Ekstrakcija ključnih riječi iz tekstova na talijanskom jeziku

Pokos, Marija

Sveučilište u Rijeci – Odjel za informatiku

Diplomski dvopredmetni studij informatike i talijanskog jezika i književnosti

Marija Pokos

Ekstrakcija ključnih riječi iz tekstova na talijanskome jeziku

Diplomski rad

Rijeka, rujan 2018.

Sveučilište u Rijeci – Odjel za informatiku

Diplomski dvopredmetni studij informatike i talijanskog jezika i književnosti

Marija Pokos

Ekstrakcija ključnih riječi iz tekstova na talijanskome jeziku

Diplomski rad

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

Rijeka, rujan 2018.

ODJEL ZA INFORMATIKU
SVEUČILIŠTE U RIJECI

Rijeka, 08.02.2018.

Zadatak za diplomski rad

Pristupnica: **Marija Pokos**

Naziv diplomskog rada**: Ekstrakcija ključnih riječi iz tekstova na talijanskome jeziku**

Naziv diplomskog rada na eng. jeziku: **Keyword Extraction from Italian Texts**

Sadržaj zadatka:

Glavni cilj diplomskog rada jest istražiti postojeće sustave, postupke i skupove podatka za ekstrakciju ključnih riječi iz talijanskih tekstova. U praktičnome dijelu diplomskoga rada će se testirati postupci ekstrakcije ključnih na vlastitome skupu podataka. Vrednovanje postupaka će se izvršit pomoću uobičajenih mjera u postupcima ekstrakcije ključnih riječi. Također u diplomskome radu će se proučiti i primijeniti alati namijenjeni za procjenu čitljivosti talijanskoga teksta, te procijeniti mogućnost njihove primjene prilikom ekstrakcije ključnih riječi.

Mentorica:

Izv. prof. dr. sc. Sanda Martinčić-Ipšić

Voditeljica za diplomske radove:

Dr. sc. Martina Ašenbrener Katić

Martina Ašenbrener Katić

Zadatak preuzet:

Pokos Marija

(Marija Pokos)

Ekstrakcija ključnih riječi iz tekstova na talijanskome jeziku

Sažetak

Ekstrakcija ključnih riječi je metoda kojom se automatski identificira skup izraza koji najbolje opisuju dokument. Cilj ovog rada je istražiti sustave za ekstrakciju ključnih riječi te predstaviti rad algoritma RAKE i algoritma Maui.

U uvodnom dijelu rada objasnit će se teorijski dio o ekstrakciji ključnih riječi, lingvističko stajalište te procesi lematizacije i korjenovanja. Time će se dobiti podloga za daljnje istraživanje.

Kako bi se pobliže objasnili algoritmi, testirat će ih se na vlastitome skupu podataka. Radi se o novinskim tekstovima umjetničkog sadržaja prevedenim na talijanski jezik. Skup podataka na kojim se radi testiranje još su i stop riječi (engl. stopwords) talijanskog jezika i ključne riječi označene od strane čovjeka.

Nakon analize algoritama objašnjeno je i prikazano vrednovanje postupaka pomoću uobičajenih mjera u postupcima ekstrakcije ključnih riječi, odnosno pomoću mjera precision, recall i F1.

Rezultati pokazuju da ključne riječi koje je označio algoritam Maui imaju više sličnosti s ključnim riječima koje je označio čovjek. Ključne riječi koje je označio algoritam RAKE, također imaju sličnosti, no u manjoj mjeri.

Ključne riječi: Ekstrakcija ključnih riječi, ključne riječi, RAKE, Maui, Precision, Recall, F1

Kyeword Extraction from Italian Texts

Abstract

Keyword extraction is a method that identifies a set of terms that describes a document in the most effective way. The goal of this thesis is to research the systems for keyword extraction and to show how the algorithms RAKE and Maui are working.

First, the theoretical part of keyword extraction will be explained, and then the linguistic standpoint and the process of lemmatization and stemming. This will provide a solid background for further research.

To explain the algorithms more closely, they will be tested on a private set of data. The set of data consists of texts with the topic of art translated to Italian. One part of data are also Italian stopwords and keywords annotated by a human.

After the analysis of algorithms, the evaluation of procedures with help of the usual measures in the procedures of keyword extraction is explained and shown. Those measures are precision, recall and F1.

The results have shown that keywords annotated by the algorithm Maui are more similar to the keywords annotated by human. The keywords annotated by algorithm RAKE also have some similarity but in lesser extent.

Keywords: Keyword extraction, keywords, RAKE, Maui, Precision, Recall, F1

L'estrazione di parole chiave dai testi in lingua italiana

Sommario

L’estrazione di parole chiave è un metodo con il quale si identifica automaticamente un set di dati che descrivono meglio un documento. Lo scopo di questa tesi è di analizzare il funzionamento dei sistemi per l’estrazione di parole chiave e mostrare come funzionano gli algoritmi RAKE e Maui.

Nell’introduzione del lavoro si spiegherà la parte teorica dell’estrazione di parole chiave, il punto di vista linguistico e i processi di lemmatizzazione e stemming. Questo farà da base per l’ulteriore ricerca.

Per spiegarli meglio, gli algoritmi saranno testati attraverso il set di dati predeterminato. Si tratta di testi giornalistici che trattano d’arte e che sono stati tradotti in lingua italiana. I set di dati sui quali si fa la ricerca sono da un lato le parole “stopwords” e dall’altro le parole chiave scelte dagli studenti.

Dopo l’analisi degli algoritmi è spiegata e mostrata la valutazione dei procedimenti attraverso i consueti sistemi nei processi di estrazione di parole chiave, cioè con l’aiuto dei sistemi precision, recall e F1.

I risultati mostrano che le parole chiave assegnate dall’algoritmo Maui assomigliano più alle parole chiave assegnate dall’uomo. Anche le parole chiave assegnate dall’algoritmo RAKE hanno anche una certa similitudine, ma in misura più piccola.

Parole chiave: estrazione di parole chiave, parole chiave, RAKE, Maui, Precision, Recall, F1

Sadržaj

1. Uvod

Ekstrakcija ključnih riječi novi je pojam koji se počeo pojavljivati na području informatike unatrag nekih desetak godina. Ekstrakcija ključnih riječi je metoda kojom se automatski identificira skup termina koji najbolje opisuju dokument (Beliga i sur., 2018).

U prvom dijelu ovog rada objasnit će se teorijska pozadina ekstrakcije ključnih riječi, indeksiranja, pretraživanja informacija (Information retrievala) i klasifikacije dokumenata. Ukratko će se objasniti i procesi lematizacije i korjenovanja (stemminga) koji ujedno i čine pozadinu koja bi mogla olakšati i poboljšati proces računalne ekstrakcije ključnih riječi.

U drugom dijelu govorit će se o provedenom eksperimentu. Istraživana su dva algoritma za ekstrakciju ključnih riječi RAKE (engl. Rapid Automatic Keyword Extraction) (Airpair, 2018) i Maui (Airpair, 2018). Ti algoritmi bili su testirani na određenom skupu podataka, odnosno na novinskim tekstovima umjetničkog sadržaja prevedenih na talijanski jezik.

U radu su prikazani i uspoređeni rezultati kroz različite aspekte. Razlika u rezultatima uvjetovana različitim timovima koji su označavali tekst, razlika u rezultatima nakon čišćenja teksta od nepotrebnih elemenata i razlika koja nastaje zbog korištenja dva različita algoritma RAKE i Maui.

Na kraju je provedena evaluacija pomoću uobičajenih mjera za mjerenje uspješnosti postupka ekstrakcije ključnih riječi (precison, recall i F1) te je analizirano i raspravljeno u kojoj su mjeri algoritmi uspjeli izvršiti svoj zadatak.

2. Ekstrakcija ključnih riječi

2.1. Općenito

Zadatak ekstrakcije ključnih riječi je da se automatski identificira skup izraza koji najbolje opisuju dokument (Beliga i sur., 2018)..Obično se ekstrakcija riječi primjenjivala na jednom zasebnom dokumentu, no danas se sve više teži k zahtjevnijim zadacima. Odnosno, želi se postići da se ekstrakcija riječi koristi na cijelom skupu dokumenata ili na cijeloj web stranici (Beliga i sur., 2018).

Potreba za ekstrakcijom ključnih riječi nekada nije bila toliko velika, no rastom tekstualnih sadržaja na internetu današnja potreba za ekstrakcijom ključnih riječi sve je veća. Tako je ekstrakcija ključnih riječi u fokusu istraživanja već zadnjih desetak godina, a razlog tome je i eksponecijalno povećavanje količine podatka u eri „big-“ana" (Beliga i sur., 2018).

Može se reći da je ekstrakcija ključnih riječi bitna u svim pogledima, osobito u radu s tekstom. Osobe često pretražuju dokumente koji su im potrebni za njihove radove, seminare ili prezentacije te pomoću ključnih riječi mogu lakše i brže prosuditi hoće li im dokument koji su trenutno otvorili biti od koristi ili ne (Airpair, 2018). Ključne riječi bitne su i za kreatore web stranica jer na taj način mogu grupirati dokumente sličnog sadržaja (Airpair, 2018).

Ekstrakcija riječi se bazira na statističkom, lingvističkom i strojnom učenju (Beliga i sur., 2018). Isto tako tipični algoritam za ekstrakciju ključnih riječi ima tri komponente. Prvo se može govoriti o selekciji potencijalnih riječi. U tom koraku ekstrahiraju se sve moguće riječi, fraze i termini koji bi potencijalno mogli biti ključne riječi (Airpair, 2018). Drugi korak je izračun svojstva koji kvantificira potencijal riječi, izraza i fraza da postanu ključna riječ. Za svaku riječ, frazu ili termin koji predstavljaju potencijalnu ključnu riječ potrebno je izračunati svojstvo koje bi moglo odrediti da se radi o ključnoj riječi (Airpair, 2018). Primjerice neka riječ koja se pojavljuje u naslovu filma, mogla bi lako postati ključna riječ. Te zadnji korak je označavanje i bodovanje ključnih riječi. Sve riječi koje čine kandidate mogu biti bodovane prema određenoj formuli ili korištenjem strojnog učenja kojim bi se odredilo da riječ kandidat zapravo i je ključna riječ. Rezultat ili limit riječi na kraju se koristi kako bi se odredio konačni skup ključnih riječi (Airpair, 2018).

2.2. Primjena za indeksiranje, pretraživanje i klasifikaciju dokumenta

Primjena ekstrakcije riječi, pretraživanja informacija (Information retrieval), indeksiranja i klasifikacije dokumenata još je mlada grana koja se tek počela razvijati na području informatike.

Indeksiranje dokumenata na računalu je postupak u kojem informacijski sustav pregledava i istražuje dokumente te ih kategorizira kako bi se oni mogli lakše pretraživati (Wikipedia, 2018.) Najprije se počelo razvijati na području gdje se pojavljivalo mnogo dokumenta ili knjiga. Kao primjer mogu se navesti knjižnice koje su zapošljavale ljude samo kako bi označavali glavne teme nekih dokumenata (Medelyan, 2009). Danas u big data eri indeksiranje više nije potrebno samo u knjižnicama već je nužno i u tvrtkama, poduzećima, udrugama ili organizacijama, zbog rastuće količine tekstova koje moraju pohraniti i pretraživati. Stoga se počelo razvijati algoritme koji će automatski označavati u cijelosti ili djelomično teme tekstova (Medelyan, 2009).

