi s človeškimi povzetki.

Na **Sliki 19** je prikazan histogram porazdelitve vrednosti ocene podobnosti povzemanja z arhetipi in človeškim povzemanjem z uporabo Rouge-3. Podobno je na **Sliki 20** prikazan histogram porazdelitve vrednosti ocene podobnosti semantičnega in človeškega povzemanja z uporabo Rouge-3. Histograma sta si po obliki podobna.



Slika 19: Histogram ocen podobnosti povzemanja z arhetipi in človeškim povzemanjem z uporabo Rouge-3. Na abcisni osi grafa so zapisane ocene podobnosti (F-mera) med povzemanji z arhetipi in človeškimi povzetki. Ocene so grupirane v 10 skupin. Na ordinatni osi je zavedeno število primerov ocen podobnosti (F-mera), ki pripadajo eni od 10 skupin. Graf je izdelan na podlagi **Priloge 4**.



Slika 20: Histogram ocen podobnosti povzemanja semantičnega in človeškega povzemanja z uporabo Rouge-3. Na abcisni osi grafa so zapisane ocene podobnosti (F-mera) med semantičnim povzemanjem in človeškimi povzetki. Ocene so grupirane v 10 skupin. Na ordinatni osi je zavedeno število primerov ocen podobnosti (F-mera), ki pripadajo eni od 10 skupin. Graf je izdelan na podlagi **Priloge 7**.

V **Tabeli 7** so prikazane povprečne vrednosti in standardni odkloni F-mere pri uporabi vseh treh n-gramov za ocenjevanje podobnosti povzetka z arhetipi in človeškimi povzetki. Podobno so v **Tabeli 8** in v **Tabeli 9** prikazani padci podobnosti pri semantičnem povzemanju v primerjavi s človeškimi povzetki ter pri povzemanju z arhetipi in semantičnem povzemanjem.

AA/človeški	povpr. vr. (F-mera)	st.dev. (F-mera)
Rouge-1	0,1918	0,0868
Rouge-2	0,0387	0,0380
Rouge-3	0,0130	0,0194

Tabela 7: Primerjava kvalitete povzetkov z arhetipi in ročnih povzetkov.

LSA/človeški	povpr. vr. (F-mera)	st.dev. (F-mera)
Rouge-1	0,1636	0,0715
Rouge-2	0,0366	0,0334
Rouge-3	0,0135	0,0237

Tabela 8: Primerjava kvalitete ročnih povzetkov s semantičnim povzemanjem.

AA/LSA	povpr. vr. (F-mera)	st.dev. (F-mera)
Rouge-1	0,2754	0,0915
Rouge-2	0,0961	0,0793
Rouge-3	0,0618	0,0727

Tabela 9: Primerjava kvalitete semantičnega povzemanja z povzemanjem z arhetipi.

Pri 95% intervalu zaupanja smo izvedli tri ločene parne T-test, kjer smo preverili ničelno hipotezo, da ni statistične značilne razlike med ocenami povzemanj AA/človeški pri kombinaciji parov n-gramov (1,2), (2,3) in (1,3). Pri vseh treh smo dobili P-vrednost<0,05 in s tem potrdili, da se kvaliteta povzemanja statistično poslabša, če pri ocenjevanju uspešnosti povzemanja namesto unigramov uporabimo bigrame ali trigrame.

POGLAVJE 6

6. Zaključek

V magistrski nalogi smo pregledali področje raziskovanja povzemanja naravnih besedil in tako določili smernice za izdelavo programa SimpleX za povzemanje naravnih besedil s pomočjo semantične analize. Izbrani sta bili metodologiji latentne semantične analize in leksikalnega veriženja. Pregledali smo obstoječe knjižnice in digitalne vire za podporo pri procesiranju naravnih jezikov. Izbrani sta bili knjižnica CoreNLP Univerze Stanford za predprocesiranje in knjižnica TML za izvajanje algoritma LSA. Uporabljeni slovarji so bili sloWNet, slovenski leksikon, slovenski morfološki slovar in oblikoskladenjske specifikacije JOS. Razvili smo samostojen spletni program, ki vključuje semantično povzemanje in vključuje ekstraktivno povzemanje s pomočjo analize arhetipov. Ocenili smo oba načina povzemanja in ju primerjali s človeškimi povzetki ter med seboj. Prvo hipotezo lahko potrdimo, ker smo s parnim T-testom uspeli dokazati, da obstaja statistično značilna razlika med povzemanjem z arhetipi in semantičnim povzemanjem. Drugo hipotezo smo preverjali preko trenda ujemanja podobnosti med sistemskimi povzetki in ročnimi povzetki medtem, ko povečujemo dolžino n-gramov. Pri povzemanju z analizo arhetipov v primerjavi s človeškimi povzetki smo pokazali, da obstaja statistično značilna razlika med ocenami, ko povečujemo n-grame in na ta način iščemo tesno povezanost sistemskega in človeškega povzetka. Na podlagi rezultata parnega T-testa med različnimi n-grami tako ne moremo potrditi druge hipoteze.

V nalogi smo izpostavili težave, povezane z dvoumno lematizacijo in dvoumnimi oblikoskladenjskimi zapisi. V tem pogledu predlagamo uporabo programa PosTaggerTrain za označevanje besedila [63]. Poleg boljšega označevanja besedila

predlagamo uporabo anotacije predstavitve pomena (angl. abstract meaning representation). Anotacija omogoča raziskovanje razmerij med besedami, ki so v relaciji, in določajo pomen. Anotacija AMR zagotavlja, da vse različice in kombinacije zgeneriranih stavkov ohranjajo isti pomen. Z drugače generiranimi stavki bi se lahko bolj približali človeškim povzetkom.

Povzetek strokovnih prispevkov magistrskega dela je naslednji:

- celovit opis področja povzemanja naravnih jezikov;
- statistično vrednotenje načina povzemanja akademskega dela dr. Ercana Canhasija v primerjavi z lastno predstavitvijo povzemanja s pomočjo semantike in digitalnih virov na življenjskem vzorcu dokumentov;
- implementacija integracijskega ogrodja za lažje raziskovanje naravnih jezikov;
- predstavitev izzivov pri raziskovanju in semantični obravnavi naravnih besedil napisanih v slovenskem jeziku..

7. Literatura in internetni viri

[1] Hans P. Luhn. The Automatic Creation of Literature Abstracts. In IBM Journal of Research Development, vol. 2, no. 2, pp. 159-165. 1958.

[2] H. P. Edmundson. New methods in automatic extracting. In Journal of the ACM, vol. 16, no. 2, pp. 264-285. 1969.

[3] Cue word or phrase. Available at: http://grammar.about.com/od/c/g/Cue-Word-Or-Phrase.htm. Accessed 11 August, 2016.

[4] Phyllis B. Baxendale: Man-made Index for Technical Literture – an experiment. In IBM Journal of Research Development, vol. 2, no. 4, pp. 354-361. 1958.

[5] Gerard Salton: Automatic text processing. Addison-Wesley Publishing Company. 1988.

[6] Josef Steinberger. Text Summarization within the LSA Framework. Doctoral Thesis. Pilsen, 2007.

[7] Christopher D. Manning, Hinrich Schutze. Foundations of Statistical Natural Language Processing. The MIT Press. Second printing. 1999.

[8] Christian Borgelt, Matthias Steinbrecher, Rudolf Kruse. Graphical Models: Representations for Learning, Reasoning and Data Mining, Second Edition. Wiley. 2009.

[9] Basic Language Structures [Shippensburg University web site]. Available at: http://webspace.ship.edu/cgboer/basiclangstruct.html. Accessed January 11, 2016.

[10] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh. A Comprehensive Survey on Text Summarization Systems. IEEE. 2009.

[11] W.T. Visser, M.B. Wieling. Sentence-based Summarization of Scientific Documents. University of Groningen. 2005.

[12] Jiang-Liang Hou, Yong-Jhih Chen. Development and Application of Optimization Model for Customized Text Summarization. Proceeding of the 17th IEEE. 2013.

[13] Zoubin Ghahramani. An Introduction to Hidden Markov Models and Bayesian Networks. International Journal of Pattern Recognition and Artifical Intelligence. 2001.

[14] Lili Kotlerman, Ido Dagan, Maya Gorodetsky, Ezra Daya. Sentence Clustering via Projection over Term Clusters. First Joint Conference on Lexical and Computational Semantics. 2012.

[15] Vipul Dalal, Dr. Latesh Malik. A Survey of Extractive and Abstractive Automatic Text Summarization Techniques. 6th International Conference on Emerging Trends in Engineering and Technology. 2013.

[16] Rhetorical Structure Theory [Wikipedia3 web site]. Available at: https://en.wikipedia.org/wiki/Rhetorical_Structure_Theory. Accessed 11 January, 2016.

[17] Intro to Rhetorical Structure Theory [Simon Fraser University]. Available at: http://www.sfu.ca/rst/01intro/intro.html. Accessed 11 January, 2016.

[18] William C. Mann, Sandra A. Thompson. Rhetorical Structure Theory: A Theory Of Text Organization. ISI/RS, pp. 87-190. 1987.

[19] Stop Words. Available at: https://en.wikipedia.org/wiki/Stop_words. Accessed 9 August, 2016.

[20] Most Common Words in English. Available at: https://en.wikipedia.org/wiki/Most_common_words_in_English. Accessed 9 August, 2016.

[21] Rasha Mohammed Badry, Ahmed Sharaf Eldin, Doaa Saad Elzanfally. Text Summarization within the Latent Semantic Analysis Framework: Comparitive Study. International Journal of Computer Applications, vol. 81, no. 11. 2013.

[22] Polysemy (words and meanings). Available at http://grammar.about.com/od/pq/g/polysemyterm.htm. Accessed 10 August, 2016.

[23] Tf-Idf. Available at: https://en.wikipedia.org/wiki/Tf%E2%80%93idf. Accessed at 11 August, 2016.

[24] Jon M. Kleinberg. Authoritative sources in a hyper-linked environment. In Journal of the ACM, vol. 46, no. 5, pp. 604-632. 1999.

[25] R. Mihalcea and P. Tarau. Text-rand bringing order into texts. In Proceeding of the Conference in Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004.

[26] R. Mihalcea, P. Tarau. An Algorithm for Language Independent Single and Multiple Document Summarization. In Proceedings of the International Joint Conference on Natural Language Processing. Korea, 2005.

[27] Victoria McCargar. Statistical Approaches to Automatic Text Summarization. Bul. Am. Soc. Info. Sci. Tech., vol. 30, pp. 21–25. 2004.

[28] Hongyan Jing: Sentence Reduction for Automatic Text Summarization. In Proceedings of the 6th Applied Natural Language Processing Conference, pp. 310-315. Seattle, USA, 2000.

[29] Kevin Knight, Daniel Marcu. Statistics-Based Summarization – Step one: Sentence Compression. In Proceeding of The 17th National Conference of the American Association for Artificial Intelligence, pp. 703-710. Austin, USA, 2000.

[30] Caroline Sporleder, Mirella Lapata. Discourse chunking and its application to sentence compression. In Proceeedings of HLT/EMNLP, pp. 257-264. Vancouver, Canada, 2005.

[31] Josef Steinberger, Karel Ježek. Sentence Compression for the LSA-based Summarizer. In Proceedings of the 7th International Conference on Information Systems Implementation and Modelling, pp. 141-148. Prerov, Czech Republic, 2006.

[32] Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, Eleazar Eskin. Towards Multidocument Summarization by Reformulation: Progress and Prospects. In Proceedings of AAAI-99, pp. 453-460. Orlando, USA, 1999.

[33] Hongyan Jing, Kathleen McKeown. Cut and Paste Based Text Summariation. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computional Linguistics, pp. 178-185. Seattle, USA, 2000.

[34] Kenji Ono, Kazuo Sumita, Seiji Miike. Abstract generation based on rhetorical structure extraction. In Proceedings of the 15 th International Conference on Computional Linguistics, volume 1, pp. 344-384. Kyoto, Japan, 1994.

[35] Anaphora (linguistics). Available at https://en.wikipedia.org/wiki/Anaphora_ (linguistics). Accessed 12 August, 2016.

