Sveučilište u Rijeci
University of Rijeka
http://www.uniri.hr

Polytechnica: Journal of Technology Education, Volume 4, Number 1 (2020)
Politehnika: Časopis za tehnički odgoj i obrazovanje, Volumen 4, Broj 1 (2020)

Politehnika
Polytechnica
http://www.politehnika.uniri.hr
cte@uniri.hr

DOI: https://doi.org/10.36978/cte.4.1.1

Izvorni znanstveni članak
Original scientific paper
UDK 004.912

Automatska ekstrakcija ključnih riječi iz teksta standardnim računalnim postupcima

Slobodan Beliga

Sveučilište u Rijeci, Odjel za informatiku

Sveučilište u Rijeci, Centar za umjetnu inteligenciju i kibernetičku sigurnost

Laboratorij za obradu prirodnog govora i jezika

Radmile Matejčić 2, 51 000 Rijeka, Hrvatska

sbeliga@inf.uniri.hr

Sažetak

Automatska ekstrakcija ključnih riječi iz teksta aktualan je istraživački problem u području računalne analize prirodnog jezika i pretraživanja informacija. Iako su razvijene brojne metode za ekstrakciju ključnih riječi iz teksta, njihova učinkovitost ovisna je o brojnim faktorima poput pristupa kojim su konstruirane, domene na koju su prilagođene, vrste jezika ili zadataka za koji su konstruirane i sl., a samim time prostor za napredak u smislu nadogradnje i poboljšanja, uvijek postoji. U ovom radu objašnjene su i rekonstruirane dvije postojeće metode – RAKE i MAUI, a koje su standardni predstavnici nenadzirane i nadzirane skupine metoda. Eksperimentalno je ispitano mogu li metode uspješno ekstrahirati ključne riječi iz tekstova pisanih na talijanskom jeziku, na kojem do sada nisu usporedno testirane. Za potrebe eksperimenta prikupljeni su i ručno označeni talijanski tekstovi. Efikasnost MAUI metode pokazala se perspektivnijom u odnosu na RAKE metodu što je već ranije potvrđeno u eksperimentu ekstrakcije ključnih riječi iz tekstova pisanih na engleskom jeziku.

Ključne riječi: *automatska ekstrakcija ključnih riječi; RAKE; MAUI; talijanski jezik.*

1 Uvod

Zadatak automatske ekstrakcije ključnih riječi ima za cilj u zadanom tekstu identificirati skup riječi koje sažeto, jezgrovito i kompaktno opisuju tematiku teksta (Beliga, Meštrović, Martinčić-Ipšić, 2014, 2018; Mihalcea, Tarau, 2004). Postupci ekstrakcije ključnih riječi učestalo se koriste i okosnica su u brojnim aplikacijama za računalnu analizu prirodnog jezika, poput automatskog indeksiranja pojmova ili dokumenata, grupiranja ili razvrstavanja dokumenata, pretraživanja i dohvaćanja informacija, automatskog sažimanja tekstova i sl.

S obzirom na radnu okolinu u kojoj djeluju, metode za ekstrakciju ključnih riječi u grubo dijelimo u tri osnovne skupine, a to su nadzirane (engl. *supervised*), nenadzirane (engl. *unsupervised*) i polunadzirane (engl. *semi-supervised*) metode. Osim takve grube podjele, suvremeni pristupi često metode dijele i na statističke metode, lingvističke metode, metode bazirane na strojnom učenju, neuralnim mrežama, grafovima i sl. (Beliga, 2019; Beliga, Meštrović, Martinčić-Ipšić, 2015; Merrouni, Frikh, Ouhbi, 2016). Nadzirane metode zahtijevaju unaprijed definirane ključne riječi na temelju kojih uče svoj model. Za nenadzirane metode to nije potrebno, one izlučuju ključne riječi neposredno iz zadanog teksta bez potrebe da uče na prethodno definiranim ključnim riječima za pojedini tekst.