Information retrieval (IR) je pronalaženje materijala, obično dokumenata, nestrukturirane prirode, primjerice tekstova, koji zadovoljavaju informaciju potrebnu iz velikog skupa podataka, obično pohranjenih u računalima (Cambridge University Press, 2009).

IR se počeo razvijati isto kao i indeksiranje na području knjižnica i na mjestima gdje su se pojavljivali veliki skupovi podataka. No, danas se način života promijenio te je IR potreban skoro svima u različitim životnim situacijama. Najjednostavniji primjeri korištenja IR-a su google tražilica ili pretraživanje e-mail sandučića (Cambridge University Press, 2009).

Klasifikacija dokumenata kao što i samo ime govori je svrstavanje dokumenata u neke klase. Koristi se kod računalne analize prirodnog jezika, a na taj način može se odrediti jesu li komentari na neki film, knjigu, časopis ili članak pozitivni ili negativni. Također, pomoću klasifikacije teksta moguće je odrediti spol ili dob neke osobe koja je napisala određeni tekst.

Kod definiranja klasifikacije teksta možemo reći da je klasifikacija postupak dodjeljivanja klase c iz skupa s konačnim brojem elemenata C, prema određenom pravilu, neoznačenom dokumentu d koji pripada skupu dokumenata D (James i sur., 2013.)

3. Metodologija

3.1. Priprema teksta

Kako bi se tekst mogao računalno analizirati i kako bi se iz njega moglo algoritamski izvaditi ključne riječi, najprije ga je bilo potrebno pripremiti. Skup podataka na kojem se radila analiza su tekstovi pisani na hrvatskom i engleskom jeziku koji su kasnije prevedeni i na talijanski jezik. Radi se o tekstovima većinom umjetničkog sadržaja, a ekstrakcija ključnih riječi se vršila na talijanskom prijevodu svih tekstova u kolekciji. U dobivenom skupu podataka nalazili su se već označene ključne riječi te oznake tipa *<figcaption> i </figcaption>* koje je bilo potrebno ukloniti. Ponekad za pripremu tekstova bilo bi potrebno provesti lematizaciju i stemming, no u ovom slučaju to nije bilo potrebno. Dobiveni skup podataka očistio se od ključnih riječi pisanih unutar teksta te od nepotrebnih oznaka navedenih ranije.

Drugi skup podataka koji je bio potreban kod analize bile su ključne riječi. Ključne riječi dobivene su na hrvatskom jeziku, a prema istim uputama ih je označavalo osam timova studenta. Ključne riječi su označavali studenti neovisno jedni od drugih te stoga nije bilo nužno da svaka datoteka sadrži označene sve identične ključne riječi. Budući da se u ovom radu radi analiza talijanskih tekstova i ekstrakcija ključnih riječi iz talijanskih tekstova bilo je potrebno prevesti i ključne riječi na talijanski jezik. Datoteke s ključnim riječima na talijanskom jeziku potrebne su za analizu teksta pomoću Maui algoritma.

3.2. Alati za lematizaciju i korijenovanje talijanskoga jezika

Lematizacija je proces kojim se pronalazi osnovni oblik riječi, odnosno lemma. Lematizacijom kao rezultat dobivamo riječi koje su zapisane u rječniku, odnosno imenice u nominativu jednine, infinitiv glagola i pozitiv pridjeva.

TreeTagger (Centrum für informations und sprachverarbeitung, 2018) program ispisuje POS[1] oznaku i prepoznaje lemmu. Rezultat se ispisuje u 3 stupca. Prvi stupac je riječ iz teksta, drugi POS oznaka i treći lemma. Prikaz rezultata može se dobiti u command promptu ili ispisan u .txt datoteci. Nedostatak programa je što ne prepoznaje složenije oblike riječi. Primjerice za lemmu glagola *scoprirà* ispisuje isti oblik i prepoznaje riječ kao glagol u imperativu, a zapravo je lemma *scoprire*, a vrsta riječi je glagol u futuru.

Drugi program koji omogućava ispis POS oznaka i pronalazi lemmu je LinguA (Italian Natural Language Processing Lab, 2018) moguće je pronaći online. Program, osim što zapisuje POS

[1] Parts-of-speech (POS) je obilježavanje vrste riječi, odnosno automatsko određivanje gramatičkih obilježja svake riječi u tekstu.

oznake i pronalazi lemmu, razdvaja rečenice, radi sintaksno parsiranje i prikazuje sintaksno stablo. Nedostatak ovog alata je što se u .txt datoteci može preuzeti samo dio u kojem se pronalazi lemma i POS oznaka. Nije moguće preuzeti cijelo sintaksno stablo (syntactic tree) niti dio u kojem su razdvojene rečenice.

Treći program, DyLan TextTools v2.1.9 (Istituto di Linguistica Computazionale, 2018), prikazuje POS oznake i pronalazi lemmu, a može se također pronaći online. Program omogućuje korisniku da unese tekst nakon čega se tekst analizira. Program razdvaja tekst na rečenice, iz rečenice izdvaja tokene, označava svaki token sa POS oznakom i pronalazi lemmu. Mana ovog programa je što se ni jedan rezultat ne može preuzeti u .txt obliku, dok mu je prednost što dodatno radi analizu čitljivosti teksta prema različitim kriterijima.

Stemming je metoda za korjenovanje. Tom metodom želi se grupirati riječi na nivou morfološkog korijena. Metodu možemo još nazvati i morfološko grupiranje. Cilj je otkriti i grupirati sve oblike iste riječi u korpusu.

Gotovo svi pronađeni stemmeri rade samo za engleski jezik. Najpoznatiji stemmer je Porter stemmer, radi na bazi JavaScripta, a funkcionalan je za engleski jezik. U ovom slučaju isprobano je radi li djelomično točno i za talijanski jezik.

Stemmer pronalazi nastavke tipične za engleski jezik kao što su –ed, -les, -ive i -e. Jedini nastavak koji bi mogao funkcionirati u talijanskom jeziku je -e, ali u većini slučajeva ni on nije valjan, tako da to nije primjerno rješenje za talijanski jezik.

Drugi program koji radi i za talijanski jezik radi pomoću Python-ovog NLTK-a.

Moguće je izabrati stemmer za određeni jezik, a zatim program izbaci stemmirani tekst. Mana ovog programa je što je moguće unijeti samo do 50 000 znakova. Za razliku od Porter stemmera, ovaj stemmer pronalazi relativno dobre nastavke za talijanski jezik te dobro prepoznaje morfološki korijen. Problem su riječi koje imaju više značenja. Na primjer, riječ *Nato*, stemmer prepoznaje kao glagol, a ne kao imenicu te zbog toga miče nastavak –o. Problem stvaraju i talijanska slova ì, à, è, ò, i ù. Iako program ne bi trebao maknuti nastavak s riječi *giovedì* i s riječi *realtà*, nastavci su ipak ispušteni.

4. Metode i alati za ekstrakciju ključnih pojmova

U ovom poglavlju objasnit će se princip rada RAKE i Maui metode. Odnosno rad dviju metoda koje se koriste kod ekstrakcije ključnih riječi. Metode su u ovom slučaju bile korištene kod tekstova na talijanskom jeziku.

4.1. RAKE

RAKE (engl. Rapid Automatic Keyword Extraction) je jednostavna biblioteka unutar Pythona koju se vrlo jednostavno koristi (Airpair, 2018). Možemo reći da je to algoritam koji automatski provodi ekstrakciju ključnih riječi iz dokumenata. RAKE je dobro poznat i vrlo korišten kao NLP tehnika, no njegov ishod ovisi o puno faktora. Možemo reći da ovisi o jeziku na kojem je sadržaj pisan, domeni sadržaja i o svrsi ključnih riječi (Hackage, 2018).

Kod razvijanja RAKE metode zapravo se željela razviti metoda ekstrakcije ključnih riječi koja je iznimno učinkovita, a jednostavna . Tako se RAKE metoda temelji na zapažanju da ključne riječi sadrže više riječi, no rijetko sadrže interpunkcijske znakove, stop riječi (engl. *stopwords)* ili nepunoznačnice (Hackage, 2018). Stop riječi uglavnom nisu uključene u analize jer ne daju dovoljno informacija ili u drugom slučaju nemaju značenje. Također, takve riječi nisu uključene u analizu jer se često pojavljuju u tekstovima, a nemaju bitniju ulogu kod odabira ključnih riječi.

Kako bi se RAKE metoda ispravno odvijala potrebna su tri parametra, odnosno potrebna je lista stop riječi na jeziku kojem se radi analiza (u ovom slučaju stop riječi talijanskog jezika), skup podataka koji označava kraj fraze te skup podataka koji označava kraj riječi. RAKE metoda koristi te parametre kako bi podijelila tekst u kandidate ključnih riječi koje se slijedno pojavljuju u tekstu.

4.2. Maui

Maui (engl. Multi-purpose automatic topic indexing) je proširena verzija algoritma Kea te se nalazi unutar GNU GPL-a kao licencirana biblioteka pisana u Javi (Airpair. 2018). Algoritam Maui može izvoditi zadatke kao što su ekstrakcija ključnih riječi (engl. keyword extraction), automatsko tagiranje (engl. automatic tagging), indeksiranje subjekata (engl. subject indexing) te može izdvojiti najbitnije koncepte i entitete iz Wikipedije (engl. extracting most relevant concepts and entities from Wikipedia) (Medelyan, 2009).

Za razliku od RAKE metode Maui algoritmu potrebno je predati datoteke koje sadrže tekstove i ključne riječi na kojima će algoritam učiti, a nakon toga vršiti i analizu.

5. Eksperiment

Cilj ovog rada bio je istražiti postojeće sustave, postupke i skupove podataka za ekstrakciju ključnih riječi iz talijanskih tekstova. Kod eksperimenta testirali su se postupci ekstrakcije ključnih riječi na zadanome skupu podataka kroz algoritme RAKE i Maui.

5.1. Tekstovi i priprema podataka

RAKE algoritam većinu stvari radi samostalno i za obradu podataka potrebni su tekstovi na kojima se testira te popis stop riječi (engl. stopwords). U ovom slučaju testiranje se vršilo na 35 tekstova koji govore o hrvatskoj umjetnosti i dizajnu.