[36] Regina Barzilay, Michael Elhadad. Using Lexical Chains for Text Summarization. In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp. 10-17. Madrid, Spain, 1997.

[37] Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Celebi, Danyu Liu, Elliott Drabek. Evaluation Challanges in Large-scale Document Summarization. In Proceeding of the 41st meeting of the Association for Computional Linguistics, pp 375-382. Sapporo, Japan, 2003.

[38] Adele Cutler, Leo Breiman. Archetypal Analysis. Technometrics, vol. 36, no. 4, pp.

338-347. November, 1994.

[39] Manuel J. A. Eugster and Friedrich Leisch. Weighted and Robust Archetypal Analysis Computational Statistics and Data Analysis. Computational Statistics and Data Analysis, vol. 55, no. 3, pp. 1215-1225. 2011.

[40] Giovanni C. Porzio, Giancarlo Ragozini, Domenico Vistocco. On the use of archetypes as benchmarks. Applied Stochastic Models in Business and Industry, vol. 24, no. 5, pp. 419–437. 2008.

[41] Ben H. P. Chan, Daniel A. Mitchell, Lawrence E. Cram. Archetypal analysis of galaxy spectra. Monthly Notice of the Royal Astronomical Society, vol. 338, pp. 790 -795. 2003.

[42] Christian Bauckhage, Christian Thurau. Making archetypal analysis practical. In Proceedings of the 31st DAGM Symposium on Pattern Recognition, pp. 272–281, 2009.

[43] Stephan Morgenthaler. A survey of robust statistics. Statistical Methods and Applications, vol. 15, no. 3, pp. 271 – 293. 2007.

[44] Chin-Yew Lin, Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of HLT-NAACL. Edmonton, Canada, 2003.

[45] Ani Nenkova, Rebecca Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In Document Understanding Conference. Vancouver, Canada, 2005.

[46] Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Celebi, Danyu Liu, Elliot Drabek: Evaluation Challenges in Large-scale Document Summarization. In Proceeding of the 41st meeting of the Association for Computional Linguistics, pp. 375-382. Sapporo, Japan, 2003.

[47] Andrew Morris, George Kasper, Dennis Adams. The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance. In Information Systems Research, vol. 3, no. 1, pp. 17-35. 1992.

[48] Gong Y. and Liu X.. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. Proceedings of SIGIR'01. 2001.

[49] Ercan Canhasi. Graph-based models for multi-document summarization. Doctoral dissertation. The Faculty of Computer and Information Science at the University of Ljubljana. 2014

[50] Java ROUGE Implementation. Available at https://bitbucket.org/nocgod/jrouge/wiki/Home. Accessed 10 February 2016.

[51] Slovene Lexicon. Available at http://oznacevalnik.slovenscina.eu/Media/LemmatizerTrain/SloveneLexicon.txt.zip. Accessed 29 May, 2016.

[52] Dobrovoljc, Kaja; Krek, Simon; Holozan, Peter; Erjavec, Tomaž and Romih, Miro, 2015, Morphological lexicon Sloleks 1.2, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1039.

[53] Oblikoskladenjske specifikacije JOS. Available at http://nl.ijs.si/jos/josMSD-sl.html. Accessed 27 May 2016.

[54] Fišer, Darja, 2015, Semantic lexicon of Slovene sloWNet 3.1, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1026.

[55] The Stanford NLP Group. Available at http://nlp.stanford.edu/. Accessed 30 May 2016.

[56] A List of Summarization Projects. Avaliable at http://web.science.mq.edu.au/~swan/summarization/projects_full.htm. Accessed 30 May 2016.

[57] NLP research at Columbia. Available at http://www1.cs.columbia.edu/nlp/index.cgi. Accessed 30 May 2016.

[58] TRESTLE project. Available at http://nlp.shef.ac.uk/trestle/. Accessed 30 May 2016.

[59] SweSum. Avaliable at http://swesum.nada.kth.se/index-eng.html. Accessed 30 May 2016.

[60] NL group of University of Southern California. Avaliable at http://nlg.isi.edu/research/. Accessed 30 May 2016.

[61] Text To Matrix Generator. Available at http://scgroup20.ceid.upatras.gr:8000/tmg/. Accessed 5 February 2016.

[62] N-gram. Available at https://en.wikipedia.org/wiki/N-gram. Accessed 25 August 2016.

[63] PosTaggerTrain. Available at http://source.ijs.si/mgrcar/obeliks/blob/master/README.md. Accessed 25 August 2016.

8. Priloge

Priloga 1: Seznam izvornih besedil.

Naziv dokumenta	Naslov	Avtor	Fakulteta
001-kas-1110000-149	STRES NA DELOVNEM MESTU V PODJETJIH JAVNEGA SEKTORJA IN V ZASEBNEM SEKTORJU	Tadej Hat	UNIVERZA MARIBOR EKONOMSKO-POSLOVNA FAKULTETA
002-kas-16000-150	STATISTIČNA ANALIZA SPREMEMB STRUKTURE VPISNE POPULACIJE VISOKOŠOLSKIH USTANOV	Laura Urek	UNIVERZA V LJUBLJANI FAKULTETA ZA UPRAVO
003-kas-837001-151	CENTRALNI SISTEM ZA SHRANJEVANJE DOKUMENTOV V DRUŽBI MOBITEL D.D.	Ana Dobrecović	UNIVERZA V MARIBORU FAKULTETA ZA ORGANIZACIJSKE VEDE
004-kas-8381001-152	GEODETSKE MERITVE IN ČASOVNO PLANIRANJE MONITORINGA PRI GRADNJI GARAŽNE HIŠE	Marko Žižek	UNIVERZA V MARIBORU FAKULTETA ZA GRADBENIŠTVO EKONOMSKO-POSLOVNA FAKULTETA
005-kas-8418001-153	ANALIZA CILJNIH SKUPIN KUPCEV GORENJKE	Igor Milić	UNIVERZA V MARIBORU FAKULTETA ZA ORGANIZACIJSKE VEDE
006-kas-38001-154	FIZIKALNO MODELIRANJE MENJALNIKA Z DVOJNO SKLOPKO	Nebojša Ilić	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
007-kas-8810002-155	VPLIV TEMPERATURE SUBSTRATA NA PLENILSKO VEDENJE LIČINKE VOLKCA EUROLEON NOSTRAS	Gregor Hauptman	UNIVERZA V MARIBORU FAKULTETA ZA NARAVOSLOVJE IN MATEMATIKO
008-kas-38003-156	BIVALIŠČA V PRIMERU NARAVNIH KATASTROF: ZASNOVA MANJŠEGA MOBILNEGA OBJEKTA	Tadej Junger	UNIVERZA V MARIBORU FAKULTETA ZA GRADBENIŠTVO
009-kas-7003-157	PRIDOBIVANJE OBVEŠČEVALNIH PODATKOV IZ JAVNIH VIROV	Matej Domajnko	UNIVERZA V MARIBORU FAKULTETA ZA VARNOSTNE VEDE
010-kas-8384003-158	MLADI IN NJIHOV ODNOS DO NASILJA NAD STAREJŠIMI	Sergeja Keblič	UNIVERZA V MARIBORU FAKULTETA ZA ZDRAVSTVENE VEDE

011-kas-8381004-159	MODELIRANJE VEČNAMENSKE TORBE	Katarina Kaiser	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
012-kas-8386004-160	NAČRTOVANJE VELIKIH ŠPORTNIH DOGODKOV - S KAKŠNIMI VARNOSTNIMI IZZIVI SE SOOČAJO ORGANIZATORJI?	Tjaša Corn	UNIVERZA V MARIBORU FAKULTETA ZA VARNOSTNE VEDE
013-kas-8905005-161	MULTIFUNKCIONALNA VRHNJA OBLAČILA	Vesna Uršič	UNIVERZA V LJUBLJANI NARAVOSLOVNOTEHNIŠKA FAKULTETA
014-kas-8386005-162	NALOGE IN PRISTOJNOSTI IZVRŠITELJA OPOROKE	Ksenija Šalamon	UNIVERZA V MARIBORU PRAVNA FAKULTETA
015-kas-5005-163	UPORABA RASTNIH REGULATORJEV PRI PRIDELAVI MARELIC	Deja Mlakar	UNIVERZA V MARIBORU FAKULTETA ZA KMETIJSTVO IN BIOSISTEMSKE VEDE VINOGRADNIŠTVO, VINARSTVO IN SADJARSTVO
016-kas-1130005-164	TEŽAVA REVITALZACIJE MESTNIH SREDIŠČ: PRIMER MESTA KRANJA	Manca Štefe	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
017-kas-1129005-165	(NE)POROČANJE RTV O FAŠISTIČNIH DOGODKIH V SLOVENIJI	Katarina Mikulić	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
018-kas-844006-166	DELNE PRAZNITVE V ELEMENTIH STIKALNIH NAPRAV	Barbara Damjan	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
019-kas-838006-167	ANALIZA PROCESOV IZDELAVE KOVINSKIH OBELEŽIJ	Miran Skutnik	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
020-kas-5006-168	PREDELAVA MESA KOT DOPOLNILNA DEJAVNOST NA KMETIJI – OD IDEJE DO REGISTRACIJE DEJAVNOSTI	Anka Pregl	UNIVERZA V MARIBORU FAKULTETA ZA KMETIJSTVO IN BIOSISTEMSKE VEDE ŽIVINOREJA
021-kas-8535007-169	IDENTIFIKACIJA PARAMETROV ENOSMERNEGA MOTORJA ZA NAMENE VODENJA	Enida Suljić	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
022-kas-8386007-170	LOKALIZACIJA AGENTA Z ALGORITMOM MONTE CARLO	Andrej Dolenc	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO

023-kas-8382007-171	GEOMETRIJSKA PARAMETRIZACIJA PAHA STISKALNICE	Alen Uršnik	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
024-kas-37007-172	VLOGA ANGLEŠČINE KOT OSREDNJEGA JEZIKA EU	David Pajk	UNIVERZA NA PRIMORSKEM FAKULTETA ZA MANAGEMENT KOPER
025-kas-1129007-173	KULT GRDEGA SKOZI OBJEKTIV DIANE ARBUS	Zorica Lazić	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
026-kas-848008-174	IZKORIŠČANJE TOPLOTNEGA POTENCIALA IZPUŠNIH PLINOV S STIRLINGOVIM MOTORJEM	Edis Vehabović	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
027-kas-842008-175	UPORABA NEVRONSKEGA OMREŽJA V ELEKTROENERGETSKEM SISTEMU DIPLOMSKA NALOGA	Miha Krajnc	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
028-kas-8383008-176	PRIMERJAVA ODPRTOKODNIH SIS ZA VODENJE ELEKTRONSKIH ZAPISOV O PACIENTIH	Klemen Žnidar	UNIVERZA V MARIBORU FAKULTETA ZA ORGANIZACIJSKE VEDE
029-kas-38008-177	ELEKTRIČNI VLAČILEC ZA ŠPORTNA LETALA	Andrej Dolinar	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
030-kas-15008-178	HIDRAVLIČNA RAZISKAVA GRADBENE JAME HE BREŽICE NA FIZIČNEM MODELU	Martin Bombač	UNIVERZA V LJUBLJANI FAKULTETA ZA GRADBENIŠTVO IN GEODEZIJO
031-kas-1128008-179	RAELJANSKO GIBANJE V SLOVENIJI	Vesna Prijatelj	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
032-kas-848009-180	OPTIMIRANJE TESNOSTI SEKLJALNIKA MUM4 Z METODOLOGIJO 6 SIGMA	Franc Retko	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
033-kas-845009-181	ALGORITMI ZA ISKANJE POTI V RAČUNALNIŠKIH IGRAH	Sašo Bašič	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
034-kas-841009-182	MODNO NAČRTOVANJE V CINEMA 4D	Martina Čeh Ambruš	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO

035-kas-8382009-183	UPORABA ORODJA UNITY3D ZA IZGRADNJO RAČUNALNIŠKE IGRE	Tomaž Kočevar	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
036-kas-5009-184	PARAMETRI REPRODUKCIJE IN PORODA PRI KRAVAH DOJILJAH LISASTE PASME	Majda Štelcer	UNIVERZA V MARIBORU FAKULTETA ZA KMETIJSTVO IN BIOSISTEMSKE VEDE ŽIVINOREJA
037-kas-1129009-185	RESNIČNOSTNI ŠOV KMETIJA: PRODUKCIJSKI VIDIKI	Ana Grobelnik	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
038-kas-1128009-186	SOCIOLOŠKI VIDIKI NOVIH REPRODUKTIVNIH TEHNOLOGIJ	Eva Otoničar	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
039-kas-848010-187	PROGRAMIRANJE IZDELAVE AVTOMOBILSKEGA DISTANČNIKA NA CNC STROJU	Metod Pečoler	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
040-kas-842010-188	PREDLOG MARKETINŠKEGA NAČRTA ZA PROJEKT "DRAVA KOT PRILOŽNOST"	Petra Langbauer	UNIVERZA V MARIBORU EKONOMSKO-POSLOVNA FAKULTETA
041-kas-8386010-189	ASTROFOTOGRAFIJA	Boštjan Selinšek	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
042-kas-8385010-190	INFORMACIJSKA PODPORA LOGISTIKI POŠTE 3102 CELJE	Bane Nikolić	UNIVERZA V MARIBORU FAKULTETA ZA LOGISTIKO
043-kas-1135010-191	JEZIKOVNA POLITIKA SLOVENSKEGA VISOKEGA ŠOLSTVA: STALIŠČA DO RABE JEZIKOV IN PRIPOROČILA ZA NJENO UREJANJE	Monika Kalin Golob	UNIVERZA V LJUBLJANI FILOZOFSKA FAKULTETA
044-kas-848011-192	VODENJE CNC REZKALNEGA STROJA Z UPORABO G-KODE	Rok Kovše	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
045-kas-8383011-193	SANACIJA OGRODJA PRIJEMALA MANIPULATORJA ALUMINIJASTIH BRAM	Tilen Štefane	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
046-kas-8382011-194	DIMENZIONIRANJE NOSILCA IZPUŠNEGA SISTEMA MOTORNEGA KOLESA	Gregor Rešek	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO

047-kas-5011-195	UGOTAVLJANJE PRETOKA VODE V ROZGAH PODLAG VINSKE TRTE V ČASU MIROVANJA IN SOLZENJA	Zvonka Rabič	UNIVERZA V MARIBORU FAKULTETA ZA KMETIJSTVO IN BIOSISTEMSKE VEDE VINOGRADNIŠTVO, VINARSTVO IN SADJARSTVO
048-kas-4011-196	VPLIV STOPNJE ZRELOSTI NA VSEBNOST SKUPNIH FENOLOV V IZBRANIH SORTAH SLIV	Katja Lukač	UNIVERZA V MARIBORU FAKULTETA ZA KMETIJSTVO IN BIOSISTEMSKE VEDE
049-kas-1152011-197	INFRASTRUKTURA SLOVENŠČINE	Kozma Ahačič	UNIVERZA V LJUBLJANI FILOZOFSKA FAKULTETA
050-kas-1130011-198	KORPORACIJE IN NJIHOVO UPRAVLJANJE: PRIMERJAVA ZDA IN EU	Teja Bedrač	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
051-kas-8905012-199	ODDELEK ZA TEKSTILSTVO, GRAFIKO IN OBLIKOVANJE ZGIBANJE KAŠIRANE LEPENKE ZA 180˚	Tomislav Peperko	UNIVERZA V LJUBLJANI NARAVOSLOVNOTEHNIŠKA FAKULTETA
052-kas-1152012-200	PROBLEMI DIGITALIZACIJE SLOVENSKE LITERARNE IN KULTURNE DEDIŠČINE	Marijan Dović	UNIVERZA V LJUBLJANI
053-kas-1144012-369	PRODAJA KMETIJSKIH ZEMLJIŠČ	Barbara Bališ	UNIVERZA V MARIBORU PRAVNA FAKULTETA
054-kas-8905013-202	IZDELAVA FOTOGRAFIJ IN GRAFIČNO OBLIKOVANJE GLASBENEGA KOMPAKTNEGA DISKA	Magdalena Žnidar Jermen	UNIVERZA V LJUBLJANI NARAVOSLOVNOTEHNIŠKA FAKULTETA
055-kas-848013-203	DRUŽBENI AKTIVIZEM NA SPLETNIH SOCIALNIH OMREŽJIH V DRŽAVAH ARABSKEGA SVETA	Alen Cvišić	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
056-kas-844013-204	VPLIV GORIV NA KARAKTERISTIKE SODOBNEGA DIZELSKEGA MOTORJA	Branko Grah	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
057-kas-8384013-205	ANALIZA PROMETNEGA ŠTEVCA NA KOROŠKI CESTI V MARIBORU	Jernej Titan	UNIVERZA V MARIBORU FAKULTETA ZA GRADBENIŠTVO
058-kas-8383013-206	RAZVOJ RAČUNALNIŠKIH IGER S POMOČJO GRAFIČNEGA POGONA UNITY	Teo Babič	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
059-kas-5013-207	HIDROPONSKA PRIDELAVA JAGOD DIPLOMSKO DELO	Mateja Košič	UNIVERZA V MARIBORU FAKULTETA ZA KMETIJSTVO IN BIOSISTEMSKE VEDE

060-kas-1130013-208	VLOGA NARKO KARTELOV V MEHIŠKI DRUŽBI	Jošt Kralj	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
061-kas-1129013-209	DARILNOSTNA EKONOMIJA KOT NOVA DRUŽBENA PARADIGMA?	Igor Šiška	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
062-kas-9014-210	SMERNICE ZA PRIPRAVO MEDICINSKE OPREME OB MNOŽIČNIH NESREČAH	Jernej Kocbek	UNIVERZA V MARIBORU FAKULTETA ZA ZDRAVSTVENE VEDE
063-kas-8905014-211	NAČRTOVANJE IN IZDELAVA EMBALAŽE ZA POSLOVNA DARILA	Nina Rudolf	UNIVERZA V LJUBLJANI NARAVOSLOVNOTEHNIŠKA FAKULTETA
064-kas-852014-212	ANALIZA ENERGETSKE UČINKOVITOSTI POSLOVNE STAVBE KOSTAK, D. D.	Tina Zorko	UNIVERZA V MARIBORU FAKULTETA ZA ENERGETIKO
065-kas-38014-213	PRIMERJAVA REZULTATOV ROČNEGA IN AVTOMATIZIRANEGA NAČINA IZBIRE PARAMETROV STROJNEGA VEZENJA	Janja ŠPEGLIČ	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
066-kas-1152014-214	DRUGAČEN POGLED NA SLOVARSKE DEFINICIJE: OPISATI, POJASNITI, RAZLOŽITI?	Polona Gantar	UNIVERZA V LJUBLJANI FILOZOFSKA FAKULTETA
067-kas-1118014-368	ZAPOSLOVANJE IN POKLICNA REHABILITACIJA INVALIDOV	Gregor Srdarev	UNIVERZA V MARIBORU FAKULTETA ZA ORGANIZACIJSKE VEDE
068-kas-8905015-216	GLASBENI VIDEOSPOT	Irena Košir	UNIVERZA V LJUBLJANI NARAVOSLOVNOTEHNIŠKA FAKULTETA
069-kas-8384015-217	PREVERJANJE PRAVILNE UPORABE TEHNOLOGIJE 12-KANALNEGA SNEMANJA EKG SIGNALA V KLINIČNEM OKOLJU	Nina Jaušovec	UNIVERZA V MARIBORU FAKULTETA ZA ZDRAVSTVENE VEDE
070-kas-37015-218	RAZVOJ MODELA: OD UPORABNIKA SPODBUJENO INOVIRANJE	Mitja Ruzzier	UNIVERZA NA PRIMORSKEM FAKULTETA ZA MANAGEMENT KOPER
071-kas-1152015-219	ODVISNOSTNO POVRŠINSKOSKLADENJSKO OZNAČEVANJE SLOVENŠČINE: SPECIFIKACIJE IN OZNAČENI KORPUSI	Nina Ledinek	INSTITUT JOŽEF STEFAN LJUBLJANA
072-kas-848016-220	NUMERIČNA SIMULACIJA KROŽNE IN PARABOLIČNE IZVEDBE SONČNEGA KOLEKTORJA	Matjaž Gajšek	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
073-kas-8420016-221	VLOGA IZVEDENCA ZA RAČUNOVODSTVO V KAZENSKIH POSTOPKIH	Brigita Golnar	UNIVERZA V MARIBORU EKONOMSKO-POSLOVNA FAKULTETA

074-kas-8408016-222	POLICIJSKA POOBLASTILA IN VARNOST ČLOVEKOVIH PRAVIC PRI KONTROLI PROMETA	Tjaša Magdič	UNIVERZA V MARIBORU FAKULTETA ZA VARNOSTNE VEDE
075-kas-8383016-223	ALGORITMI ZA REŠEVANJE KOSOVNE SESTAVLJANKE	Andrej Prajndl	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
076-kas-8382016-224	IZDELAVA POKROVA SKLOPA KONDENZATORJA	Jure Lorenčič	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
077-kas-1152016-225	SLOVENSKO ZGODOVINSKO SLOVAROPISJE S KONCEPTUALNO-RAZVOJNEGA VIDIKA	Majda Merše	INŠTITUT ZA SLOVENSKI JEZIK FRANA RAMOVŠA
078-kas-1110016-226	PRIMERJAVA POTNIŠKEGA ZRAČNEGA IN CESTNEGA TRANSPORTA NA DOLOČENI RELACIJI	Katja Kraner	UNIVERZA V MARIBORU EKONOMSKO-POSLOVNA FAKULTETA MARIBOR
079-kas-848017-227	DOLOČITEV PROCESNIH ZNAČILNOSTI AERACIJSKEGA MEŠALA	Simon Urbas	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
080-kas-8384017-228	PRERAČUN ZOBNIŠKEGA GONILA MEŠALNEGA VENTILA	Žan Pfifer	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
081-kas-8017-229	PREHOD NA IP TELEFONIJO V EKONOMSKO-TRGOVSKI ŠOLI	Gregor Belovič	UNIVERZA V MARIBORU FAKULTETA ZA ORGANIZACIJSKE VEDE
082-kas-4017-230	VPLIV RAZLIČNIH NAČINOV OSNOVNE OBDELAVE TAL ZA SETEV KORUZE NA PORABO GORIVA IN ZDRS	Peter Šbul	UNIVERZA V MARIBORU FAKULTETA ZA KMETIJSTVO IN BIOSISTEMSKE VEDE BIOSISTEMSKO INŽENIRSTVO
083-kas-1152017-231	REPOZITORIJ ROKOPISOV SLOVENSKEGA SLOVSTVA – ODDALJENA BLIŽINA?	Matija Ogrin	INŠTITUTZA SLOVENSKO LITERATURO IN LITERARNE VEDE
084-kas-1113017-232	NARAVA, INŽENIRJI IN NUMERIČNE SIMULACIJE	Jure Ravnik	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
085-kas-8905018-233	LIZBONA SKOZI FOTOGRAFSKI OBJEKTIV	Blaž Janez Erjavec	UNIVERZA V LJUBLJANI NARAVOSLOVNOTEHNIŠKA FAKULTETA
086-kas-8534018-234	IMPLEMENTACIJA UREJEVALNIKA MEST Z UPORABO KNJIŽNICE LIBGDX	Luka Horvat	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO

087-kas-8420018-235	INTEGRACIJA METOD MODELIRANJA IN SIMULACIJ V SISTEME SCADA	Benjamin Mastnak	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
088-kas-37018-236	ANALIZA INOVACIJSKE KLIME V ZASEBNEM SEKTORJU V SLOVENIJI	Cene Bavec	UNIVERZA NA PRIMORSKEM FAKULTETA ZA MANAGEMENT KOPER
089-kas-8534019-237	STISKANJE TRIKOTNIŠKIH MREŽ PO METODAH DE FLORIANIJEVE S SODELAVCI IN GUMHOLD-STRASSERJA	Patrik Kokol	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO
090-kas-848019-238	ELEKTRONSKI SISTEM ZA VODENJE IZMENIČNIH MOTORJEV – PROGRAMSKA OPREMA ZA MIKROKRMILNIK TMS320F28335	Darko Podržaj	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
091-kas-8383019-239	PRENOVA APLIKACIJE NUDENJA IT STORITEV	Andriana Georgieva	UNIVERZA V MARIBORU FAKULTETA ZA ORGANIZACIJSKE VEDE
092-kas-5019-240	MREŽNO PLANIRANJE NA PRIMERU IZGRADNJE KOMUNALNE KANALIZACIJE S ČISTILNO NAPRAVO	Danijela Gomboc	UNIVERZA V MARIBORU FAKULTETA ZA KMETIJSTVO IN BIOSISTEMSKE VEDE MANAGEMENT V AGROŽIVILSTVU IN RAZVOJ PODEŽELJA
093-kas-4019-241	VPLIV VEČLETNE ALTERNATIVNE OBDELAVE TAL NA ZBITOST TAL	Klemen Bedrač	UNIVERZA V MARIBORU FAKULTETA ZA KMETIJSTVO IN BIOSISTEMSKE VEDE BIOSISTEMSKO INŽENIRSTVO
094-kas-38019-242	MOBILNI OBJEKTI: ZASNOVA 'VAGONSKEGA' HOSTLA	Vedran Vugrin	UNIVERZA V MARIBORU FAKULTETA ZA GRADBENIŠTVO
095-kas-37019-243	PRIDOBIVANJE ZNANJA V SLOVENSKIH MALIH IN SREDNJE VELIKIH PODJETJIH	Gomezelj Omerzel	UNIVERZA NA PRIMORSKEM FAKULTETA ZA MANAGEMENT KOPER
096-kas-848020-244	TEKSTILIJE ZA ZVOČNO ZAŠČITO PROSTOROV	Nina Ačko	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
097-kas-8385021-245	ŽITNI KROGI	Dejan Ščuri	UNIVERZA V MARIBORU FAKULTETA ZA ELEKTROTEHNIKO, RAČUNALNIŠTVO IN INFORMATIKO

098-kas-8383021-246	STROJ ZA RAZŠIRJANJE TANKOSTENSKIH CEVI	Denis Krajnc	UNIVERZA V MARIBORU FAKULTETA ZA STROJNIŠTVO
099-kas-1128021-247	SOCIOLOŠKI POGLED NA NOVE ŽENSKE IDENTITETE: PRIMER ŽENSK, KI SE NE ODLOČIJO ZA MATERINSTVO	Evelin Trafela	UNIVERZA V LJUBLJANI FAKULTETA ZA DRUŽBENE VEDE
100-kas-8905022-248	OSVEŠČANJE PREDŠOLSKIH OTROK SKOZI PRIMERE RECIKLIRANJA PAPIRNATIH BRISAČ	Natalija Grabner	UNIVERZA V LJUBLJANI NARAVOSLOVNOTEHNIŠKA FAKULTETA

Priloga 2: Ocene primerjave povzemanja z arhetipi in človeškimi povzetki z Rouge-1.

AA/človeški (Rouge-1)	Recall	Precision	F measure
001-kas-1110000-149	0,2203	0,1831	0,1997
002-kas-16000-150	0,2857	0,0962	0,1440
003-kas-837001-151	0,2992	0,1237	0,1750
004-kas-8381001-152	0,3750	0,0982	0,1556
005-kas-8418001-153	0,4304	0,2965	0,3497
006-kas-38001-154	0,2976	0,0972	0,1465
007-kas-8810002-155	0,3899	0,3309	0,3577
008-kas-38003-156	0,3145	0,0586	0,0988
009-kas-7003-157	0,3533	0,1923	0,2490
010-kas-8384003-158	0,2954	0,2602	0,2767
011-kas-8381004-159	0,3795	0,1047	0,1641
012-kas-8386004-160	0,2741	0,2101	0,2379
013-kas-8905005-161	0,2356	0,1790	0,2032
014-kas-8386005-162	0,1607	0,1367	0,1477
015-kas-5005-163	0,4359	0,2244	0,2962
016-kas-1130005-164	0,3690	0,1753	0,2376
017-kas-1129005-165	0,4121	0,2341	0,2986
018-kas-844006-166	0,3506	0,0982	0,1534
019-kas-838006-167	0,3796	0,0714	0,1202
020-kas-5006-168	0,0588	0,0197	0,0296
021-kas-8535007-169	0,3723	0,1897	0,2513
022-kas-8386007-170	0,2667	0,0500	0,0841
023-kas-8382007-171	0,3148	0,1559	0,2085
024-kas-37007-172	0,3750	0,1536	0,2179
025-kas-1129007-173	0,3750	0,2177	0,2755
026-kas-848008-174	0,2010	0,0727	0,1068
027-kas-842008-175	0,3646	0,0568	0,0983
028-kas-8383008-176	0,0699	0,0342	0,0459

029-kas-38008-177	0,2742	0,1173	0,1642
030-kas-15008-178	0,1280	0,0930	0,1077
031-kas-1128008-179	0,3939	0,2360	0,2947
032-kas-848009-180	0,2853	0,1621	0,2067
033-kas-845009-181	0,2981	0,0877	0,1350
034-kas-841009-182	0,4211	0,1395	0,2095
035-kas-8382009-183	0,3553	0,0944	0,1491
036-kas-5009-184	0,4808	0,3073	0,3749
037-kas-1129009-185	0,3099	0,1894	0,2351
038-kas-1128009-186	0,3247	0,1739	0,2265
039-kas-848010-187	0,3556	0,1046	0,1617
040-kas-842010-188	0,3016	0,3296	0,3146
041-kas-8386010-189	0,3417	0,0714	0,1181
042-kas-8385010-190	0,2893	0,1122	0,1617
043-kas-1135010-191	0,1786	0,0344	0,0577
044-kas-848011-192	0,3384	0,1101	0,1658
045-kas-8383011-193	0,4408	0,1167	0,1845
046-kas-8382011-194	0,4683	0,0978	0,1618
047-kas-5011-195	0,4662	0,2254	0,3038
048-kas-4011-196	0,4893	0,3772	0,4260
049-kas-1152011-197	0,2447	0,0385	0,0665
050-kas-1130011-198	0,3958	0,1921	0,2572
051-kas-8905012-199	0,2350	0,1732	0,1994
052-kas-1152012-200	0,1087	0,0246	0,0401
053-kas-1144012-369	0,3037	0,2351	0,2650
054-kas-8905013-202	0,2629	0,1560	0,1957
055-kas-848013-203	0,3885	0,1789	0,2447
056-kas-844013-204	0,3462	0,1985	0,2522
057-kas-8384013-205	0,3065	0,0482	0,0833
058-kas-8383013-206	0,3202	0,1045	0,1576
059-kas-5013-207	0,3115	0,1407	0,1938
060-kas-1130013-208	0,3138	0,1955	0,2405
061-kas-1129013-209	0,3675	0,2328	0,2850
062-kas-9014-210	0,1364	0,0746	0,0965
063-kas-8905014-211	0,2982	0,2167	0,2510
064-kas-852014-212	0,3806	0,0837	0,1372
065-kas-38014-213	0,3796	0,1452	0,2100
066-kas-1152014-214	0,1582	0,1865	0,1709
067-kas-1118014-368	0,3320	0,1353	0,1923
068-kas-8905015-216	0,3390	0,1447	0,2028
069-kas-8384015-217	0,4096	0,1588	0,2288
070-kas-37015-218	0,4157	0,1145	0,1795
071-kas-1152015-219	0,4773	0,1006	0,1661
072-kas-848016-220	0,2778	0,0976	0,1445
073-kas-8420016-221	0,3883	0,2455	0,3007

074-kas-8408016-222	0,2293	0,2360	0,2326
075-kas-8383016-223	0,4274	0,0923	0,1517
076-kas-8382016-224	0,4479	0,0793	0,1347
077-kas-1152016-225	0,2705	0,0589	0,0968
078-kas-1110016-226	0,3045	0,1742	0,2216
079-kas-848017-227	0,1260	0,0426	0,0636
080-kas-8384017-228	0,5526	0,1393	0,2225
081-kas-8017-229	0,3624	0,1887	0,2482
082-kas-4017-230	0,4344	0,4100	0,4218
083-kas-1152017-231	0,3029	0,1101	0,1615
084-kas-1113017-232	0,1304	0,3492	0,1899
085-kas-8905018-233	0,0000	0,0000	0,0000
086-kas-8534018-234	0,4043	0,0658	0,1132
087-kas-8420018-235	0,3526	0,0967	0,1517
088-kas-37018-236	0,3750	0,1298	0,1929
089-kas-8534019-237	0,0000	0,0000	0,0000
090-kas-848019-238	0,1650	0,0642	0,0924
091-kas-8383019-239	0,3896	0,2312	0,2902
092-kas-5019-240	0,3104	0,2215	0,2576
093-kas-4019-241	0,4635	0,2740	0,3442
094-kas-38019-242	0,2436	0,0508	0,0835
095-kas-37019-243	0,2218	0,0844	0,1223
096-kas-848020-244	0,3583	0,2356	0,2842
097-kas-8385021-245	0,2653	0,0409	0,0707
098-kas-8383021-246	0,4274	0,1890	0,2620
099-kas-1128021-247	0,3558	0,1853	0,2434
100-kas-8905022-248	0,4137	0,2038	0,2731

Priloga 3: Ocene primerjave povzemanja z arhetipi in človeškimi povzetki z Rouge-2.

AA/človeški (Rouge-2)	Recall	Precision	F measure
001-kas-1110000-149	0,0066	0,0054	0,0060
002-kas-16000-150	0,0500	0,0168	0,0251
003-kas-837001-151	0,0357	0,0147	0,0209
004-kas-8381001-152	0,0933	0,0239	0,0381
005-kas-8418001-153	0,2047	0,1405	0,1659
006-kas-38001-154	0,0422	0,0138	0,0207
007-kas-8810002-155	0,1014	0,0858	0,0929
008-kas-38003-156	0,0656	0,0122	0,0205
009-kas-7003-157	0,0723	0,0392	0,0509
010-kas-8384003-158	0,0562	0,0495	0,0527

011-kas-8381004-159	0,0854	0,0234	0,0367
012-kas-8386004-160	0,0176	0,0135	0,0153
013-kas-8905005-161	0,0237	0,0179	0,0204
014-kas-8386005-162	0,0159	0,0136	0,0146
015-kas-5005-163	0,1290	0,0662	0,0875
016-kas-1130005-164	0,0972	0,0462	0,0626
017-kas-1129005-165	0,0396	0,0225	0,0287
018-kas-844006-166	0,1184	0,0329	0,0515
019-kas-838006-167	0,0660	0,0121	0,0205
020-kas-5006-168	0,0178	0,0059	0,0089
021-kas-8535007-169	0,0786	0,0399	0,0529
022-kas-8386007-170	0,0000	0,0000	0,0000
023-kas-8382007-171	0,0709	0,0352	0,0470
024-kas-37007-172	0,0586	0,0238	0,0339
025-kas-1129007-173	0,0440	0,0255	0,0323
026-kas-848008-174	0,0573	0,0206	0,0303
027-kas-842008-175	0,0532	0,0081	0,0141
028-kas-8383008-176	0,0037	0,0018	0,0024
029-kas-38008-177	0,0122	0,0051	0,0072
030-kas-15008-178	0,0170	0,0122	0,0142
031-kas-1128008-179	0,1011	0,0607	0,0758
032-kas-848009-180	0,0444	0,0251	0,0321
033-kas-845009-181	0,0146	0,0040	0,0063
034-kas-841009-182	0,1170	0,0378	0,0571
035-kas-8382009-183	0,0267	0,0070	0,0111
036-kas-5009-184	0,1519	0,0970	0,1184
037-kas-1129009-185	0,0265	0,0161	0,0201
038-kas-1128009-186	0,0327	0,0175	0,0228
039-kas-848010-187	0,0225	0,0066	0,0102
040-kas-842010-188	0,0470	0,0508	0,0488
041-kas-8386010-189	0,0424	0,0087	0,0144
042-kas-8385010-190	0,0542	0,0206	0,0299
043-kas-1135010-191	0,0091	0,0017	0,0028
044-kas-848011-192	0,1122	0,0362	0,0547
045-kas-8383011-193	0,1600	0,0418	0,0663
046-kas-8382011-194	0,0806	0,0166	0,0275
047-kas-5011-195	0,1837	0,0884	0,1194
048-kas-4011-196	0,2318	0,1787	0,2018
049-kas-1152011-197	0,0326	0,0052	0,0090
050-kas-1130011-198	0,0449	0,0202	0,0278
051-kas-8905012-199	0,0327	0,0237	0,0275
052-kas-1152012-200	0,0074	0,0016	0,0026
053-kas-1144012-369	0,0528	0,0408	0,0460
054-kas-8905013-202	0,0287	0,0171	0,0214
055-kas-848013-203	0,0833	0,0374	0,0515