Konačno, treća podjela metoda za ekstrakciju ključnih riječi, dijeli metode na osnovi složenosti zadatka koje obavlja. Naime, u jednostavnijoj varijanti metoda ekstrahira riječi iz svakog pojedinog teksta zasebno, dok u složenijoj varijanti, metoda ekstrahira ključne riječi iz velike kolekcije dokumenata istovremeno (primjerice cijele kolekcije tekstova pojedine domene, svih web stranica pojedinog web sjedišta i sl.) te skup dokumenata promatra kao jedan dokument. Razlika u dobivenom rezultatu takvih ekstrakcija ključnih riječi je u tome što u jednostavnijoj varijanti metoda na izlazu za svaki dokument ponudi zasebne ključne riječi (zadatak ekstrakcije usmjeren je na dokument), dok u složenijem zadatku za cijelu kolekciju tekstova ponudi jednu listu ključnih riječi (zadatak ekstrakcije usmjeren je na kolekciju).

U ovome radu ispitat će se i usporediti dvije standardne metode, od kojih jedna pripada skupini nadziranih, a druga skupini nenadziranih metoda. Na taj način suprotstavit će se i ispitati dva oprečna pristupa u istome zadatku – ekstrakciji ključnih riječi iz zadanog ulaznog teksta. Istraživanja takve prirode već su poznata od ranije. Međutim, u ovom istraživanju važno je napomenuti da će metode biti ispitane na podacima na kakvima do sada nisu bile ispitivane. Odnosno, ispitat će mogućnost ekstrakcije ključnih riječi standardnim metodama iz tekstova pisanih na talijanskom jeziku. Koliko je poznato, do sada metode nikad nisu bile usporedno testirane na talijanskim tekstovima. Sva istraživanja vezana za ekstrakciju ključnih riječi najčešće su provedena na tekstovima pisanim na engleskom jeziku. Poznat je samo mali broj istraživanja koji je proveden za neke od jezika poput španjolskog i francuskog (O. Medelyan, 2009), portugalskog (Abilhoa, De Castro, 2014), češkog (Kozłowski, Kozłowski, 2014), hrvatskog (Beliga, Martinčić-Ipšić, 2017; Beliga, Meštrović, Martinčić-Ipšić, 2016), srpskog (Beliga, Kitanović, Stanković, Martinčić-Ipšić, 2018) i sl. Međutim, ta istraživanja nužno ne uspoređuju uspješnost dvije spomenute metode, već se pojedina od njih bave specifičnom problematikom istraživanja vezanog uz ekstrakciju ključnih riječi.

Ostatak rada organiziran je kako slijedi. U drugome poglavlju definirani su glavni istraživački ciljevi ovog istraživanja, kao i motivacija za rad. Treće poglavlje opisuje metodologiju, ali i detaljnu arhitekturu sustava MAUI i RAKE kao i postupak evaluacije po kojem će njihova funkcionalnost biti izmjerena i ispitana. U četvrtome poglavlju opisani su eksperimentalni podaci i parametri metoda koje su korištene u eksperimentu. Analiza rezultata ekstrakcije ključnih riječi prikazana je u petome poglavlju. Šesto poglavlje donosi neke primjere ekstrahiranih ključnih riječi, a sedmo zaključak.

2 Standardne metode i istraživački ciljevi

Iako suvremene metode za automatsku ekstrakciju ključnih riječi koriste različite tehnike i modele poput grafova i kompleksnih mreža, neuralnih mreža i dubokog učenja, ovaj rad ispitati će dvije metode iz skupa dobro poznatih i standardnih metoda, koje će u ovom istraživanju po prvi puta biti primijenjene i uspoređene za talijanski jezik. Samim time, dobit će se osnovni uvid u funkcionalnost tih metoda na podacima koji su različiti od tekstova pisanih na engleskom jeziku, za koji su metode inicijalno razvijene. Osim toga, znanstveni rezultati ovog istraživanja, iskazani standardnim mjerama uspješnosti iz područja pretraživanja i dohvaćanja informacija, osigurat će osnovne numeričke rezultate (engl. *baseline*) koji su potrebni za uspoređivanje uspješnosti svakog daljnjeg istraživanja nekih drugih (sofisticiranijih) metoda i njihove uspješnosti ekstrakcije na talijanskome jeziku.