Broj teksta	Broj riječi u neočišćenom tekstu	Broj riječi u očišćenom tekstu	Broj ključnih riječi	
2	998	988	Tim 1	6
			Tim 2	7
			Tim 3	5
			Tim 4	5
			Tim 5	5
			Tim 6	5
			Tim 7	6
			Tim 8	5
			Autor	5
3	292	283	Tim 1	7
			Tim 2	5
			Tim 3	4
			Tim 4	5
			Tim 5	5
			Tim 6	6
			Tim 7	5
			Tim 8	5
			Autor	6
4	1195	1151	Tim 1	5
			Tim 2	8
			Tim 3	4
			Tim 4	8

			Tim 5	8
			Tim 6	5
			Tim 7	8
			Tim 8	5
			Autor	6
5	1163	1149	Tim 1	7
			Tim 2	6
			Tim 3	6
			Tim 4	8
			Tim 5	6
			Tim 6	6
			Tim 7	5
			Tim 8	5
			Autor	
6	543	533	Tim 1	6
			Tim 2	7
			Tim 3	4
			Tim 4	6
			Tim 5	5
			Tim 6	4
			Tim 7	6
			Tim 8	5
			Autor	5
12	1174	1157	Tim 1	6
			Tim 2	5
			Tim 3	6
			Tim 4	7
			Tim 5	7
			Tim 6	6
			Tim 7	6
			Tim 8	7
			Autor	8
14	868	858	Tim 1	7

			Tim 2	5
			Tim 3	6
			Tim 4	8
			Tim 5	8
			Tim 6	6
			Tim 7	5
			Tim 8	5
			Autor	6
15	663	650	Tim 1	6
			Tim 2	6
			Tim 3	6
			Tim 4	6
			Tim 5	6
			Tim 6	5
			Tim 7	5
			Tim 8	4
			Autor	6
16	507	496	Tim 1	5
			Tim 2	8
			Tim 3	6
			Tim 4	6
			Tim 5	6
			Tim 6	7
			Tim 7	6
			Tim 8	8
			Autor	7
17	876	863	Tim 1	5
			Tim 2	6
			Tim 3	6
			Tim 4	6
			Tim 5	8
			Tim 6	8
			Tim 7	7

			Tim 8	8
			Autor	7
22	287	276	Tim 1	4
			Tim 2	5
			Tim 3	5
			Tim 4	7
			Tim 5	4
			Tim 6	8
			Tim 7	7
			Tim 8	7
			Autor	8
24	618	606	Tim 1	4
			Tim 2	8
			Tim 3	5
			Tim 4	5
			Tim 5	8
			Tim 6	5
			Tim 7	6
			Tim 8	7
			Autor	5
25	1549	1542	Tim 1	5
			Tim 2	7
			Tim 3	8
			Tim 4	8
			Tim 5	6
			Tim 6	6
			Tim 7	8
			Tim 8	8
			Autor	4
26	726	707	Tim 1	5
			Tim 2	7
			Tim 3	7
			Tim 4	5

			Tim 5	8
			Tim 6	8
			Tim 7	8
			Tim 8	8
			Autor	6
27	356	336	Tim 1	5
			Tim 2	6
			Tim 3	4
			Tim 4	5
			Tim 5	5
			Tim 6	6
			Tim 7	6
			Tim 8	7
			Autor	8
29	562	548	Tim 1	4
			Tim 2	7
			Tim 3	7
			Tim 4	5
			Tim 5	7
			Tim 6	5
			Tim 7	6
			Tim 8	7
			Autor	6
30	405	389	Tim 1	5
			Tim 2	4
			Tim 3	6
			Tim 4	6
			Tim 5	6
			Tim 6	8
			Tim 7	6
			Tim 8	7
			Autor	4
31	317	312	Tim 1	5

			Tim 2	8
			Tim 3	6
			Tim 4	5
			Tim 5	8
			Tim 6	6
			Tim 7	7
			Tim 8	6
			Autor	4
32	463	456	Tim 1	6
			Tim 2	6
			Tim 3	8
			Tim 4	5
			Tim 5	7
			Tim 6	5
			Tim 7	8
			Tim 8	7
			Autor	4
34	387	374	Tim 1	5
			Tim 2	8
			Tim 3	5
			Tim 4	6
			Tim 5	7
			Tim 6	5
			Tim 7	4
			Tim 8	5
			Autor	5
38	606	590	Tim 1	5
			Tim 2	8
			Tim 3	6
			Tim 4	5
			Tim 5	8
			Tim 6	6
			Tim 7	6

			Tim 8	7
			Autor	6
41	564	546	Tim 1	6
			Tim 2	7
			Tim 3	5
			Tim 4	7
			Tim 5	8
			Tim 6	4
			Tim 7	5
			Tim 8	7
			Autor	6
45	414	402	Tim 1	5
			Tim 2	4
			Tim 3	6
			Tim 4	4
			Tim 5	5
			Tim 6	5
			Tim 7	8
			Tim 8	6
			Autor	6
49	407	394	Tim 1	4
			Tim 2	8
			Tim 3	6
			Tim 4	4
			Tim 5	6
			Tim 6	5
			Tim 7	6
			Tim 8	6
			Autor	5
50	835	809	Tim 1	4
			Tim 2	6
			Tim 3	8
			Tim 4	5

			Tim 5	7
			Tim 6	5
			Tim 7	8
			Tim 8	7
			Autor	9
54	501	491	Tim 1	5
			Tim 2	8
			Tim 3	7
			Tim 4	6
			Tim 5	7
			Tim 6	8
			Tim 7	6
			Tim 8	5
			Autor	4
55	583	572	Tim 1	5
			Tim 2	5
			Tim 3	6
			Tim 4	5
			Tim 5	4
			Tim 6	6
			Tim 7	7
			Tim 8	6
			Autor	4
56	676	651	Tim 1	8
			Tim 2	5
			Tim 3	4
			Tim 4	6
			Tim 5	4
			Tim 6	6
			Tim 7	5
			Tim 8	5
			Autor	7
60	495	473	Tim 1	5

			Tim 2	5
			Tim 3	4
			Tim 4	5
			Tim 5	5
			Tim 6	6
			Tim 7	7
			Tim 8	5
			Autor	7
62	934	897	Tim 1	7
			Tim 2	6
			Tim 3	4
			Tim 4	6
			Tim 5	8
			Tim 6	7
			Tim 7	7
			Tim 8	6
			Autor	7
71	504	485	Tim 1	8
			Tim 2	5
			Tim 3	4
			Tim 4	7
			Tim 5	6
			Tim 6	6
			Tim 7	4
			Tim 8	5
			Autor	8
78	935	907	Tim 1	8
			Tim 2	8
			Tim 3	4
			Tim 4	5
			Tim 5	5
			Tim 6	6
			Tim 7	6

			Tim 8	6
			Autor	8
81	485	467	Tim 1	6
			Tim 2	5
			Tim 3	5
			Tim 4	4
			Tim 5	8
			Tim 6	4
			Tim 7	8
			Tim 8	6
			Autor	8
85	547	520	Tim 1	8
			Tim 2	6
			Tim 3	8
			Tim 4	5
			Tim 5	6
			Tim 6	7
			Tim 7	8
			Tim 8	7
			Autor	10
90	686	650	Tim 1	4
			Tim 2	6
			Tim 3	6
			Tim 4	7
			Tim 5	6
			Tim 6	5
			Tim 7	7
			Tim 8	8
			Autor	12

Tekstovi su originalno pisani hrvatskim jezikom, a prevedeni su na engleski i talijanski jezik. Važno je za napomenuti da su tekstovi na kojima se testirala metoda za ekstrakciju ključnih riječi pregledani od strane prevoditelja, dok kod nekih tekstova na kojima je algoritam učio možda ima i grešaka jer nisu u potpunosti obrađeni i pregledani.

Testiranje se provodilo u dva ciklusa. Prvi puta u tekstu su se nalazile ključne riječi označene od strane autora i dodatne oznake za slike poput <*figcaption*> i <*/figcaption*>, a drugi puta tekst je bio očišćen od tih elemenata.

Maui algoritam treba nešto više podataka kako bi izvršio ekstrakciju ključnih riječi. Maui algoritmu potrebno je predati skup podataka na kojima će algoritam učiti, odnosno trenirati, a zatim i podatke na kojima će se vršiti testiranje, odnosno ekstrakcija ključnih riječi. Podaci na kojima je algoritam učio bile su ključne riječi koje su označavali studenti te tekstovi koji su bili prevedeni s hrvatskog na talijanski. Dio tekstova je pregledan od strane prevoditelja, a drugom dijelu postoji mogućnost pogreške jer tekstovi nisu u potpunosti obrađeni.

Ključne riječi označivali su studenti podijeljeni u 8 timova. Svaki tim je označio ključne riječi za 35 tekstova te nije bilo nužno da svi imaju označene iste riječi. Tako je nastalo 280 datoteka u kojima su se nalazile ključne riječi označene od strane čovjeka u rasponu od 4 do 8 riječi po datoteci. Ključne riječi koje su označili studenti bile su na hrvatskom jeziku, te ih je bilo potrebno prevesti na talijanski jezik. Kod prevođenja težilo se da prijevodi ključnih riječi budu identični onima koji se pojavljuju u tekstu. Problem kod prevođenja nastaje kada riječ ili fraza koja se prevodi s hrvatskog jezika na talijanski jezik ima više riječi. Kao primjer možemo navesti fraze *Vučedolska kultura* koja se prevodi kao *cultura di Vučedol, Muzej uli*č*ne umjetnosti* preveden kao *Museo dell'atre stradale, ideja igranja* prevedena kao *idea di giocare*. U ovim slučajevima u prijevod je dodana jedna riječ više. Razlog tome je što u talijanskom jeziku ne postoje nastavci za padeže kao u hrvatskome jeziku. Zatim postoje primjeri gdje se izrazi ne mogu doslovno prevesti. Zbog toga je *folklor* preveden izrazom *danza popolare.*

Nakon što su ključne riječi bile prevedene bilo je potrebno datoteke u kojima se nalaze pretvoriti u .key datoteke. Razlog tome je da Maui metoda, u datoteci koja sadržava podatke za učenje, mora imati tekstove na kojima će algoritam učiti te ključne riječi u obliku .key datoteke. Dodatno, ključne riječi koje su prevodili različiti timovi bile su razvrstane u različite datoteke.

Kako bi se analiza kod Maui algoritma izvršila, potrebno je .key i .txt datoteke nazvati istim imenom. Ukoliko datoteke nemaju isti naziv, analiza se neće izvršiti, odnosno, kao izlaz dobit će se prazna datoteka.

S time je završila priprema za ekstrakciju ključnih riječi pomoću algoritma Maui. Nakon pripreme i prije početka ekstrakcije bilo je potrebno razvrstati podatke za obradu i dvije različite datoteke. Jedna datoteka sadržavala je podatke za učenje, a druga podatke za testiranje. U datoteci za učenje nalazilo se 35 datoteka s ključnim riječima koje je označio čovjek i 85 tekstova koji nisu u potpunosti obrađeni. U datoteci za testiranje nalazilo se 35 tekstova obrađenih od strane profesionalnog prevoditelja. Bitno je da podataka na kojima algoritam uči bude više od onih na kojima se algoritam testira.

Metodologija testiranja

Testiranje se izvodilo devet puta. Kod svakog novog testiranja u datoteci za učenje mijenjale su se .key datoteke na način da se svaki puta stave ključne riječi drugog tima. Deveto testiranje provodilo se s ključnim riječima koje su se već originalno nalazile u tekstovima, označene od strane autora.

5.2. Rezultati

5.2.1. Rake

Prikaz liste stop riječi

Datoteka s listom stop riječi za talijanski jezik sadrži 660 riječi. To su riječi koje nemaju nikakvo značenje za sadržaj teksta (Wikipedia,2018). Primjere nekih od tih riječi moguće je vidjeti u tablici.