056-kas-844013-204	0,0903	0,0516	0,0657
057-kas-8384013-205	0,0984	0,0153	0,0264
058-kas-8383013-206	0,0511	0,0165	0,0250
059-kas-5013-207	0,0426	0,0191	0,0264
060-kas-1130013-208	0,0374	0,0230	0,0284
061-kas-1129013-209	0,0636	0,0403	0,0493
062-kas-9014-210	0,0065	0,0035	0,0046
063-kas-8905014-211	0,0276	0,0201	0,0233
064-kas-852014-212	0,0758	0,0164	0,0270
065-kas-38014-213	0,0421	0,0158	0,0230
066-kas-1152014-214	0,0053	0,0063	0,0058
067-kas-1118014-368	0,0444	0,0179	0,0256
068-kas-8905015-216	0,0427	0,0181	0,0255
069-kas-8384015-217	0,0968	0,0375	0,0540
070-kas-37015-218	0,0732	0,0200	0,0314
071-kas-1152015-219	0,1462	0,0304	0,0504
072-kas-848016-220	0,0102	0,0036	0,0053
073-kas-8420016-221	0,0802	0,0513	0,0625
074-kas-8408016-222	0,0173	0,0179	0,0176
075-kas-8383016-223	0,0656	0,0139	0,0229
076-kas-8382016-224	0,1277	0,0218	0,0373
077-kas-1152016-225	0,0083	0,0018	0,0030
078-kas-1110016-226	0,0355	0,0202	0,0258
079-kas-848017-227	0,0077	0,0027	0,0040
080-kas-8384017-228	0,2333	0,0582	0,0932
081-kas-8017-229	0,0743	0,0385	0,0508
082-kas-4017-230	0,1831	0,1728	0,1778
083-kas-1152017-231	0,0874	0,0323	0,0472
084-kas-1113017-232	0,0147	0,0395	0,0215
085-kas-8905018-233	0,0000	0,0000	0,0000
086-kas-8534018-234	0,0109	0,0017	0,0030
087-kas-8420018-235	0,0390	0,0106	0,0166
088-kas-37018-236	0,0672	0,0231	0,0344
089-kas-8534019-237	0,0000	0,0000	0,0000
090-kas-848019-238	0,0000	0,0000	0,0000
091-kas-8383019-239	0,1275	0,0754	0,0947
092-kas-5019-240	0,0746	0,0524	0,0614
093-kas-4019-241	0,1963	0,1150	0,1449
094-kas-38019-242	0,0130	0,0022	0,0037
095-kas-37019-243	0,0163	0,0062	0,0089
096-kas-848020-244	0,1129	0,0745	0,0897
097-kas-8385021-245	0,0000	0,0000	0,0000
098-kas-8383021-246	0,0935	0,0411	0,0571
099-kas-1128021-247	0,0935	0,0487	0,0640
100-kas-8905022-248	0,0797	0,0391	0,0525

Priloga 4: ocene primerjave povzemanja z arhetipi in človeškimi povzetki z Rouge-3.

AA/človeški (Rouge-3)	Recall	Precision	F measure
001-kas-1110000-149	0,0000	0,0000	0,0000
002-kas-16000-150	0,0112	0,0037	0,0056
003-kas-837001-151	0,0040	0,0016	0,0023
004-kas-8381001-152	0,0676	0,0169	0,0271
005-kas-8418001-153	0,1172	0,0787	0,0938
006-kas-38001-154	0,0244	0,0080	0,0120
007-kas-8810002-155	0,0382	0,0321	0,0349
008-kas-38003-156	0,0250	0,0046	0,0078
009-kas-7003-157	0,0303	0,0164	0,0213
010-kas-8384003-158	0,0058	0,0052	0,0055
011-kas-8381004-159	0,0123	0,0034	0,0053
012-kas-8386004-160	0,0022	0,0017	0,0019
013-kas-8905005-161	0,0026	0,0019	0,0022
014-kas-8386005-162	0,0020	0,0017	0,0018
015-kas-5005-163	0,0422	0,0216	0,0285
016-kas-1130005-164	0,0420	0,0195	0,0266
017-kas-1129005-165	0,0061	0,0035	0,0044
018-kas-844006-166	0,0333	0,0092	0,0144
019-kas-838006-167	0,0096	0,0017	0,0029
020-kas-5006-168	0,0000	0,0000	0,0000
021-kas-8535007-169	0,0360	0,0181	0,0241
022-kas-8386007-170	0,0000	0,0000	0,0000
023-kas-8382007-171	0,0451	0,0223	0,0299
024-kas-37007-172	0,0091	0,0037	0,0052
025-kas-1129007-173	0,0000	0,0000	0,0000
026-kas-848008-174	0,0211	0,0075	0,0110
027-kas-842008-175	0,0109	0,0016	0,0029
028-kas-8383008-176	0,0000	0,0000	0,0000
029-kas-38008-177	0,0000	0,0000	0,0000
030-kas-15008-178	0,0024	0,0017	0,0020
031-kas-1128008-179	0,0480	0,0284	0,0356
032-kas-848009-180	0,0089	0,0050	0,0064
033-kas-845009-181	0,0049	0,0012	0,0019
034-kas-841009-182	0,0430	0,0135	0,0206
035-kas-8382009-183	0,0000	0,0000	0,0000
036-kas-5009-184	0,0417	0,0266	0,0324
037-kas-1129009-185	0,0000	0,0000	0,0000

038-kas-1128009-186	0,0066	0,0035	0,0046
039-kas-848010-187	0,0000	0,0000	0,0000
040-kas-842010-188	0,0094	0,0101	0,0098
041-kas-8386010-189	0,0086	0,0017	0,0029
042-kas-8385010-190	0,0294	0,0110	0,0161
043-kas-1135010-191	0,0000	0,0000	0,0000
044-kas-848011-192	0,0515	0,0162	0,0246
045-kas-8383011-193	0,0405	0,0105	0,0166
046-kas-8382011-194	0,0164	0,0033	0,0055
047-kas-5011-195	0,0616	0,0295	0,0399
048-kas-4011-196	0,1315	0,1014	0,1145
049-kas-1152011-197	0,0111	0,0017	0,0030
050-kas-1130011-198	0,0000	0,0000	0,0000
051-kas-8905012-199	0,0025	0,0018	0,0021
052-kas-1152012-200	0,0000	0,0000	0,0000
053-kas-1144012-369	0,0138	0,0106	0,0120
054-kas-8905013-202	0,0000	0,0000	0,0000
055-kas-848013-203	0,0182	0,0078	0,0109
056-kas-844013-204	0,0227	0,0130	0,0165
057-kas-8384013-205	0,0333	0,0051	0,0088
058-kas-8383013-206	0,0000	0,0000	0,0000
059-kas-5013-207	0,0078	0,0035	0,0048
060-kas-1130013-208	0,0027	0,0016	0,0020
061-kas-1129013-209	0,0213	0,0135	0,0165
062-kas-9014-210	0,0000	0,0000	0,0000
063-kas-8905014-211	0,0000	0,0000	0,0000
064-kas-852014-212	0,0385	0,0082	0,0135
065-kas-38014-213	0,0094	0,0036	0,0052
066-kas-1152014-214	0,0000	0,0000	0,0000
067-kas-1118014-368	0,0163	0,0065	0,0093
068-kas-8905015-216	0,0043	0,0018	0,0025
069-kas-8384015-217	0,0380	0,0146	0,0211
070-kas-37015-218	0,0123	0,0034	0,0053
071-kas-1152015-219	0,0469	0,0096	0,0160
072-kas-848016-220	0,0000	0,0000	0,0000
073-kas-8420016-221	0,0323	0,0210	0,0254
074-kas-8408016-222	0,0017	0,0018	0,0017
075-kas-8383016-223	0,0333	0,0070	0,0116
076-kas-8382016-224	0,0543	0,0090	0,0154
077-kas-1152016-225	0,0000	0,0000	0,0000
078-kas-1110016-226	0,0000	0,0000	0,0000
079-kas-848017-227	0,0039	0,0014	0,0020
080-kas-8384017-228	0,1014	0,0250	0,0402
081-kas-8017-229	0,0102	0,0053	0,0069
082-kas-4017-230	0,0847	0,0799	0,0822

083-kas-1152017-231	0,0490	0,0182	0,0265
084-kas-1113017-232	0,0019	0,0052	0,0028
085-kas-8905018-233	0,0000	0,0000	0,0000
086-kas-8534018-234	0,0000	0,0000	0,0000
087-kas-8420018-235	0,0000	0,0000	0,0000
088-kas-37018-236	0,0381	0,0131	0,0195
089-kas-8534019-237	0,0000	0,0000	0,0000
090-kas-848019-238	0,0000	0,0000	0,0000
091-kas-8383019-239	0,0461	0,0271	0,0342
092-kas-5019-240	0,0194	0,0136	0,0160
093-kas-4019-241	0,0947	0,0550	0,0696
094-kas-38019-242	0,0000	0,0000	0,0000
095-kas-37019-243	0,0041	0,0015	0,0022
096-kas-848020-244	0,0432	0,0286	0,0344
097-kas-8385021-245	0,0000	0,0000	0,0000
098-kas-8383021-246	0,0205	0,0089	0,0124
099-kas-1128021-247	0,0390	0,0202	0,0266
100-kas-8905022-248	0,0182	0,0089	0,0120

Priloga 5: ocena primerjave semantičnega povzemanja s človeškimi povzetki z Rouge-1.

LSA/človeški (Rouge-1)	Recall	Precision	F measure
001-kas-1110000-149	0,4185	0,1429	0,2130
002-kas-16000-150	0,4505	0,0851	0,1431
003-kas-837001-151	0,4094	0,0887	0,1459
004-kas-8381001-152	0,3947	0,0599	0,1040
005-kas-8418001-153	0,3247	0,1694	0,2226
006-kas-38001-154	0,3452	0,0430	0,0765
007-kas-8810002-155	0,4765	0,1897	0,2713
008-kas-38003-156	0,3871	0,0378	0,0689
009-kas-7003-157	0,3832	0,1203	0,1831
010-kas-8384003-158	0,4556	0,1852	0,2634
011-kas-8381004-159	0,4699	0,0662	0,1161
012-kas-8386004-160	0,3816	0,1300	0,1940
013-kas-8905005-161	0,3403	0,0796	0,1290
014-kas-8386005-162	0,5952	0,2415	0,3436
015-kas-5005-163	0,4872	0,1638	0,2452
016-kas-1130005-164	0,5517	0,0748	0,1318
017-kas-1129005-165	0,6788	0,0735	0,1327
018-kas-844006-166	0,5195	0,0634	0,1130