U ovome radu definirani su sljedeći istraživački ciljevi:

1. Ekstrahirat će se ključne riječi iz talijanskih tekstova koristeći standardne metode MAUI i RAKE.
2. Komparativno će se analizirati uspješnost ekstrakcije, metoda RAKE i MAUI, koristeći standardne metode matematičke statistike i mjere iz strojnog učenja za vrednovanje uspješnosti dohvaćenih informacija.
3. Ispitat će se postiže li nadzirana metoda (MAUI), u terminima preciznosti, odziva i *F*1 mjere, bolje rezultate nego nenadzirana (RAKE) na talijanskom skupu podataka.

3 Metodologija

U ovom poglavlju objašnjene su dvije metode koje se ustaljeno koriste u zadatku ekstrakcije ključnih riječi. Prva (MAUI) pripada skupini nadziranih (engl. *supervised*), a druga (RAKE) skupini nenadziranih (engl. *unspupervised*) metoda. Odabrane su ponajprije zbog toga što svaka predstavlja osnovnu metodu (preteču kasnije nastalim metodama) skupine kojoj pripada.

3.1 MAUI

MAUI je akronim koji dolazi od engleskog naziva ***M**ulti-purpose **A**utomatic **T**opic **I**ndexing*, a predstavlja algoritam za automatsko indeksiranje tema iz teksta (O. Medelyan, 2009). Algoritam u osnovi sadrži 4 programske komponente otvorenog kôda. MAUI se bazira na standardnoj metodi iz

skupine nadziranih metoda, koja je ujedno i preteča u rješavanju zadatka automatske ekstrakcije ključnih riječi (Tonkin, Tourte, 2014). Riječ je o metodi KEA (engl. *keyphrase extraction algorithm*), tvoraca skupine autora Witten i sur., a koja je postala standard u području automatske ekstrakcije ključnih riječi odnosno ključnih fraza (Witten, Paynter, Frank, Gutwin, Nevill-Manning, 1999). Postupak filtriranja ključnih fraza i ekstrakcija n-grama iz KEA-e preuzeta je u MAUI te dodatno proširena sljedećim elementima:

- novim algoritmom za mapiranje bilo kojeg teksta u bilo koji kontrolirani vokabular u SKOS (engl. *Simple Knowledge Organization System*) formatu,
- novim algoritmom za mapiranje bilo kojeg teksta s pojmovima sadržanim u Wikipediji,
- nove značajke kao što su položaj posljednje pojave i širenja, semantička sličnost, inverzna frekvencija Wikipedije, ukupna frazeologičnost (engl. *keyphraseness*) Wikipedije i općenitost.

Ostale pojedinosti vezane za rad KEA metode mogu se pronaći u (Witten i sur., 1999).

Druga komponenta je **WEKA** (engl. *Waikato Environment for Knowledge Analysis*) razvijena na Sveučilištu u Waikatu, a nudi alate za dubinsku analizu podataka i strojno učenje (Witten, Frank, Hall, Pal, 2016). U kontekstu MAUI algoritma, WEKA se koristi za izradu modela za indeksiranje tema i njegovu primjenu na nove dokumente. Naivni Bayesov algoritam zamijenjen je vrećom stabala odlučivanja (engl. *bagged decision trees*) u implementaciji novog klasifikatora.

Treća komponenta je **JENA** softver (McBride, 2001), tj. RDF API (engl. *Resource Description Framework Application Programming Interface*) pisan u programskom jeziku Java koji služi za implementaciju RDF modela i specifikaciju sintakse u skladu s W3C standardom ("W3C - Standards," n.d.). MAUI koristi JENA-u kod uključivanja vanjskih kontroliranih rječnika budući da su tezaurusi formatirani prema RDF standardu.

Četvrta komponenta je **Wikipedia Miner** (Milne, Witten, 2013), alat otvorenog kôda za dubinsku analizu podataka iz Wikipedije. Za potrebe MAUI-a, Wikipedia Miner pretvara depoe Wikipedije (engl. Wikipedia dumps) u format prikladan za pohranu unutar MySQL baze podataka. Osim toga, služi i za računanje semantičke sličnosti između pojedinih članaka Wikipedije na temelju kojih MAUI razdvaja fraze iz dokumenata na članke Wikipedije i određuje semantičke značajke. Sve 4 spomenute komponente u koheziji tvore algoritam za indeksiranje pojedinih tema i ekstrakciju ključnih riječi.