Tablica 1 Lista stop riječi talijanskog jezika (Github, 2018)

ahimè	da	qualcuno	stessero	uno
ai	dagl	quale	stessi	uomo
al	dagli	quali	stessimo	va
alcuna	dai	qualunque	stesso	vai
alcuni	dal	quando	steste	vale
alcuno	dall	quanta	stesti	vari
all	dalla	quante	stette	varia
alla	dalle	quanti	stettero	varie
alle	dallo	quanto	stetti	vario
allo	dappertutto	quantunque	stia	verso
allora	davanti	quarto	stiamo	vi
altre	degl	quasi	stiano	via
altri	degli	quattro	stiate	vicino
altrimenti	dei	quel	sto	visto
altro	del	quella	su	vita
altrove	dell	quelle	sua	voi
altrui	della	quelli	subito	volta
anche	delle	quello	successivamente	volte
ancora	dello	quest	successivo	vostra
anni	dentro	questa	sue	vostre
anno	detto	queste	sugl	vostri
ansa	deve	questi	sugli	vostro
anticipo	devo	questo	sui	ã¨
assai	di	qui	sul	è

Rezultati analize tekstova u kojima su sadržane ključne riječi od strane autora

Primjer teksta

<figcaption> Alice Pedroletti- Senza Titolo </figcaption>
<figcaption> Contrazione- Scamnology </figcaption>
<figcaption> Ana Vuzadarić- Hertaforming </figcaption>
<figcaption> Mediterranea 16- Library </figcaption>
<figcaption> Virginia Zanetti- Walking on the Water Miracle und Utopia </figcaption>
Al fine, la manifestazione con le sue innumerevoli performance, concerti, presentazioni e laboratori e diventata un luoho d'incontro e di scambio d'idee, punti di vista diversi, esperienze e conoscenze condizionate dei diversi contesti della localita di provenienza dei partecipanti. Lo sforzo comune a lungo termine dei curatoti durante giorni d'apertura della mostra e sfociato nel dialogo con gli espositori, e ne gli errori anticipati nello titolo stesso di questo nuovo formato della Biennale dei giovani. Errori come un modo di prendere le distanze dagli schemi e standard tradizionali si sono di mostrati come il luogo che offre nuove possibilitá alla formazione collettiva ed individuale degli artisti in un periodo di crisi generale del campo artistico o della sua transizione in una nuovo, ancora che distaricolata forma.
I TAG
Ancona, il biennale, gli eventi, il mediterraneo, i giovani, l'arte

Slika 1 Primjer neočišćenog teksta pod brojem 4

Tablica 2 Ključne riječi koje je označio čovjek za tekst pod brojem 4

<table>
<tr>
<td>
4_tal_T1.txt | 4_tal_T2.txt
1 Ancona
2 biennale
3 arte
4 artisti giovani
5 Mole Vanvitelliana
Slika 2 Ključne riječi koje je označio 1. tim
</td>
<td>
4_tal_T1.txt | 4_tal_T2.txt | 4_tal
1 Mole Vanvitelliana
2 BJCEM
3 reti di solidarieta
4 prestazioni
5 presenza della persona
6 archeologia
7 esperienze collettive
8 Library
Slika 3 Ključne riječi koje je označio 2. tim
</td>
</tr>
<tr>
<td>
1 Ancona
2 arti
3 reti di solidarieta
4 Library
Slika 4 Ključne riječi koje je označio 3. tim
</td>
<td>
1 Ancona
2 artisti giovanni
3 biennale
4 curatore
5 Mediterranea 16
6 mediterraneo
7 mostre
8 arte
Slika 5 Ključne riječi koje je označio 4. tim
</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</table>

```
1  Ancona
2  relazioni libere
3  arte
4  Mediterranea
5  prestazioni
6  reti di solidarieta
7  attivismo
8  mostre
```

Slika 6 Ključne riječi koje je označio 5. tim

```
1  Ancona
2  biennale
3  arte
4  mostra
5  artisti giovani
```

Slika 7 Ključne riječi koje je označio 6. tim

```
1  Ancona
2  mostre
3  mediterranea 16
4  arte
5  curatori
6  sistema
7  rete di solidarieta
8  In Honor of
```

Slika 8 Ključne riječi koje je označio 7. tim

```
1  artista
2  Ancona
3  mostra
4  manifestazione
5  curatore
```

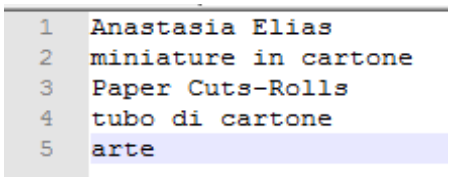
Slika 9 Ključne riječi koje je označio 9. tim

4-tal.txt | 4-tal_izlaz.txt

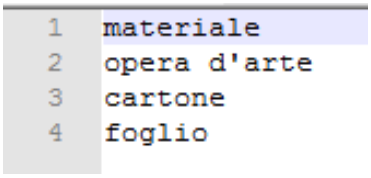
```
1  Keywords: [('ancona', 1.625), ('campo', 1.1666666666666667), ('biennale', 1.0)]
```

Slika 10 rezultati Rake algoritma za tekst pod brojem 4

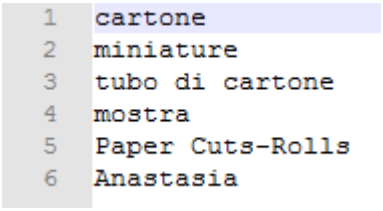
Tablica 3 Ključne riječi koje je označio čovjek za tekst pod brojem 30



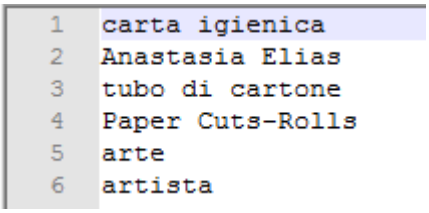
Slika 11 Ključne riječi koje je označio 1. tim



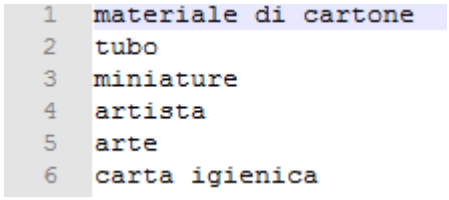
Slika 12 Ključne riječi koje je označio 2. tim



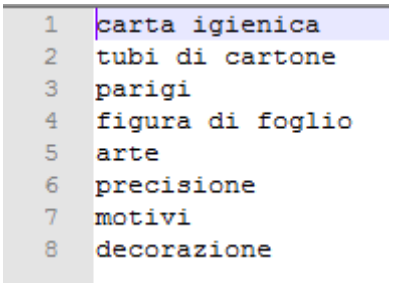
Slika 13 Ključne riječi koje je označio 3. tim



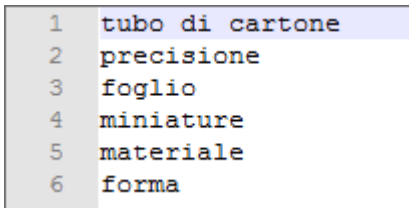
Slika 14 Ključne riječi koje je označio 4. tim



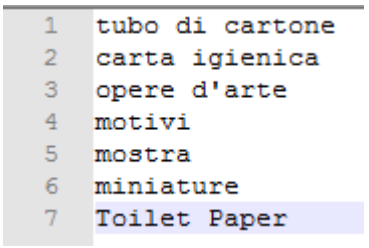
Slika 15 Ključne riječi koje je označio 5. tim



Slika 16 Ključne riječi koje je označio 6. tim



Slika 17 Ključne riječi koje je označio 7. tim



Slika 18 Ključne riječi koje je označio 8. tim

```
1  Keywords: [('cartone artistici </figcaption>', 7.909090909090909), (' </figcaption>
   esempi', 4.0), ('cartone', 1.9090909090909092), ('rotoli', 1.0)]
```

Slika 19 rezultati RAKE algoritma za neočišćeni tekst pod brojem 30

Rezultati analize tekstova očišćenih od ključnih riječi i dodatnih elemenata

Primjer teksta

```
21  Alice Pedroletti- Senza Titolo
22  Contrazione- Scamnology
23  Ana Vuzadarić- Hertaforming
24  Mediterranea 16- Library
25  Virginia Zanetti- Walking on the Water Miracle und Utopia
26 Al fine, la manifestazione con le sue innumerevoli performance, concerti, presentazioni e laboratori e diventata un luoho d'incontro e di scambio d'idee, punti
   di vista diversi, esperienze e conoscenze condizionate dei diversi contesti della localita di provenienza dei  partecipanti. Lo sforzo comune a lungo termine
   dei curatoti durante giorni d'apertura della mostra e sfociato nel dialogo con gli espositori, e ne gli errori anticipati nello titolo stesso di questo nuovo
   formato della Biennale dei giovani. Errori come un modo di prendere le distanze dagli schemi e standard tradizionali si sono di mostrati come il luogo che
   offre nuove possibilitá alla formazione collettiva ed individuale degli artisti in un periodo di crisi generale  del campo artistico o della sua transizione in
   una nuovo, ancora che distaricolata forma.
27
28
```

Slika 20 Primjer očišćenog teksta pod brojem 4

Rezultat

```
1  Keywords: [('ancona', 1.4285714285714286), ('campo', 1.1666666666666667), ('biennale', 1.0)]
```

Slika 21 Rezultati RAKE algoritma za očišćeni tekst pod brojem 4

```
1  Keywords: [('cartone artistici', 3.5), ('cartone', 1.5), ('rotoli', 1.0), ('esempi', 1,0)]
```

Slika 22 Rezultati RAKE algoritma za očišćeni tekst pod brojem 30

5.2.2. Maui

Rezultati ovisni o ključnim riječima prvog tima: usporedba čovjek - stroj

Tablica 4 Rezultati ovisni o ključnim riječima prvog tima

Maui_izlaz_1tim_81-tal.txt

```
1  Keyword: piazza 0.10415524367113067
2  Keyword: Superflex 0.01630373681848151
3  Keyword: TOPOTEK1 0.01630373681848151
4  Keyword: architettura 0.01630373681848151
5  Keyword: Rossa 0.012621379129297986
6  Keyword: parco 0.012621379129297986
7  Keyword: Superkilen 0.010643359459990948
8  Keyword: BIG 0.010643359459990948
```

Slika 23 Ključne riječi - RAKE algoritam za tekst pod brojem 81

81_tal_T1.txt

```
1  Superkilen
2  parco urbano
3  piazza rossa
4  BIG
5  Superflex
6  TOPOTEK1
```

Slika 24 Ključne riječi - studenti za tekst pod brojem 81

Rezultati ovisni o ključnim riječima drugog tima: usporedba čovjek - stroj

Tablica 5 Rezultati ovisni o ključnim riječima drugog tima

Maui_izlaz_tim2_81-tal.txt

```
1  Keyword: Copenhagen 0.016793171907817418
2  Keyword: Rossa 0.004796868161951809
3  Keyword: piazza Nera 0.004796868161951809
4  Keyword: Ă 0.004796868161951809
5  Keyword: creazione di 0.0043396885151423285
6  Keyword: parco 0.0043396885151423285
7  Keyword: piazza 0.0043396885151423285
8  Keyword: Piazza Rossa 0.003753501645421555
```