019-kas-838006-167	0,3519	0,0395	0,0710
020-kas-5006-168	0,3941	0,0929	0,1504
021-kas-8535007-169	0,7660	0,1494	0,2500
022-kas-8386007-170	0,3500	0,0343	0,0625
023-kas-8382007-171	0,5852	0,1653	0,2577
024-kas-37007-172	0,4286	0,0828	0,1387
025-kas-1129007-173	0,4813	0,0914	0,1537
026-kas-848008-174	0,3918	0,1022	0,1620
027-kas-842008-175	0,4792	0,0406	0,0748
028-kas-8383008-176	0,3971	0,0740	0,1247
029-kas-38008-177	0,4113	0,1081	0,1711
030-kas-15008-178	0,5266	0,2466	0,3359
031-kas-1128008-179	0,4972	0,1142	0,1858
032-kas-848009-180	0,3765	0,0979	0,1553
033-kas-845009-181	0,3462	0,0638	0,1078
034-kas-841009-182	0,3684	0,0812	0,1331
035-kas-8382009-183	0,4737	0,0533	0,0959
036-kas-5009-184	0,4560	0,1469	0,2222
037-kas-1129009-185	0,4444	0,0861	0,1442
038-kas-1128009-186	0,4935	0,1164	0,1884
039-kas-848010-187	0,4111	0,0772	0,1301
040-kas-842010-188	0,3156	0,2036	0,2475
041-kas-8386010-189	0,4333	0,0392	0,0718
042-kas-8385010-190	0,3884	0,0801	0,1328
043-kas-1135010-191	0,5000	0,0338	0,0633
044-kas-848011-192	0,4444	0,0667	0,1159
045-kas-8383011-193	0,5789	0,0630	0,1137
046-kas-8382011-194	0,7302	0,0847	0,1518
047-kas-5011-195	0,6554	0,1334	0,2217
048-kas-4011-196	0,5726	0,2752	0,3717
049-kas-1152011-197	0,4255	0,0425	0,0772
050-kas-1130011-198	0,6190	0,0895	0,1564
051-kas-8905012-199	0,5150	0,1329	0,2113
052-kas-1152012-200	0,1159	0,0219	0,0368
053-kas-1144012-369	0,4566	0,1704	0,2481
054-kas-8905013-202	0,3429	0,0958	0,1498
055-kas-848013-203	0,4892	0,1068	0,1753
056-kas-844013-204	0,5449	0,1560	0,2425
057-kas-8384013-205	0,6129	0,0387	0,0728
058-kas-8383013-206	0,4944	0,0421	0,0775
059-kas-5013-207	0,3846	0,0842	0,1381
060-kas-1130013-208	0,4787	0,0900	0,1515
061-kas-1129013-209	0,5361	0,1201	0,1963
062-kas-9014-210	0,4416	0,1388	0,2112
063-kas-8905014-211	0,4358	0,2097	0,2832

064-kas-852014-212	0,4627	0,0602	0,1065
065-kas-38014-213	0,3426	0,1072	0,1634
066-kas-1152014-214	0,2394	0,2326	0,2359
067-kas-1118014-368	0,4640	0,1450	0,2210
068-kas-8905015-216	0,3644	0,0925	0,1475
069-kas-8384015-217	0,6702	0,0707	0,1279
070-kas-37015-218	0,4940	0,0807	0,1387
071-kas-1152015-219	0,2879	0,0823	0,1279
072-kas-848016-220	0,4343	0,0827	0,1389
073-kas-8420016-221	0,4734	0,1606	0,2399
074-kas-8408016-222	0,4000	0,1574	0,2259
075-kas-8383016-223	0,4194	0,0520	0,0925
076-kas-8382016-224	0,3750	0,0356	0,0651
077-kas-1152016-225	0,2623	0,0780	0,1203
078-kas-1110016-226	0,4551	0,1488	0,2243
079-kas-848017-227	0,3359	0,1222	0,1792
080-kas-8384017-228	0,5658	0,0707	0,1257
081-kas-8017-229	0,5705	0,1991	0,2951
082-kas-4017-230	0,5861	0,2563	0,3566
083-kas-1152017-231	0,2788	0,0718	0,1142
084-kas-1113017-232	0,1151	0,3285	0,1705
085-kas-8905018-233	0,5035	0,1021	0,1698
086-kas-8534018-234	0,5745	0,0507	0,0931
087-kas-8420018-235	0,5128	0,0673	0,1190
088-kas-37018-236	0,5833	0,1316	0,2147
089-kas-8534019-237	0,3158	0,0223	0,0417
090-kas-848019-238	0,3981	0,0694	0,1182
091-kas-8383019-239	0,3117	0,1135	0,1664
092-kas-5019-240	0,4231	0,1392	0,2095
093-kas-4019-241	0,3073	0,1103	0,1623
094-kas-38019-242	0,2949	0,0363	0,0646
095-kas-37019-243	0,7661	0,1484	0,2487
096-kas-848020-244	0,6096	0,1650	0,2597
097-kas-8385021-245	0,4694	0,0288	0,0543
098-kas-8383021-246	0,4516	0,0995	0,1630
099-kas-1128021-247	0,5833	0,0728	0,1294
100-kas-8905022-248	0,4964	0,1255	0,2003

Priloga 6: ocena primerjave semantičnega povzemanja s človeškimi povzetki z Rouge-2.

LSA/človeški (Rouge-2)	Recall	Precision	F measure
001-kas-1110000-149	0,0531	0,0181	0,0270
002-kas-16000-150	0,0778	0,0146	0,0245
003-kas-837001-151	0,0556	0,0120	0,0197
004-kas-8381001-152	0,0400	0,0060	0,0104
005-kas-8418001-153	0,0570	0,0296	0,0390
006-kas-38001-154	0,0241	0,0030	0,0053
007-kas-8810002-155	0,0942	0,0374	0,0536
008-kas-38003-156	0,0492	0,0047	0,0086
009-kas-7003-157	0,0422	0,0132	0,0201
010-kas-8384003-158	0,0620	0,0252	0,0358
011-kas-8381004-159	0,0732	0,0102	0,0179
012-kas-8386004-160	0,0837	0,0284	0,0425
013-kas-8905005-161	0,0474	0,0110	0,0179
014-kas-8386005-162	0,1554	0,0629	0,0896
015-kas-5005-163	0,1032	0,0346	0,0518
016-kas-1130005-164	0,0833	0,0112	0,0198
017-kas-1129005-165	0,2256	0,0243	0,0439
018-kas-844006-166	0,1842	0,0222	0,0397
019-kas-838006-167	0,0566	0,0063	0,0113
020-kas-5006-168	0,0533	0,0125	0,0202
021-kas-8535007-169	0,4857	0,0942	0,1578
022-kas-8386007-170	0,0339	0,0033	0,0060
023-kas-8382007-171	0,2612	0,0734	0,1146
024-kas-37007-172	0,0541	0,0104	0,0174
025-kas-1129007-173	0,0566	0,0107	0,0180
026-kas-848008-174	0,0833	0,0216	0,0343
027-kas-842008-175	0,1064	0,0088	0,0163
028-kas-8383008-176	0,0370	0,0069	0,0116
029-kas-38008-177	0,0488	0,0127	0,0202
030-kas-15008-178	0,2136	0,0998	0,1360
031-kas-1128008-179	0,1461	0,0334	0,0544
032-kas-848009-180	0,0592	0,0153	0,0243
033-kas-845009-181	0,0485	0,0089	0,0150
034-kas-841009-182	0,0213	0,0047	0,0076
035-kas-8382009-183	0,0667	0,0074	0,0134
036-kas-5009-184	0,0663	0,0213	0,0322
037-kas-1129009-185	0,0529	0,0102	0,0171
038-kas-1128009-186	0,1176	0,0276	0,0447
039-kas-848010-187	0,0449	0,0084	0,0141

040-kas-842010-188	0,0376	0,0242	0,0295
041-kas-8386010-189	0,0169	0,0015	0,0028
042-kas-8385010-190	0,0583	0,0119	0,0198
043-kas-1135010-191	0,1273	0,0085	0,0159
044-kas-848011-192	0,0612	0,0091	0,0159
045-kas-8383011-193	0,2267	0,0244	0,0440
046-kas-8382011-194	0,3226	0,0369	0,0662
047-kas-5011-195	0,2381	0,0482	0,0802
048-kas-4011-196	0,2060	0,0988	0,1335
049-kas-1152011-197	0,0435	0,0043	0,0078
050-kas-1130011-198	0,1198	0,0172	0,0301
051-kas-8905012-199	0,1508	0,0388	0,0617
052-kas-1152012-200	0,0294	0,0055	0,0092
053-kas-1144012-369	0,1284	0,0478	0,0697
054-kas-8905013-202	0,0517	0,0144	0,0225
055-kas-848013-203	0,1232	0,0267	0,0439
056-kas-844013-204	0,1548	0,0441	0,0687
057-kas-8384013-205	0,2459	0,0153	0,0288
058-kas-8383013-206	0,1136	0,0096	0,0177
059-kas-5013-207	0,0310	0,0067	0,0111
060-kas-1130013-208	0,0642	0,0120	0,0202
061-kas-1129013-209	0,1212	0,0270	0,0442
062-kas-9014-210	0,0784	0,0245	0,0374
063-kas-8905014-211	0,0968	0,0465	0,0628
064-kas-852014-212	0,1818	0,0233	0,0414
065-kas-38014-213	0,0000	0,0000	0,0000
066-kas-1152014-214	0,0160	0,0155	0,0158
067-kas-1118014-368	0,0968	0,0301	0,0459
068-kas-8905015-216	0,0427	0,0108	0,0172
069-kas-8384015-217	0,1505	0,0157	0,0285
070-kas-37015-218	0,0854	0,0138	0,0238
071-kas-1152015-219	0,0154	0,0043	0,0068
072-kas-848016-220	0,0612	0,0116	0,0194
073-kas-8420016-221	0,1123	0,0380	0,0568
074-kas-8408016-222	0,0692	0,0272	0,0390
075-kas-8383016-223	0,0492	0,0060	0,0107
076-kas-8382016-224	0,0213	0,0020	0,0036
077-kas-1152016-225	0,0333	0,0098	0,0152
078-kas-1110016-226	0,1806	0,0588	0,0887
079-kas-848017-227	0,0538	0,0195	0,0286
080-kas-8384017-228	0,1200	0,0148	0,0264
081-kas-8017-229	0,1757	0,0610	0,0906
082-kas-4017-230	0,2222	0,0969	0,1350
083-kas-1152017-231	0,0680	0,0174	0,0277
084-kas-1113017-232	0,0090	0,0256	0,0133

085-kas-8905018-233	0,0986	0,0199	0,0331
086-kas-8534018-234	0,0652	0,0056	0,0104
087-kas-8420018-235	0,1039	0,0135	0,0239
088-kas-37018-236	0,1681	0,0377	0,0615
089-kas-8534019-237	0,0000	0,0000	0,0000
090-kas-848019-238	0,0588	0,0102	0,0173
091-kas-8383019-239	0,0261	0,0095	0,0139
092-kas-5019-240	0,1050	0,0344	0,0518
093-kas-4019-241	0,0471	0,0169	0,0248
094-kas-38019-242	0,0130	0,0016	0,0028
095-kas-37019-243	0,4634	0,0892	0,1496
096-kas-848020-244	0,2204	0,0594	0,0936
097-kas-8385021-245	0,1042	0,0063	0,0118
098-kas-8383021-246	0,1301	0,0285	0,0467
099-kas-1128021-247	0,2645	0,0328	0,0584
100-kas-8905022-248	0,0652	0,0164	0,0262

Priloga 7: ocena primerjave semantičnega povzemanja s človeškimi povzetki z Rouge-3.