Dodatno, osim ekstrakcije ključnih riječi iz teksta, MAUI može obavljati i zadatak određivanja terminologije u tekstu uz pomoć kontroliranog vokabulara ili tezaurusa, indeksirati subjekte, ekstrahirati terminologiju iz teksta te automatski označavati i ekstrahirati koncepte iz teksta.

U postupku generiranja kandidata za ključne riječi (tj. indeksiranja tema), MAUI algoritam provodi postupak učenja modela za indeksiranje iz ručno označenih tema (ljudski eksperti su manualno označili) i primjenu naučenog modela koji određuje teme dokumentima koji nemaju prethodno označene teme, tj. onima na kojima nije učio. Drugim riječima, implementira nadzirani pristup strojnog učenja (engl. *supervised machine learning approach*).

3.2 RAKE

RAKE je akronim koji dolazi od engleskog naziva ***R**apid **A**utomatic **K**eyword **E**xtraction*, a predstavlja metodu za automatsko ekstrahiranje ključnih riječi iz teksta (Berry, Kogan, 2010). Autori metode, Rose, Engel, Cramer i Coweley (2010) pri razvoju metode bili su motivirani idejom da metoda bude ekstremno učinkovita, primjenjiva na individualnim dokumentima i dinamičnim kolekcijama, ali primjenjiva i na različitim domenama i vrstama tekstova. RAKE spada u skupinu nenadziranih metoda jer u postupku ekstrahiranja ključnih riječi ne iziskuje inicijalne anotacije.

Njihove opservacije rezultirale su zaključkom da ključne riječi češće sadrže funkcijske riječi (sekvence riječi koje nose sadržaj), a rjeđe zaustavne riječi (engl. *stop words*, one koje ne nose ili nose minimalno leksičko značenje, poput veznika, čestica, prijedloga i sl. kao i interpunkcijskih znakova). U sustavima za dohvaćanje informacija (engl. *Information Retrieval Systems*) zaustavne riječi se obično isključuju iz indeksa kao i iz različitih vrsta analiza teksta jer je poznato da su zbog visoke frekvencije pojavljivanja neinformativne i nepotrebne u takvim analizama i zadacima.

Ulazni parametri za RAKE metodu su (Rose, Engel, Cramer, Cowley, 2010):

- lista zaustavnih riječi (npr. *i, ili, a, ali, nego, već, no, o, ne, oj, …*),
- lista znakova koji se koriste kao razdjelnici fraza i
- lista znakova koji se koriste kao graničnici među riječima.

Zaustavne riječi i oznake koje razgraničuju (omeđuju) fraze koriste se za particioniranje teksta u riječi koje predstavljaju kandidate za ključne riječi. Izbacivanjem zaustavnih riječi i interpunkcija u kandidate za ključne riječi svrstavaju se samo riječi koje su puni nositelji sadržaja i to zajedno s očuvanom informacijom o susjedstvu drugih riječi bez potrebe korištenja kliznog prozora. Takvim pojednostavljenim načinom selekcije riječi metoda se

automatski prilagođava stilu i sadržaju teksta te omogućuje adaptivno i granularno mjerenje pojavljivanja susjednih riječi (parova riječi) koje će se koristiti kao kandidati za ključne riječi.

RAKE u **prvoj fazi** parsira tekst u kandidate za ključne riječi. Riječi iz teksta smješta u listu riječi s obzirom na graničnike među riječima. Takva lista riječi se potom dijeli u sekvence uzastopnih riječi s obzirom na graničnike među frazama i zaustavne riječi. Riječi unutar jedne sekvence razmatraju se kao jedna ključna riječ. Primjerice za rečenicu: „*Protokolarno otvorenje EPK 2020 održat će se u HNK Ivana pl. Zajca u popodnevnim satima.*“ metoda će parsirati tekst u: „*Protokolarno – otvorenje – EPK 2020 – HNK Ivana pl. Zajca – popodnevnim satima*“. Parsiranjem su zaustavne riječi *održat, će, se, u, u* izbačene i zamijenjene znakom (–), a susjedstvo između fraza i sekvenca od više riječi koje sadrže prazan znak (razmak) je zadržana. Na taj način sekvenca od 2 riječi *EPK 2020* promatra se kao jedan kandidat koji može postati ključna riječ (tj. ključna fraza) iako je sastavljena od 2 riječi.