Slika 25 Ključne riječi - RAKE algoritam za tekst pod brojem 81

```
1  Copenhagen
2  piazza rossa
3  Norrebro
4  parco urbano
5  Superkilen
```

Slika 26 Ključne riječi - studenti za tekst pod brojem 81

Rezultati ovisni o ključnim riječima trećeg tima: usporedba čovjek - stroj

Tablica 6 Rezultati ovisni o ključnim riječima trećeg tima

Maui_izlaz_3tim_81-tal.txt

```
1  Keyword: piazza 0.11456484799752083
2  Keyword: creazione di 0.06881538623079816
3  Keyword: Piazza Rossa 0.056129418501280684
4  Keyword: architettura 0.03994088910879794
5  Keyword: piazza Nera 0.03151877606130664
6  Keyword: Rossa 0.028446579440722923
7  Keyword: Ă 0.018720467527547167
8  Keyword: parco 0.018588017196421343
```

Slika 27 Ključne riječi - RAKE algoritam za tekst pod brojem 81

```
1  progetto Superkilen
2  parco urbano
3  piazza rossa
4  piazza verde
5  nero
```

Slika 28 Ključne riječi - studenti za tekst pod brojem 81

Rezultati ovisni o ključnim riječima četvrtog tima: usporedba čovjek – stroj

Tablica 7 Rezultati ovisni o ključnim riječima četvrtog tima

Maui_izlaz_4tim_81-tal.txt

```
1  Keyword: piazza 0.11456484799752083
2  Keyword: creazione di 0.06881538623079816
3  Keyword: Piazza Rossa 0.056129418501280684
4  Keyword: architettura 0.03994088910879794
5  Keyword: piazza Nera 0.03151877606130664
6  Keyword: Rossa 0.028446579440722923
7  Keyword: Ã 0.018720467527547167
8  Keyword: parco 0.018588017196421343
```

Slika 29 Ključne riječi - RAKE algoritam za tekst pod brojem 81

```
1  progetto Superkilen
2  parco urbano
3  piazza rossa
4  quartiere di Norrebro
```

Slika 30 Ključne riječi - studenti za tekst pod brojem 81

Rezultati ovisni o ključnim riječima petog tima: usporedba čovjek - stroj

Tablica 8 Rezultati ovisni o ključnim riječima petog tima

Maui_izlaz_5tim_81-tal.txt

```
1  Keyword: progetto 0.18232510321097925
2  Keyword: creazione di 0.17283134991081683
3  Keyword: Piazza Rossa 0.10810455556825505
4  Keyword: Superkilen 0.10616887957063743
5  Keyword: parco 0.053048197276020226
6  Keyword: piazza 0.04491167753784088
7  Keyword: creazione 0.025566648391394213
8  Keyword: Superflex 0.01586155096543647
```

Slika 31 Ključne riječi - RAKE algoritam za tekst pod brojem 81

```
1  Superkilen
2  Copenhagen
3  piazza rossa
4  parco
5  progetto
6  BIG
7  Superflex
8  TOPOTEK1
```

Slika 32 Ključne riječi - studenti za tekst pod brojem 81

Rezultati ovisni o ključnim riječima šestog tima: usporedba čovjek – stroj

Tablica 9 Rezultati ovisni o ključnim riječima šestog tima

Maui_izlaz_6tim_81-tal.txt

```
1  Keyword: creazione di 0.12481947744973303
2  Keyword: Piazza Rossa 0.12003874579783576
3  Keyword: spazio 0.07726925207994104
4  Keyword: Copenhagen 0.07719782350851247
5  Keyword: architettura 0.07080399262523568
6  Keyword: Rossa 0.03689711658619572
7  Keyword: parco 0.031052707583891982
8  Keyword: piazza Nera 0.031014763645019244
```

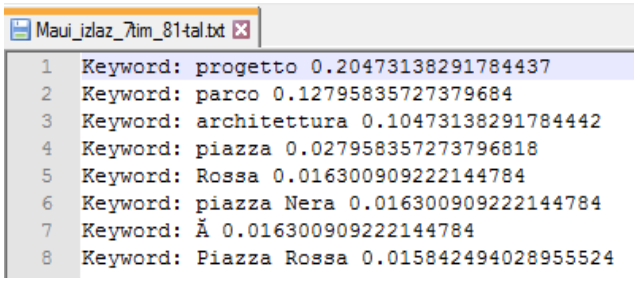
Slika 33 Ključne riječi - RAKE algoritam za tekst pod brojem 81

```
1  Copenhagen
2  Norrebro
3  parco urbano
4  piazza rossa
```

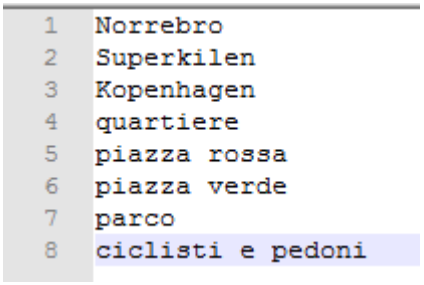
Slika 34 Ključne riječi - studenti za tekst pod brojem 81

Rezultati ovisni o ključnim riječima sedmog tima: usporedba čovjek - stroj

Tablica 10 Rezultati ovisni o ključnim riječima sedmog tima



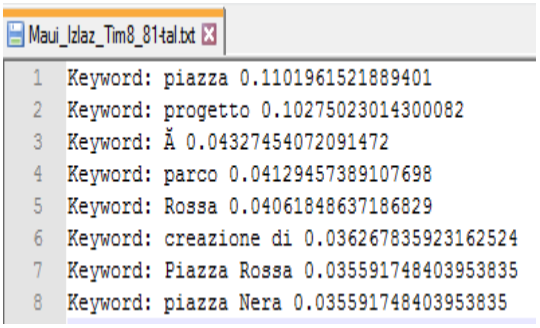
Slika 35 Ključne riječi - RAKE algoritam za tekst pod brojem 81



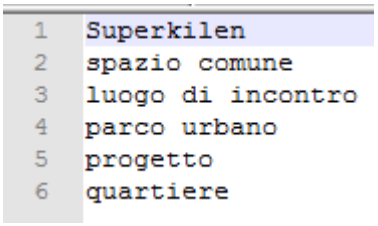
Slika 36 Ključne riječi - studenti za tekst pod brojem 81

Rezultati ovisni o ključnim riječima osmog tima: usporedba čovjek - stroj

Tablica 11 Rezultati ovisni o ključnim riječima osmog tima



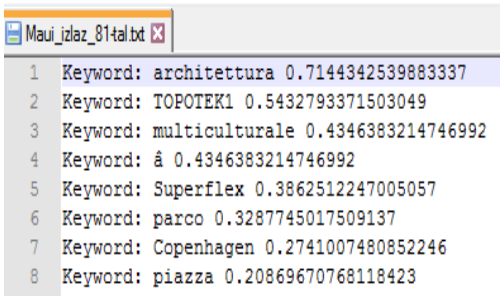
Slika 37 Ključne riječi - RAKE algoritam za tekst pod brojem 81



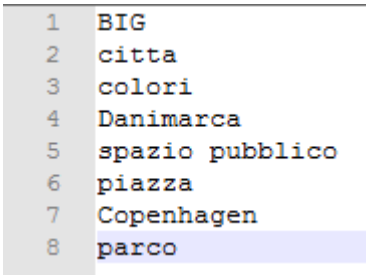
Slika 38 Ključne riječi - studenti za tekst pod brojem 81

Rezultati ovisni o ključnim riječima označenih od strane autora: usporedba čovjek – stroj

Tablica 12 Rezultati ovisni o ključnim riječima označenih od strane autora



Slika 39 Ključne riječi - RAKE algoritam za tekst pod brojem 81



Slika 40 Ključne riječi - studenti za tekst pod brojem 81

5.3. Usporedba rezultata

Tablica 13 Usporedba rezultata za tekst pod brojem 2

<table>
<tr><td>Čovjek</td><td>Maui</td></tr>
<tr><td>

```
1  Vučedol
2  capeggio
3  museo
4  architettura
5  cultura di Vučedol
6  progetto
```

Slika 41 Ključne riječi koje je označio student

</td><td>

```
1  Keyword: Vučedol 0.0891716364221803
2  Keyword: Museo 0.024504969755513653
3  Keyword: cultura di Vučedol 0.010643359459990948
4  Keyword: Ministero della cultura 0.010643359459990948
5  Keyword: architetti 0.010643359459990948
6  Keyword: pubblico 0.010643359459990948
7  Keyword: architettura 0.010643359459990948
8  Keyword: Goran 0.010080523550756405
```

Slika 42 Ključne riječi koje je označio algoritam Maui

</td></tr>
<tr><td colspan="2">RAKE – neočišćeni tekst</td></tr>
<tr><td colspan="2">

```
1  Keywords: [('vucedol', 4.714285714285714), ('progetto', 1.5), ('museo',
   2.2173913043478262), ('cultura'), 1.1428571428571428)]
```

Slika 43 Ključne riječi koje je označio algoritam RAKE kod neočišćenog teksta

</td></tr>
<tr><td colspan="2">RAKE – očišćeni tekst</td></tr>
<tr><td colspan="2">

```
1  Keywords: [('vucedol', 4.666666666666667), ('progetto', 1.5), ('museo',
   1.2083333333333333), ('vukovar', 1.2), ('cultura', 1.1428571428571428)]
```

Slika 44 Ključne riječi koje je označio algoritam RAKE kod očišćenog teksta

</td></tr>
</table>

U primjeru za tekst broj 2 u tablici 13 može se primijetiti da u svakom algoritmu ima nekih sličnih pojmova, a pojavljuju se i neki drugačiji. Riječi *Vučedol i museo* zajedničke su objema algoritmima, a nalaze se i u ključnim riječima koje su označili studenti.

Nadalje, kod algoritma Maui javlja se i ključna fraza *cultura di Vučedol* koju su i studenti označili kao ključnu frazu.

U primjeru prikazanom u tablici 13, vidi se da algoritam Maui i „čovjek“ imaju označeno više istih pojmova, odnosno njih 3 dok kod algoritma RAKE označena su samo dva ista pojma.