LSA/človeški (Rouge-3)	Recall	Precision	F measure
001-kas-1110000-149	0,0133	0,0045	0,0068
002-kas-16000-150	0,0112	0,0021	0,0035
003-kas-837001-151	0,0160	0,0034	0,0056
004-kas-8381001-152	0,0000	0,0000	0,0000
005-kas-8418001-153	0,0052	0,0027	0,0036
006-kas-38001-154	0,0122	0,0015	0,0027
007-kas-8810002-155	0,0218	0,0086	0,0124
008-kas-38003-156	0,0000	0,0000	0,0000
009-kas-7003-157	0,0000	0,0000	0,0000
010-kas-8384003-158	0,0117	0,0047	0,0067
011-kas-8381004-159	0,0123	0,0017	0,0030
012-kas-8386004-160	0,0177	0,0060	0,0090
013-kas-8905005-161	0,0106	0,0025	0,0040
014-kas-8386005-162	0,0440	0,0178	0,0253
015-kas-5005-163	0,0195	0,0065	0,0097
016-kas-1130005-164	0,0070	0,0009	0,0017
017-kas-1129005-165	0,0920	0,0099	0,0178
018-kas-844006-166	0,0667	0,0079	0,0142
019-kas-838006-167	0,0000	0,0000	0,0000
020-kas-5006-168	0,0060	0,0014	0,0023
021-kas-8535007-169	0,4029	0,0777	0,1302

022-kas-8386007-170	0,0000	0,0000	0,0000
023-kas-8382007-171	0,1353	0,0378	0,0591
024-kas-37007-172	0,0000	0,0000	0,0000
025-kas-1129007-173	0,0000	0,0000	0,0000
026-kas-848008-174	0,0211	0,0054	0,0086
027-kas-842008-175	0,0000	0,0000	0,0000
028-kas-8383008-176	0,0000	0,0000	0,0000
029-kas-38008-177	0,0082	0,0021	0,0034
030-kas-15008-178	0,1268	0,0591	0,0806
031-kas-1128008-179	0,0565	0,0129	0,0210
032-kas-848009-180	0,0179	0,0046	0,0073
033-kas-845009-181	0,0098	0,0018	0,0030
034-kas-841009-182	0,0000	0,0000	0,0000
035-kas-8382009-183	0,0135	0,0015	0,0027
036-kas-5009-184	0,0111	0,0036	0,0054
037-kas-1129009-185	0,0059	0,0011	0,0019
038-kas-1128009-186	0,0329	0,0077	0,0125
039-kas-848010-187	0,0114	0,0021	0,0035
040-kas-842010-188	0,0094	0,0061	0,0074
041-kas-8386010-189	0,0000	0,0000	0,0000
042-kas-8385010-190	0,0168	0,0034	0,0057
043-kas-1135010-191	0,0556	0,0036	0,0068
044-kas-848011-192	0,0000	0,0000	0,0000
045-kas-8383011-193	0,0541	0,0057	0,0104
046-kas-8382011-194	0,1475	0,0166	0,0299
047-kas-5011-195	0,0753	0,0152	0,0253
048-kas-4011-196	0,1293	0,0619	0,0837
049-kas-1152011-197	0,0000	0,0000	0,0000
050-kas-1130011-198	0,0301	0,0043	0,0075
051-kas-8905012-199	0,0606	0,0155	0,0247
052-kas-1152012-200	0,0000	0,0000	0,0000
053-kas-1144012-369	0,0415	0,0154	0,0224
054-kas-8905013-202	0,0058	0,0016	0,0025
055-kas-848013-203	0,0292	0,0063	0,0104
056-kas-844013-204	0,0519	0,0147	0,0230
057-kas-8384013-205	0,1333	0,0082	0,0154
058-kas-8383013-206	0,0230	0,0019	0,0035
059-kas-5013-207	0,0000	0,0000	0,0000
060-kas-1130013-208	0,0054	0,0010	0,0017
061-kas-1129013-209	0,0244	0,0054	0,0089
062-kas-9014-210	0,0263	0,0082	0,0125
063-kas-8905014-211	0,0185	0,0089	0,0120
064-kas-852014-212	0,0923	0,0117	0,0208
065-kas-38014-213	0,0000	0,0000	0,0000
066-kas-1152014-214	0,0027	0,0026	0,0026

067-kas-1118014-368	0,0488	0,0151	0,0230
068-kas-8905015-216	0,0000	0,0000	0,0000
069-kas-8384015-217	0,0435	0,0045	0,0082
070-kas-37015-218	0,0123	0,0020	0,0034
071-kas-1152015-219	0,0000	0,0000	0,0000
072-kas-848016-220	0,0103	0,0019	0,0033
073-kas-8420016-221	0,0269	0,0091	0,0136
074-kas-8408016-222	0,0243	0,0095	0,0137
075-kas-8383016-223	0,0167	0,0020	0,0036
076-kas-8382016-224	0,0000	0,0000	0,0000
077-kas-1152016-225	0,0000	0,0000	0,0000
078-kas-1110016-226	0,1234	0,0400	0,0604
079-kas-848017-227	0,0155	0,0056	0,0082
080-kas-8384017-228	0,0270	0,0033	0,0059
081-kas-8017-229	0,0748	0,0259	0,0385
082-kas-4017-230	0,1157	0,0504	0,0702
083-kas-1152017-231	0,0000	0,0000	0,0000
084-kas-1113017-232	0,0000	0,0000	0,0000
085-kas-8905018-233	0,0213	0,0043	0,0071
086-kas-8534018-234	0,0222	0,0019	0,0035
087-kas-8420018-235	0,0132	0,0017	0,0030
088-kas-37018-236	0,0763	0,0170	0,0278
089-kas-8534019-237	0,0000	0,0000	0,0000
090-kas-848019-238	0,0000	0,0000	0,0000
091-kas-8383019-239	0,0000	0,0000	0,0000
092-kas-5019-240	0,0389	0,0127	0,0192
093-kas-4019-241	0,0000	0,0000	0,0000
094-kas-38019-242	0,0000	0,0000	0,0000
095-kas-37019-243	0,4180	0,0799	0,1342
096-kas-848020-244	0,0757	0,0203	0,0320
097-kas-8385021-245	0,0213	0,0013	0,0024
098-kas-8383021-246	0,0820	0,0178	0,0293
099-kas-1128021-247	0,1494	0,0184	0,0328
100-kas-8905022-248	0,0073	0,0018	0,0029

Priloga 8: ocena primerjave semantičnega povzemanja s povzetki arhetipi z Rouge-1.

AA/LSA (Rouge-1)	Recall	Precision	F measure
001-kas-1110000-149	0,1602	0,3935	0,2274
002-kas-16000-150	0,1120	0,1995	0,1435
003-kas-837001-151	0,2346	0,4467	0,3076
004-kas-8381001-152	0,2066	0,3567	0,2616
005-kas-8418001-153	0,1868	0,2466	0,2117
006-kas-38001-154	0,1766	0,4647	0,2556
007-kas-8810002-155	0,2198	0,4685	0,2991
008-kas-38003-156	0,2157	0,4123	0,2833
009-kas-7003-157	0,2556	0,4439	0,3244
010-kas-8384003-158	0,1900	0,4113	0,2599
011-kas-8381004-159	0,2207	0,4323	0,2922
012-kas-8386004-160	0,2093	0,4710	0,2898
013-kas-8905005-161	0,1756	0,5753	0,2689
014-kas-8386005-162	0,1184	0,2485	0,1603
015-kas-5005-163	0,2866	0,4377	0,3463
016-kas-1130005-164	0,1338	0,4694	0,2081
017-kas-1129005-165	0,1261	0,6611	0,2117
018-kas-844006-166	0,4010	0,9200	0,5585
019-kas-838006-167	0,2672	0,4480	0,3346
020-kas-5006-168	0,0208	0,0296	0,0244
021-kas-8535007-169	0,2282	0,5957	0,3300
022-kas-8386007-170	0,2876	0,5519	0,3775
023-kas-8382007-171	0,2531	0,4427	0,3221
024-kas-37007-172	0,2302	0,4883	0,3129
025-kas-1129007-173	0,1574	0,4809	0,2371
026-kas-848008-174	0,1290	0,1791	0,1500
027-kas-842008-175	0,2011	0,3697	0,2605
028-kas-8383008-176	0,0726	0,1992	0,1063
029-kas-38008-177	0,2373	0,3852	0,2936
030-kas-15008-178	0,1063	0,1669	0,1298
031-kas-1128008-179	0,1861	0,4861	0,2689
032-kas-848009-180	0,2607	0,5712	0,3580
033-kas-845009-181	0,1605	0,2718	0,2011
034-kas-841009-182	0,2274	0,3427	0,2733
035-kas-8382009-183	0,2141	0,5051	0,3007
036-kas-5009-184	0,2106	0,4170	0,2798
037-kas-1129009-185	0,1767	0,5573	0,2683
038-kas-1128009-186	0,1914	0,4352	0,2659
039-kas-848010-187	0,2098	0,3282	0,2560

040-kas-842010-188	0,2530	0,4264	0,3173
041-kas-8386010-189	0,2282	0,5280	0,3186
042-kas-8385010-190	0,2496	0,4713	0,3263
043-kas-1135010-191	0,0876	0,2459	0,1291
044-kas-848011-192	0,1735	0,3733	0,2363
045-kas-8383011-193	0,2507	0,6105	0,3554
046-kas-8382011-194	0,2486	0,4479	0,3197
047-kas-5011-195	0,1898	0,4503	0,2670
048-kas-4011-196	0,2669	0,4281	0,3288
049-kas-1152011-197	0,1444	0,2272	0,1765
050-kas-1130011-198	0,1885	0,6221	0,2880
051-kas-8905012-199	0,1897	0,5405	0,2807
052-kas-1152012-200	0,3497	0,4214	0,3820
053-kas-1144012-369	0,2734	0,5671	0,3689
054-kas-8905013-202	0,2955	0,6366	0,4036
055-kas-848013-203	0,2896	0,6058	0,3914
056-kas-844013-204	0,2083	0,4171	0,2777
057-kas-8384013-205	0,3371	0,8401	0,4811
058-kas-8383013-206	0,1506	0,5778	0,2389
059-kas-5013-207	0,1827	0,3774	0,2461
060-kas-1130013-208	0,1690	0,5611	0,2595
061-kas-1129013-209	0,1808	0,5103	0,2670
062-kas-9014-210	0,1000	0,1752	0,1273
063-kas-8905014-211	0,2715	0,4101	0,3267
064-kas-852014-212	0,2553	0,4312	0,3207
065-kas-38014-213	0,3174	0,3917	0,3506
066-kas-1152014-214	0,2584	0,3203	0,2856
067-kas-1118014-368	0,2725	0,3555	0,3085
068-kas-8905015-216	0,2376	0,3988	0,2977
069-kas-8384015-217	0,1695	0,6179	0,2659
070-kas-37015-218	0,2766	0,4661	0,3472
071-kas-1152015-219	0,2316	0,1707	0,1965
072-kas-848016-220	0,2115	0,3907	0,2745
073-kas-8420016-221	0,2319	0,4324	0,3018
074-kas-8408016-222	0,1893	0,4943	0,2737
075-kas-8383016-223	0,2830	0,4918	0,3592
076-kas-8382016-224	0,1960	0,3635	0,2546
077-kas-1152016-225	0,1902	0,1388	0,1604
078-kas-1110016-226	0,1520	0,2660	0,1934
079-kas-848017-227	0,1014	0,0945	0,0977
080-kas-8384017-228	0,2278	0,4594	0,3046
081-kas-8017-229	0,2646	0,3948	0,3169
082-kas-4017-230	0,2697	0,5824	0,3687
083-kas-1152017-231	0,3329	0,4680	0,3889
084-kas-1113017-232	0,4416	0,4145	0,4276

085-kas-8905018-233	0,0000	0,0000	0,0000
086-kas-8534018-234	0,2795	0,5155	0,3625
087-kas-8420018-235	0,2938	0,6133	0,3972
088-kas-37018-236	0,1833	0,2818	0,2221
089-kas-8534019-237	0,0000	0,0000	0,0000
090-kas-848019-238	0,2657	0,5974	0,3676
091-kas-8383019-239	0,3097	0,5048	0,3839
092-kas-5019-240	0,1664	0,3573	0,2263
093-kas-4019-241	0,1748	0,2911	0,2183
094-kas-38019-242	0,1901	0,3408	0,2418
095-kas-37019-243	0,1234	0,2424	0,1636
096-kas-848020-244	0,1802	0,4372	0,2551
097-kas-8385021-245	0,2368	0,5821	0,3359
098-kas-8383021-246	0,3002	0,6017	0,4005
099-kas-1128021-247	0,1340	0,5604	0,2161
100-kas-8905022-248	0,2091	0,4077	0,2764

Priloga 9: ocena primerjave semantičnega povzemanja s povzetki arhetipi z Rouge-2.