Razlika je u tome što algoritam Maui ima skup podataka na kojem uči te je zbog toga uspio izdvojiti i frazu poput *culutra di Vučedol,* dok je algoritam RAKE uspio izvući iz teksta samo riječ *cultura.*

Tablica 14 Usporedba rezultata za tekst pod brojem 5

<table>
<tr><td>Čovjek</td><td>Maui</td></tr>
<tr><td>

```
1  farmacia
2  aula
3  Branke Benčić
4  Vodnjan
5  Matija Debeljuh
6  mostre
7  arti cotemporanee
```

Slika 45 Ključne riječi koje je označio student

</td><td>

```
1  Keyword: Farmacia 0.11884459239702308
2  Keyword: arte 0.08097040348514818
3  Keyword: mostra 0.016303736818481 51
4  Keyword: Vodnjan 0.010643359459990948
5  Keyword: artista 0.010643359459990948
6  Keyword: mostre 0.010643359459990948
7  Keyword: galleria 0.010643359459990948
8  Keyword: museo 0.010643359459990948
```

Slika 46 Ključne riječi koje je označio algoritam Maui

</td></tr>
<tr><td colspan="2">RAKE – neočišćeni tekst</td></tr>
<tr><td colspan="2">

```
1  Keywords: [('farmacia', 1.5714285714285714),('spazio', 1,41666666666666667),('serie', 1.2)]
```

Slika 47Ključne riječi koje je označio algoritam RAKE kod neočišćenog teksta

</td></tr>
<tr><td colspan="2">RAKE – očišćeni tekst</td></tr>
<tr><td colspan="2">

```
1  Keywords: [('farmacia', 1.5333333333333334),('spazio', 1,41666666666666667),('serie', 1.2)]
```

Slika 48 Ključne riječi koje je označio algoritam RAKE kod očišćenog teksta

</td></tr>
</table>

Na primjeru u tablici 14, kao i u prethodnom primjeru možemo vidjeti da ključne riječi koje je označio student imaju više zajedničkog s algoritmom Maui, a ne s algoritmom RAKE. Odnosno rezultat je 2:1 u korist algoritma Maui.

U ovom slučaju niti jedan algoritam nije označio više fraza, već je svaki od njih označio po samo jednu ključnu riječ.

I u primjeru u tablici 13 i u tablici 14, možemo vidjeti promjenu kod koeficijenata koji slijede ključne riječi kod algoritma RAKE. Za to je zaslužno čišćenje teksta od nevažnih pojmova koji su se pojavljivali u tekstovima.

Tablica 15 Usporedba rezultata za tekst pod brojem 12

<table>
<tr><th>Čovjek</th><th>Maui</th></tr>
<tr><td>

```
1  fotografia analogica
2  aTree
3  Sara Perović
4  fanzine
5  MoMA
6  Pola
```

Slika 49 Ključne riječi koje je označio student

</td><td>

```
1  Keyword: progetto 0.0785894511041958
2  Keyword: fanzine 0.024504969755513653
3  Keyword: Pola 0.024504969755513653
4  Keyword: aTree 0.018844592397023092
5  Keyword: MoMA 0.01630373681848151
6  Keyword: art 0.01630373681848151
7  Keyword: Sara Perović 0.010643359459990948
8  Keyword: fotografia analogica 0.010643359459990948
```

Slika 50 Ključne riječi koje je označio algoritam Maui

</td></tr>
<tr><td colspan="2">RAKE – neočišćeni tekst</td></tr>
<tr><td colspan="2">

```
1  Keywords: [('progetto', 1.6363636363636365),('fotofrafia', 1.625),  ('fanzine',
   1.3636363636363635), ('atree', 1.3), (internet, 1,0)]
```

Slika 51 Ključne riječi koje je označio algoritam RAKE kod neočišćenog teksta

</td></tr>
<tr><td colspan="2">RAKE – očišćeni tekst</td></tr>
<tr><td colspan="2">

```
1  Keywords: [('fotofrafia', 1.625), ('progetto', 1.5833333333333333), ('fanzine',
   1.3333333333333333), ('atree', 1.3), (internet, 1,0)]
```

Slika 52 Ključne riječi koje je označio algoritam RAKE kod očišćenog teksta

</td></tr>
</table>

Na primjeru u tablici 15, vidljivo je da algoritam Maui ponovno ima više zajedničkih riječi i fraza nego algoritam RAKE. Algoritam Maui u ovom slučaju označio je 4 ista pojma, algoritam RAKE samo 2 ista pojma.

Kao i u prethodnim primjerima i u ovom se ističe da algoritam RAKE uvijek izdvaja po samo jednu riječ kao ključnu, dok algoritam RAKE izdvaja i fraze poput „*fotografia analogica*“.

A isto tako ponovno se mijenja i koeficijent koji se pojavljuje uz ključnu riječ, nakon što je tekst očišćen od viška znakova unutar tekstova.

Tablica 16 Usporedba rezultata za tekst pod brojem 90

<table>
<tr><th>Čovjek</th><th>Maui</th></tr>
<tr><td>

```
1  conferenza
2  architetti
3  premio Piranesi
4  progetti
```

Slika 53 Ključne riječi koje je označio student

</td><td>

```
1  Keyword: Piranesi 0.2041552436711307
2  Keyword: gli 0.03775347952559917
3  Keyword: premi 0.018844592397023092
4  Keyword: progetto 0.013922784437529135
5  Keyword: facoltÃ 0.012621379129297986
6  Keyword: Pirano 0.012621379129297986
7  Keyword: Premio Piranesi 0.010643359459990948
8  Keyword: conferenza 0.010643359459990948
```

Slika 54 Ključne riječi koje je označio algoritam Maui

</td></tr>
<tr><th colspan="2">RAKE – neočišćeni tekst</th></tr>
<tr><td colspan="2">

```
1  Keywords: [('progetto', 1.3333333333333333), ('architettura', 1.2), ('pirano',
   1.1428571428571428), ('giornate', 1.0), ('piran', 1.0), ('premi', 1.0)]
```

Slika 55 Ključne riječi koje je označio algoritam RAKE kod neočišćenog teksta

</td></tr>
<tr><th colspan="2">RAKE – očišćeni tekst</th></tr>
<tr><td colspan="2">

```
1  Keywords: [('progetto', 1.2857142857142858), ('facilita'', 1.2), ('architettura',
   1.1818181818181819), ('pirano', 1.125), ('giornate', 1.0), ('piran', 1.0), ('premi', 1.0)]
```

Slika 56 Ključne riječi koje je označio algoritam RAKE kod očišćenog teksta

</td></tr>
</table>

Na primjeru u tablici 16 vidljivo je da algoritam RAKE i student imaju označeno više zajedničkih ključnih riječi. Iako je to samo jedna riječ u korist algoritma RAKE , algoritam Maui i čovjek u ovom slučaju nemaju označene iste pojmove.

Čovjek i algoritam Maui označili su pojam *premio Piranesi* kao ključnu riječ, a kod algoritma RAKE možemo primijetiti da često odabire slične riječi, tj. riječi koje su slične onima koje je označio čovjek. Primjerice *pirano, premi, piran.*

U ovom primjeru kod algoritma RAKE može se zamijetiti da se nakon čišćenja teksta promijenio red ključnih riječi, promijenio se i broj koji slijedi nakon ključne riječi, a dodane su i označene neke nove riječi kao ključna riječ.

Tablica 17 Usporedba rezultata za tekst pod brojem 26

<table>
<tr><th>Čovjek</th><th>Maui</th></tr>
<tr><td>

```
1  professione
2  politica
3  monumento
4  terra croata
5  padiglione
```

Slika 57 Ključne riječi koje je označio student

</td><td>

```
1  Keyword: politica 0.08097040348514818
2  Keyword: monumento 0.024504969755513653
3  Keyword: arte 0.01630373681848151
4  Keyword: colonne 0.012621379129297986
5  Keyword: veterani 0.012621379129297986
6  Keyword: moderna 0.010643359459990948
7  Keyword: padiglione 0.010643359459990948
8  Keyword: Croazia 0.010080523550756405
```

Slika 58 Ključne riječi koje je označio algoritam Maui

</td></tr>
<tr><th colspan="2">RAKE – neočišćeni tekst</th></tr>
<tr><td colspan="2">

```
1  Keywords: [('veterani', 1.5), ('giuria', 1.0), ('monumento', 1.0)]
```

Slika 59 Ključne riječi koje je označio algoritam RAKE kod neočišćenog teksta

</td></tr>
<tr><th colspan="2">RAKE – očišćeni tekst</th></tr>
<tr><td colspan="2">

```
1  Keywords: [('veterani', 1.4285714285714286), ('giuria', 1.0), ('monumento', 1.0)]
```

Slika 60 Ključne riječi koje je označio algoritam RAKE kod očišćenog teksta

</td></tr>
</table>

Primjer u tablici 17 samo pokazuje sličan slijed kao i prethodni primjeri. Algoritam Maui i čovjek imaju označena tri ista pojma, a algoritam RAKE i čovjek samo jedan isti pojam.

Nakon čišćenja teksta od neželjenih elemenata, promijenili su se koeficijenti nakon ključnih riječi kod algoritma RAKE.

Zajedničke riječi u ovom slučaju su *politica, monumento i padiglione* kod algoritma Maui i čovjeka, a kod algoritma RAKE i čovjeka ista je samo riječ *monumento.*

Tablica 18 Usporedba rezultata za tekst pod brojem 29

<table>
<tr><th>Čovjek</th><th>Maui</th></tr>
<tr><td>

```
1  collettivo artistico Voina
2  incubo
3  FBS
4  innovazione
```

Slika 61 Ključne riječi koje je označio student

</td><td>

```
1  Keyword: di 0.10994772142592242
2  Keyword: politica 0.01630373681848151
3  Keyword: arte 0.01630373681848151
4  Keyword: Banksy 0.012621379129297986
5  Keyword: russo 0.012621379129297986
6  Keyword: membri 0.012621379129297986
7  Keyword: Voina 0.012621379129297986
8  Keyword: Ã 0.012488577004463985
```

Slika 62 Ključne riječi koje je označio algoritam Maui

</td></tr>
<tr><td colspan="2">RAKE – neočišćeni tekst</td></tr>
<tr><td colspan="2">

```
1  Keywords: [('membri', 1.0)]
```

Slika 63 Ključne riječi koje je označio algoritam RAKE kod neočišćenog teksta

</td></tr>
<tr><td colspan="2">RAKE – očišćeni tekst</td></tr>
<tr><td colspan="2">

```
1  Keywords: [('membri', 1.0)]
```

Slika 64 Ključne riječi koje je označio algoritam RAKE kod očišćenog teksta

</td></tr>
</table>

Na primjeru u tablici 18 može se vidjeti da postoje i rezultati u kojima ne postoji niti jedna zajednička riječ ili fraza.

U ovom slučaju niti algoritam RAKE niti algoritam Maui nisu označili niti jednu isti ključnu riječ kao student.

Može se i primijetiti da se kod algoritma Maui pojavljuju i stop riječi *di* i *Ã* što nije slučaj kod algoritma RAKE. Razlog tome je što algoritam RAKE u svojoj obradi zahtjeva i listu stop riječi te ih on neće niti izdvojiti kao ključne riječi, dok je to moguća opcija kod algoritma Maui.

6. Evaluacija

Evaluacija ekstrakcije ključnih riječi obično se provodi u terminima precison, recall i F1 rezultata.

Kada želimo usporediti zapis koji je napravio čovjek i zapis koji je dobiven strojnim putem, precision se računa kao presjek ključnih riječi označenih od strane čovjeka (A) i ključnih riječi označenih od strane stroja (B). Dobiveni broj potrebno je podijeliti s brojem ključnih riječi koje je označio stroj (B) (Beliga i sur., 2018).

Zatim, recall se izračunava kao presjek ključnih riječi označenih od strane čovjeka (A) i ključnih riječi označenih od strane stroja (B). Dobiveni broj potrebno je podijeliti s brojem ključnih riječi koje je označio čovjek (B) (Beliga i sur., 2018).

Na kraju se računa rezultat F1 koji je harmonijska sredina precisona i recalla (Beliga i sur., 2018).

Formule:

Precision (Beliga i sur., 2018):

$$P = \frac{|A \cap B|}{|B|}$$

Recall (Beliga i sur., 2018):

$$R = \frac{|A \cap B|}{|A|}$$

F1 (Beliga i sur., 2018):

$$F1 = \frac{2PR}{P + R}$$

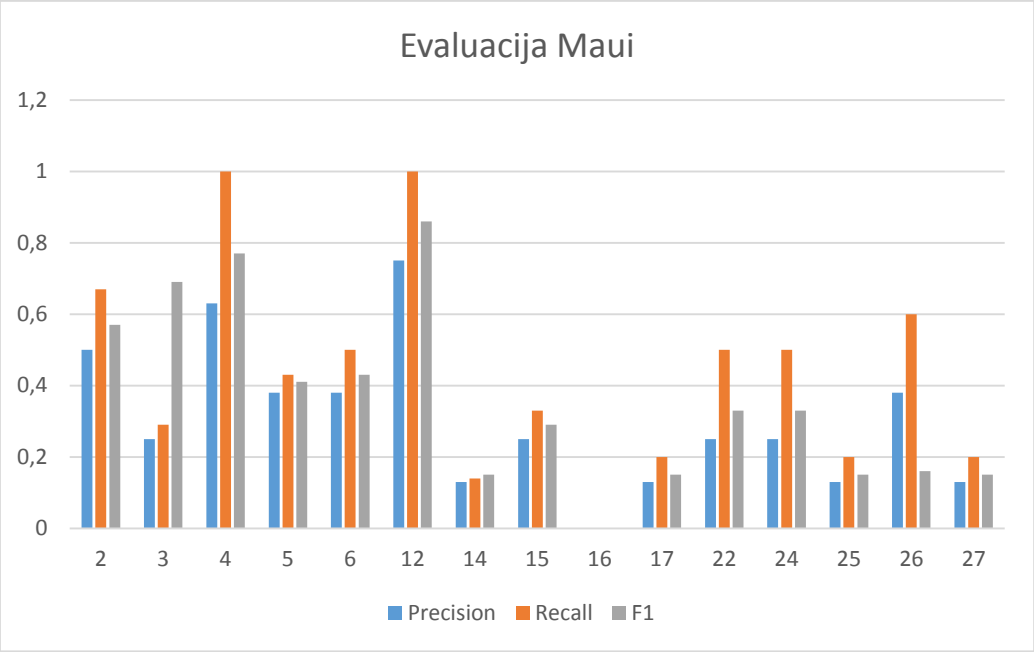
6.1. Principi

Čovjek – Maui

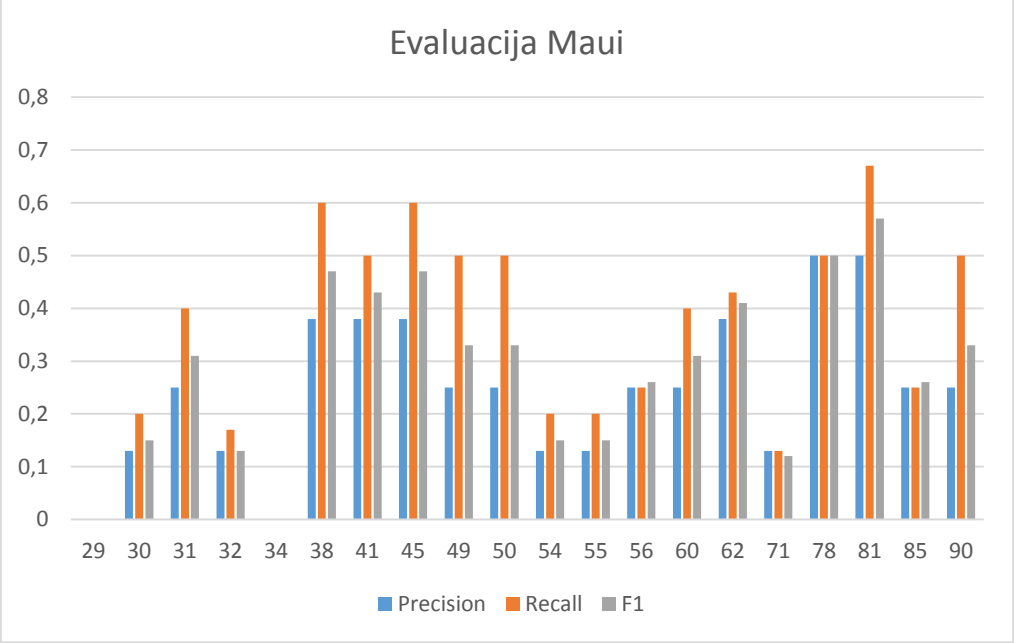
Tablica 19 Prikaz rezultata za algoritam Maui u usporedbi s čovjekom

Broj teksta	Čovjek je označio	Stroj je označio	Zajedničke riječi (presjek)	Precison	Recall	F1
2	6	8	4	0,5	0,67	0,57
3	7	8	2	0,25	0,29	0,69
4	5	8	5	0,63	1	0,77
5	7	8	3	0,38	0,43	0,41
6	6	8	3	0,38	0,5	0,43
12	6	8	6	0,75	1	0,86
14	7	8	1	0,13	0,14	0,15
15	6	8	2	0,25	0,33	0,29
16	5	8	0	0	0	0
17	5	8	1	0,13	0,2	0,15
22	4	8	2	0,25	0,5	0,33
24	4	8	2	0,25	0,5	0,33
25	5	8	1	0,13	0,2	0,15
26	5	8	3	0,38	0,6	0,16
27	5	8	1	0,13	0,2	0,15
29	4	8	0	0	0	0
30	5	8	1	0,13	0,2	0,15
31	5	8	2	0,25	0,4	0,31
32	6	8	1	0,13	0,17	0,13
34	5	8	0	0	0	0
38	5	8	3	0,38	0,6	0,47
41	6	8	3	0,38	0,5	0,43
45	5	8	3	0,38	0,6	0,47
49	4	8	2	0,25	0,5	0,33
50	4	8	2	0,25	0,5	0,33
54	5	8	1	0,13	0,2	0,15

55	5	8	1	0,13	0,2	0,15
60	5	8	2	0,25	0,4	0,31
62	7	8	3	0,38	0,43	0,41
71	8	8	1	0,13	0,13	0,12
78	8	8	4	0,5	0,5	0,5
81	6	8	4	0,5	0,67	0,57
85	8	8	2	0,25	0,25	0,26
90	4	8	2	0,25	0,5	0,33



Slika 65 Grafikon koji prikazuje mjere precision, recall i F1 od teksta 2 do teksta 27 (čovjek-Maui)



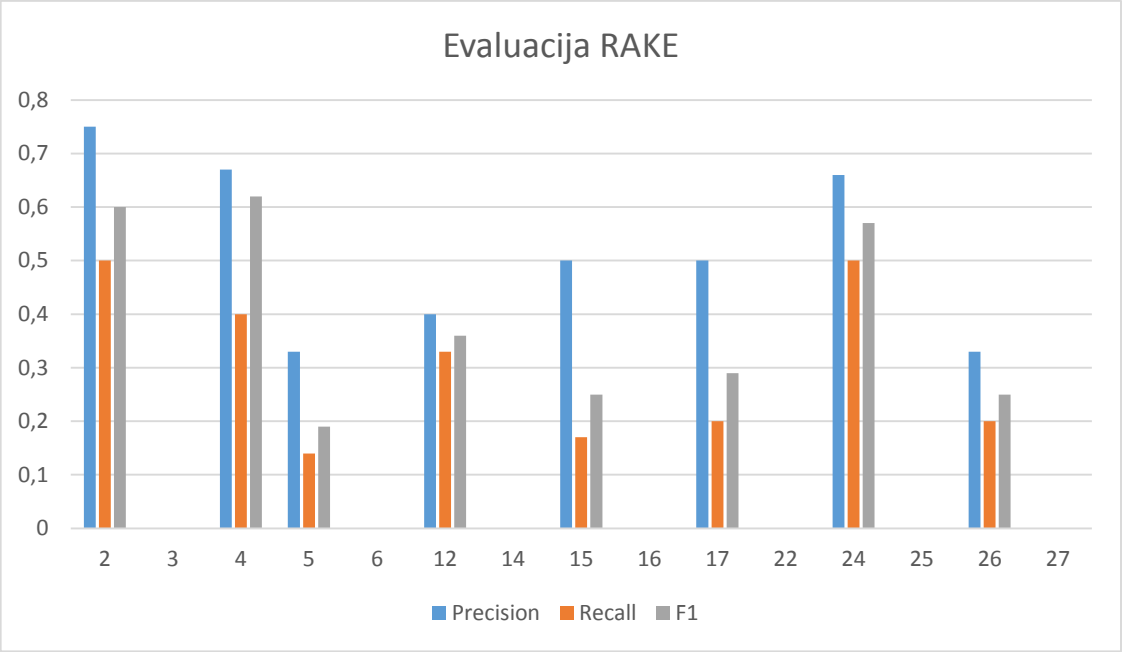
Slika 66 Grafikon koji prikazuje mjere precision, recall i F1 od teksta 29 do teksta 90 (čovjek-Maui)

Čovjek – RAKE

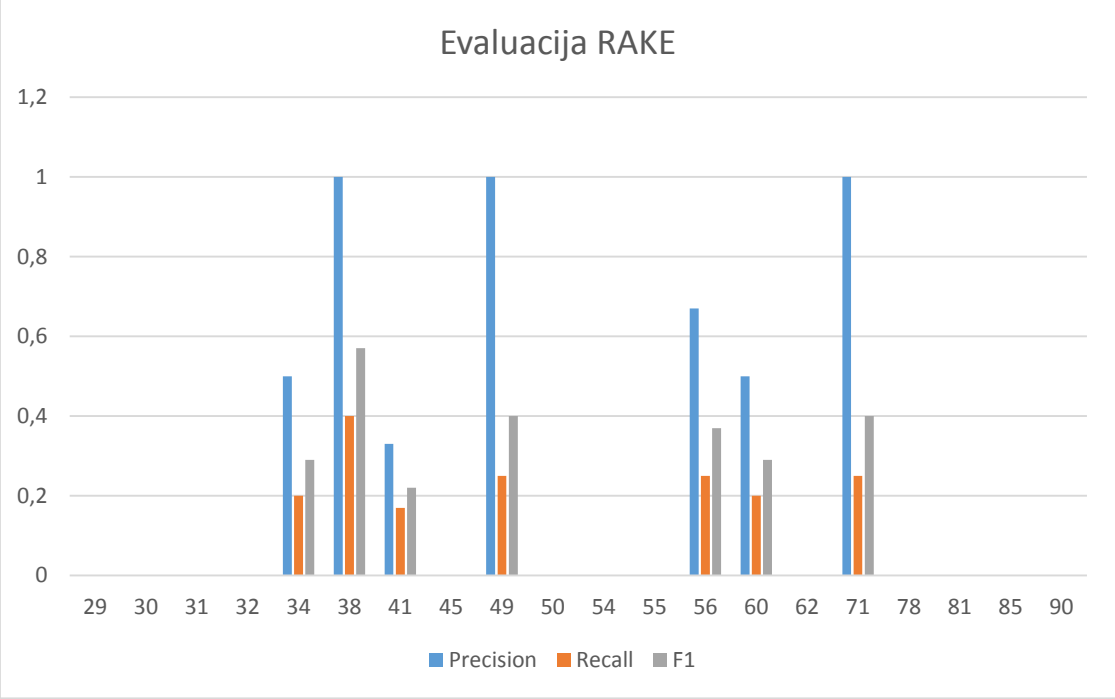
Tablica 20 Prikaz rezultata za algoritam RAKE u usporedbi s čovjekom