AA/LSA (Rouge-2)	Recall	Precision	F measure
001-kas-1110000-149	0,0369	0,0943	0,0530
002-kas-16000-150	0,0177	0,0316	0,0227
003-kas-837001-151	0,0333	0,0636	0,0437
004-kas-8381001-152	0,0260	0,0450	0,0330
005-kas-8418001-153	0,0175	0,0211	0,0191
006-kas-38001-154	0,0453	0,1207	0,0658
007-kas-8810002-155	0,0698	0,1482	0,0948
008-kas-38003-156	0,0844	0,1622	0,1110
009-kas-7003-157	0,0866	0,1517	0,1103
010-kas-8384003-158	0,0401	0,0869	0,0549
011-kas-8381004-159	0,1037	0,2038	0,1375
012-kas-8386004-160	0,0599	0,1353	0,0830
013-kas-8905005-161	0,0944	0,3132	0,1449
014-kas-8386005-162	0,0298	0,0623	0,0403
015-kas-5005-163	0,1102	0,1662	0,1324
016-kas-1130005-164	0,0304	0,1071	0,0474
017-kas-1129005-165	0,0440	0,2312	0,0740
018-kas-844006-166	0,3563	0,8193	0,4967
019-kas-838006-167	0,0792	0,1344	0,0996
020-kas-5006-168	0,0014	0,0020	0,0016

021-kas-8535007-169	0,0824	0,2154	0,1192
022-kas-8386007-170	0,1080	0,2075	0,1419
023-kas-8382007-171	0,1184	0,2070	0,1507
024-kas-37007-172	0,0579	0,1232	0,0787
025-kas-1129007-173	0,0250	0,0765	0,0377
026-kas-848008-174	0,0162	0,0224	0,0188
027-kas-842008-175	0,0309	0,0571	0,0401
028-kas-8383008-176	0,0075	0,0204	0,0110
029-kas-38008-177	0,0605	0,0982	0,0749
030-kas-15008-178	0,0159	0,0243	0,0192
031-kas-1128008-179	0,0360	0,0927	0,0518
032-kas-848009-180	0,1110	0,2440	0,1526
033-kas-845009-181	0,0178	0,0323	0,0229
034-kas-841009-182	0,0337	0,0509	0,0405
035-kas-8382009-183	0,0571	0,1351	0,0803
036-kas-5009-184	0,0328	0,0647	0,0435
037-kas-1129009-185	0,0556	0,1757	0,0844
038-kas-1128009-186	0,0498	0,1139	0,0693
039-kas-848010-187	0,0450	0,0703	0,0549
040-kas-842010-188	0,0485	0,0802	0,0604
041-kas-8386010-189	0,0664	0,1540	0,0928
042-kas-8385010-190	0,0845	0,1599	0,1105
043-kas-1135010-191	0,0308	0,0861	0,0454
044-kas-848011-192	0,0410	0,0869	0,0555
045-kas-8383011-193	0,0839	0,2048	0,1191
046-kas-8382011-194	0,0969	0,1749	0,1247
047-kas-5011-195	0,0744	0,1765	0,1046
048-kas-4011-196	0,1163	0,1866	0,1433
049-kas-1152011-197	0,0181	0,0285	0,0221
050-kas-1130011-198	0,0758	0,2546	0,1163
051-kas-8905012-199	0,0685	0,1936	0,1011
052-kas-1152012-200	0,1712	0,2028	0,1856
053-kas-1144012-369	0,1067	0,2213	0,1439
054-kas-8905013-202	0,2056	0,4452	0,2812
055-kas-848013-203	0,1533	0,3154	0,2061
056-kas-844013-204	0,0579	0,1161	0,0772
057-kas-8384013-205	0,2915	0,7277	0,4163
058-kas-8383013-206	0,0498	0,1914	0,0790
059-kas-5013-207	0,0438	0,0911	0,0592
060-kas-1130013-208	0,0571	0,1926	0,0879
061-kas-1129013-209	0,0541	0,1532	0,0799
062-kas-9014-210	0,0061	0,0108	0,0078
063-kas-8905014-211	0,0852	0,1288	0,1026
064-kas-852014-212	0,1196	0,2024	0,1504
065-kas-38014-213	0,1337	0,1677	0,1488

066-kas-1152014-214	0,0842	0,1091	0,0950
067-kas-1118014-368	0,0401	0,0523	0,0454
068-kas-8905015-216	0,0485	0,0805	0,0605
069-kas-8384015-217	0,0708	0,2607	0,1113
070-kas-37015-218	0,0809	0,1365	0,1016
071-kas-1152015-219	0,0239	0,0176	0,0203
072-kas-848016-220	0,0501	0,0927	0,0650
073-kas-8420016-221	0,0705	0,1322	0,0919
074-kas-8408016-222	0,0761	0,1976	0,1099
075-kas-8383016-223	0,1052	0,1816	0,1332
076-kas-8382016-224	0,0853	0,1590	0,1110
077-kas-1152016-225	0,0294	0,0215	0,0248
078-kas-1110016-226	0,0137	0,0239	0,0174
079-kas-848017-227	0,0153	0,0149	0,0151
080-kas-8384017-228	0,0832	0,1680	0,1113
081-kas-8017-229	0,0657	0,0979	0,0787
082-kas-4017-230	0,1032	0,2238	0,1413
083-kas-1152017-231	0,1960	0,2759	0,2291
084-kas-1113017-232	0,2802	0,2633	0,2715
085-kas-8905018-233	0,0000	0,0000	0,0000
086-kas-8534018-234	0,0536	0,0990	0,0695
087-kas-8420018-235	0,1535	0,3207	0,2076
088-kas-37018-236	0,0358	0,0550	0,0434
089-kas-8534019-237	0,0000	0,0000	0,0000
090-kas-848019-238	0,1915	0,4332	0,2655
091-kas-8383019-239	0,1303	0,2126	0,1616
092-kas-5019-240	0,0362	0,0794	0,0496
093-kas-4019-241	0,0253	0,0421	0,0316
094-kas-38019-242	0,0719	0,1297	0,0917
095-kas-37019-243	0,0039	0,0077	0,0052
096-kas-848020-244	0,0464	0,1139	0,0659
097-kas-8385021-245	0,1242	0,3059	0,1763
098-kas-8383021-246	0,1468	0,2937	0,1957
099-kas-1128021-247	0,0576	0,2453	0,0933
100-kas-8905022-248	0,0319	0,0622	0,0421

Priloga 10: ocena primerjave semantičnega povzemanja s povzetki arhetipi z Rouge-3.

AA/LSA (Rouge-3)	Recall	Precision	F measure
001-kas-1110000-149	0,0234	0,0615	0,0339
002-kas-16000-150	0,0010	0,0019	0,0013
003-kas-837001-151	0,0026	0,0050	0,0034
004-kas-8381001-152	0,0000	0,0000	0,0000
005-kas-8418001-153	0,0027	0,0032	0,0029
006-kas-38001-154	0,0119	0,0316	0,0173
007-kas-8810002-155	0,0411	0,0872	0,0558
008-kas-38003-156	0,0585	0,1126	0,0770
009-kas-7003-157	0,0472	0,0832	0,0602
010-kas-8384003-158	0,0173	0,0375	0,0237
011-kas-8381004-159	0,0698	0,1375	0,0926
012-kas-8386004-160	0,0322	0,0731	0,0447
013-kas-8905005-161	0,0675	0,2255	0,1038
014-kas-8386005-162	0,0194	0,0405	0,0262
015-kas-5005-163	0,0779	0,1165	0,0933
016-kas-1130005-164	0,0089	0,0311	0,0138
017-kas-1129005-165	0,0220	0,1161	0,0370
018-kas-844006-166	0,3331	0,7673	0,4645
019-kas-838006-167	0,0501	0,0859	0,0633
020-kas-5006-168	0,0000	0,0000	0,0000
021-kas-8535007-169	0,0472	0,1233	0,0682
022-kas-8386007-170	0,0811	0,1564	0,1067
023-kas-8382007-171	0,0945	0,1657	0,1204
024-kas-37007-172	0,0311	0,0666	0,0424
025-kas-1129007-173	0,0071	0,0220	0,0108
026-kas-848008-174	0,0000	0,0000	0,0000
027-kas-842008-175	0,0053	0,0099	0,0069
028-kas-8383008-176	0,0014	0,0036	0,0020
029-kas-38008-177	0,0309	0,0501	0,0382
030-kas-15008-178	0,0045	0,0070	0,0055
031-kas-1128008-179	0,0103	0,0256	0,0147
032-kas-848009-180	0,0775	0,1706	0,1065
033-kas-845009-181	0,0053	0,0097	0,0069
034-kas-841009-182	0,0047	0,0072	0,0057
035-kas-8382009-183	0,0401	0,0951	0,0564
036-kas-5009-184	0,0133	0,0262	0,0177
037-kas-1129009-185	0,0352	0,1116	0,0535
038-kas-1128009-186	0,0330	0,0756	0,0460
039-kas-848010-187	0,0283	0,0443	0,0345

040-kas-842010-188	0,0152	0,0242	0,0187
041-kas-8386010-189	0,0340	0,0790	0,0475
042-kas-8385010-190	0,0419	0,0796	0,0549
043-kas-1135010-191	0,0188	0,0525	0,0277
044-kas-848011-192	0,0236	0,0487	0,0317
045-kas-8383011-193	0,0295	0,0720	0,0418
046-kas-8382011-194	0,0545	0,0986	0,0702
047-kas-5011-195	0,0324	0,0773	0,0457
048-kas-4011-196	0,0732	0,1176	0,0902
049-kas-1152011-197	0,0021	0,0034	0,0026
050-kas-1130011-198	0,0491	0,1690	0,0758
051-kas-8905012-199	0,0323	0,0910	0,0477
052-kas-1152012-200	0,1291	0,1522	0,1396
053-kas-1144012-369	0,0701	0,1456	0,0946
054-kas-8905013-202	0,1715	0,3723	0,2348
055-kas-848013-203	0,1157	0,2378	0,1556
056-kas-844013-204	0,0230	0,0462	0,0307
057-kas-8384013-205	0,2490	0,6224	0,3557
058-kas-8383013-206	0,0302	0,1164	0,0479
059-kas-5013-207	0,0228	0,0475	0,0308
060-kas-1130013-208	0,0205	0,0724	0,0320
061-kas-1129013-209	0,0298	0,0847	0,0441
062-kas-9014-210	0,0000	0,0000	0,0000
063-kas-8905014-211	0,0588	0,0890	0,0708
064-kas-852014-212	0,0877	0,1485	0,1103
065-kas-38014-213	0,1079	0,1359	0,1203
066-kas-1152014-214	0,0545	0,0717	0,0619
067-kas-1118014-368	0,0126	0,0164	0,0142
068-kas-8905015-216	0,0238	0,0390	0,0295
069-kas-8384015-217	0,0501	0,1856	0,0788
070-kas-37015-218	0,0494	0,0835	0,0621
071-kas-1152015-219	0,0000	0,0000	0,0000
072-kas-848016-220	0,0280	0,0519	0,0364
073-kas-8420016-221	0,0444	0,0832	0,0579
074-kas-8408016-222	0,0599	0,1555	0,0864
075-kas-8383016-223	0,0833	0,1438	0,1055
076-kas-8382016-224	0,0557	0,1048	0,0727
077-kas-1152016-225	0,0025	0,0018	0,0021
078-kas-1110016-226	0,0000	0,0000	0,0000
079-kas-848017-227	0,0098	0,0095	0,0096
080-kas-8384017-228	0,0446	0,0901	0,0596
081-kas-8017-229	0,0247	0,0367	0,0295
082-kas-4017-230	0,0585	0,1272	0,0801
083-kas-1152017-231	0,1642	0,2310	0,1919
084-kas-1113017-232	0,2279	0,2142	0,2208

085-kas-8905018-233	0,0000	0,0000	0,0000
086-kas-8534018-234	0,0226	0,0418	0,0293
087-kas-8420018-235	0,1149	0,2404	0,1555
088-kas-37018-236	0,0085	0,0132	0,0103
089-kas-8534019-237	0,0000	0,0000	0,0000
090-kas-848019-238	0,1766	0,4006	0,2450
091-kas-8383019-239	0,1081	0,1765	0,1341
092-kas-5019-240	0,0091	0,0201	0,0125
093-kas-4019-241	0,0019	0,0032	0,0024
094-kas-38019-242	0,0578	0,1051	0,0739
095-kas-37019-243	0,0000	0,0000	0,0000
096-kas-848020-244	0,0167	0,0418	0,0238
097-kas-8385021-245	0,0760	0,1876	0,1079
098-kas-8383021-246	0,1078	0,2159	0,1438
099-kas-1128021-247	0,0284	0,1229	0,0462
100-kas-8905022-248	0,0137	0,0267	0,0181