Broj teksta	Čovjek je označio	Stroj je označio	Zajedničke riječi (presjek)	Precison	Recall	F1
2	6	4	3	0,75	0,5	0,6
3	7	0	0	0	0	0
4	5	3	2	0,67	0,4	0,62
5	7	3	1	0,33	0,14	0,19
6	6	0	0	0	0	0
12	6	5	2	0,4	0,33	0,36
14	7	1	0	0	0	0
15	6	2	1	0,5	0,17	0,25
16	5	1	0	0	0	0
17	5	2	1	0,5	0,2	0,29
22	4	0	0	0	0	0
24	4	3	2	0,66	0,5	0,57
25	5	6	0	0	0	0
26	5	3	1	0,33	0,2	0,25
27	5	2	0	0	0	0
29	4	1	0	0	0	0
30	5	4	0	0	0	0
31	5	1	0	0	0	0
32	6	2	0	0	0	0
34	5	2	1	0,5	0,2	0,29
38	5	2	1	1	0,4	0,57
41	6	3	1	0,33	0,17	0,22
45	5	0	0	0	0	0
49	4	1	1	1	0,25	0,4
50	4	3	0	0	0	0
54	5	0	0	0	0	0
55	5	4	0	0	0	0
56	8	3	2	0,67	0,25	0,37

60	5	2	1	0,5	0,2	0,29
62	7	6	0	0	0	0
71	8	2	2	1	0,25	0,4
78	8	4	0	0	0	0
81	6	0	0	0	0	0
85	8	1	0	0	0	0
90	4	6	0	0	0	0



Slika 67 Grafikon koji prikazuje mjere precision, recall i F1 od teksta 2 do teksta 27 (čovjek-RAKE)



Slika 68 Grafikon koji prikazuje mjere precision, recall i F1 od teksta 29 do teksta 90 (čovjek-RAKE)

7. Diskusija

Ekstrakcija ključnih riječi provodila se pomoću dva algoritma. Svaki od njih trebao je slične podatke kako bi izvršio ekstrakciju, no u nekim dijelovima algoritmi se i razlikuju.

Kod usporedbe rezultata algoritama RAKE i Maui s ključnim riječima koje su označili studenti, možemo primijetiti konstantnu šablonu koja se pojavljuje tokom označavanja. Neki rezultati neće imati zajedničke ključne riječi, neki će imati jednu ili dvije, a najviše do šest zajedničkih ključnih riječi.

Primjećuje se da algoritam Maui uvijek ima označeno više istih ključnih riječi kao i studenti. Razlog tome je što algoritam Maui uči na određenom skupu podataka te ima prilike „vidjeti" koje riječi bi bile bolji kandidati za ključne riječi. Ovaj primjer govori u korist algoritma Maui, no ima i on nekih mana. Primjerice, problem kod algoritma Maui je što u nekim slučajevima kao ključne riječi označava stop riječi.

To se ne događa kod algoritma RAKE. Ovaj algoritam neće označavati stop riječi kao ključne riječi jer u svojoj biblioteci sadržava listu stop riječi. Ova situacija ide u korist RAKE algoritma, no kao što se može primijetiti algoritam RAKE uvijek označi manje istih ključnih riječi kao što ih je označio čovjek. No, je li to nužno loše?

Navedimo neke primjere koje je označio čovjek i neke koje je označio algoritam Maui. U tekstu pod brojem dva javljaju se pojmovi *cultura di Vučedol* kojeg je označio student i *cultura* kojeg je označio algoritam. Moguće je vidjeti da je algoritam RAKE na tragu onog što je označio čovjek. Slični primjeri vidljivi su u tablici.

Slično se dešava i kod algoritma Maui. Kao i algoritam RAKE, algoritam Maui često je označio slične riječi kao i studenti, no iz razloga što je student označio cijelu frazu, sličan pojam nije mogao biti priznat kao „točno" označena ključna riječ.

Tablica 21 Usporedba MAUI - student

RAKE	ČOVJEK	BROJ TEKSTA
Sedia	*Poltrona*	Tekst 31
Rijeka	*Progetto di magliette di Rijeka*	Tekst 32
Negozo	*Negozio di design*	Tekst 56
Ministero, cultura	*Ministero della cultura*	Tekst 62

Tablica 22 Usporedba MAUI - student

Maui	ČOVJEK	BROJ TEKSTA
Jeane	Jeanne Claude	Tekst 3
Voina	collettivo artistico Voina	Tekst 29
cartone	tubo di cartone	Tekst 30
Progetto, magliette	progetto di magliette di Rijeka	Tekst 32
Canvas, comunity	Canvas comunity	Tekst 54
Piazza, rossa	Piazza rossa	Tekst 81
Progetto	progetti	Tekst 90

8. Zaključak

Ekstrakcija ključnih riječi nedavno se javila kao novo istraživanje i bitan je dio perioda u kojem živimo. Prije su indeksiranje, klasifikacija tekstova i IR bili potrebni samo na mjestima gdje se pojavljivalo mnogo dokumenta i podataka kao što su knjižnice, udruge i različite organizacije. Tim poslovima bavile su se stručne osobe. Dolaskom velike količine tekstualnih podataka u big-data eri pojavila se i potreba za razvijanjem algoritama koji će obavljati prije navedene procese kako bi se ljudima olakšao posao.

Kroz ovaj rad istražila su se dva algoritma za ekstrakciju ključnih riječi, a to su algoritam RAKE i algoritam Maui. Oba algoritma imaju pozitivne i negativne strane te se korisnik mora opredijeliti sam kojeg će koristiti i hoće li ga koristiti. Može odabrati ovisno o vlastitim potrebama. Kod odabira, mora imati na umu da za algoritam Maui treba imati određeni skup podataka na kojem će algoritam prvo trenirati i učiti, a kasnije treba imati i skup podataka na kojem će algoritam raditi testiranje. Nadalje, algoritam RAKE je jednostavniji za korištenje jer kod njegove analize potrebno je pronaći samo listu stop riječi za jezik na kojem se želi raditi analiza. Olakšava situaciju i činjenica da prilikom postavljanja RAKE algoritma već postoji datoteka koja sadrži listu stop riječi za španjolski, engleski i francuski jezik. U ovom slučaju datoteka s stop riječima za talijanski preuzeta je s interneta i stavljena je u postojeću datoteku.

Smatram da bi u daljnjem radu bilo potrebno pronaći još veći skup podataka na kojem će algoritmi moći učiti. Isto tako bilo bi dobro na neki način ujediniti algoritme. Na taj način izbjeglo bi se da algoritam Maui označuje stop riječi kao ključne riječi, a postiglo bi se da algoritam RAKE uspije označiti i neke fraze kao ključne riječi, a ne samo jednu zasebnu riječ.

9. Bibliografija

1) Beliga, S., Meštrović, A., i Martinčić-Ipšić, S. (2018). Lytras, M., Aljohani, N., Damiani, E., i Tai Chui, K. *Innovations, Developments, and Applications of Semantic Web and Information Systems* (str. 170-197). United States od America: IGI Global.
2) Bošnjak, R. (2017). *Usporedba metoda za klasifikaciju tekstualnih dokumenata* (Završni rad). Preuzeto s https://bib.irb.hr/datoteka/891152.Final_0036485149_42.pdf
3) James, G., Witten, D., Hastie, T., Tibishirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer. Preuzeto s: https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf
4) Medelyan, O. (2009). *Human-competitive automatic topic indexing.* New Zealand: The Universiti of Waikato.
5) Airpair (2018). *NLP keyword extraction*. Preuzeto 2.5.2018 s https://www.airpair.com/nlp/keyword-extraction-tutorial
6) Centrum für informations und sprachverarbeitung (2018). TreeTagger - a part-of-speech tagger for many languages. Preuzeto 14.5.2018. s http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
7) GitHub (2018). *Stopwords.* Preuzeto 20.7.2018 s https://github.com/stopwords-iso/stopwords-it/blob/master/stopwords-it.txt
8) Hackage (2018). *Rake*. Preuzeto 24.8.2018 s http://hackage.haskell.org/package/rake
9) Italian Natural Language Processing Lab (2018). *LinguA.* Preuzeto 27.4.2018 s http://www.italianlp.it/demo/
10) Istituto di Linguistica Computazionale (2018). *DyLan TextTools*. Preuzeto 27.4.2018 s http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest
11) Text-processing (2018). *Stemming and Lemmatization with Python NLTK.* Preuzeto 14.5.2018 s http://text-processing.com/demo/stem/
12) Wikipedia (2018). *Indeksiranje.* Preuzeto: 2.8.2018 s https://hr.wikipedia.org/wiki/Indeksiranje
13) 9ol (2018). *Javascript Porter Stemmer Online.* Preuzeto 14.5.2018 s http://9ol.es/porter_js_demo.html

10. Privitak

10.1. Rake

Potrebno je pozicionirati se u comand prompt, nakon čega je potrebno prijaviti se kao administrator. Zatim slijedi pozicioniranje u datoteku u kojoj se nalaze tekstovi za analizu i sve je spremno za rad.

/*Preuzeto: (Airpair, 2018)*/

```
Python skupup.py install
from nlp_rake import rake
stoppath = 'data/stoplists/stopwords-it.txt'
rake_object = rake.Rake(stoppath, 5, 3, 4)
sample_file = open("data/docs/fao_test_italian/2-tal.txt",
'r', encoding="iso-8859-1")
text = sample_file.read()
keywords = rake_object.run(text)
# 3. print results
print("Keywords:", keywords)
# ispis u file
fileOut = open('data/stoplists/izlaz.txt','r+')
for word in keywords:
     fileOut.write(word)
fileOut.close()
```

10.2. Maui

Potrebno je pozicionirati se u comand prompt, nakon čega je potrebno prijaviti se kao administrator. Zatim slijedi pozicioniranje u datoteku u kojoj se nalaze tekstovi za analizu i sve je spremno za rad.

/*Preuzeto: (Airpair, 2018)*/

```
java -Xmx1024m -jar maui-standalone-1.1-SNAPSHOT.jar train -l
data/docs/fao_train_italian/ -m
data/models/keyword_extraction_model -v none -o 2
java -Xmx1024m -jar maui-standalone-1.1-SNAPSHOT.jar run
data/docs/fao_test_italian/2-tal.txt -m
data/models/keyword_extraction_model -v none -n 8
```

Popis slika