Nakon identificiranja kandidata za ključne riječi i kreiranja grafa koji održava pojavu susjedstva riječi u tekstu, slijedi **druga faza**, tj. dodjeljivanje numeričke vrijednosti pojedinom kandidatu. Vrijednost pojedinog kandidata računa se kao suma vrijednosti svih riječi koje su sadržane u kandidatu. U izračunu vrijednosti za kandidata koriste se sljedeće metrike za izračun vrijednosti pojedine riječi:

- frekvencija pojavnice: *(freq(w))*,
- stupanj pojavnice: *(deg(w))* i
- omjer stupnja i frekvencije pojavnice: *(freq(w)/ deg(w))*.

Kandidati s većom sumarnom vrijednosti značajniji su kandidati za stvarne ključne riječi.

S obzirom na to da RAKE inicijalno izbacuje zaustavne riječi, a one su ponekad sastavni dio ključne fraze, u sljedećem koraku metoda identificira potrebe za vraćanjem takvih riječi, tj. korekcijom takvih slučajeva. Nadalje u trećoj fazi, metoda pretražuje parove ključnih riječi u kojima se riječi pojavljuju najmanje 2 puta u tekstu i to u istom poretku. Tada se kreira novi kandidat za ključnu riječ uključujući i zaustavnu riječ između riječi koje su prethodno bile kandidat za ključnu riječ. Primjerice, stari kandidat za ključnu riječ „glazba buka“ se u tekstu pojavljuje dva ili više puta i to s istim poretkom riječi pa metoda vraća zaustavnu riječ „i“ pa novi kandidat za ključnu riječ glasi „glazba i buka“. Vrijednost kandidata se izračunava ponovo i to tako da se sumiraju vrijednosti frekvencije, stupnja te omjera stupnja i frekvencije za sve 3 riječi u jednu jedinstvenu vrijednost. Autori metode napominju da takvih slučajeva najčešće nema puno jer je postavljen uvjet da se takvi kandidati moraju pojaviti minimalno dva puta i to u istom poretku riječi. Vjerojatnost pojavljivanja takvih slučajeva je veća u dužim tekstovima.

U zadnjoj, **četvrtoj fazi**, vrši se rangiranje ***top T*** ključnih riječi za promatrani tekst, prema dodijeljenim vrijednostima svakom kandidatu kako je opisano u prethodnoj fazi. Pritom se ***T*** definira prema preporuci autora kao jedna trećina od ukupnog broja riječi u grafu (Mihalcea, Tarau, 2004).

3.3 Evaluacija

Evaluacija uspješnosti eksperimenata ekstrakcije ključnih riječi najčešće se provodi standardnim mjerama kao što su preciznost (engl. *precision* – *P*), odziv (engl. *recall* – *R*) i njihova harmonijska sredina, tj. vrijednost *F*1 mjere.

Preciznost koju postiže računalna metoda računa se kao broj riječi koje se nalaze u presjeku skupa kojeg je označio čovjek (A) i skupa riječi koje je odredila računalna metoda (B) podijeljen s brojem riječi koje je odredila računalna metoda (B) te se matematički zapisuje kao $P = |A \cap B|/B$.

Odziv kojeg postiže računalna metoda računa se kao broj riječi koje se nalaze u presjeku skupa kojeg je označio čovjek (A) i skupa riječi koje je odredila računalna metoda (B) podijeljen s brojem riječi koje je označio čovjek (A) te se matematički zapisuje kao $R = |A \cap B|/A$.

Harmonijska sredina preciznosti (*P*) i odziva (*R*) naziva se *F*1 mjera te se matematički zapisuje kao $F1 = 2PR / P+R$.

U slučaju kada za pojedini tekst postoji više skupova ljudskih anotacija te se uspješnost svih takvih skupova uspoređuje s ključnim riječima koje je algoritam odredio automatski, jedan od pristupa računanja preciznosti, odziva i *F*1 mjere uključuje izračun prosječnih vrijednosti. U takvom pristupu objedinjuju se sve ključne riječi svih skupova koje su označili ljudi za pojedini dokument u jedan skup koji se uzima kao zlatni standard. Tada je moguće procijeniti slaganje automatski označenih ključnih riječi koje se nalaze u jedinstvenom skupu riječi s ključnim riječima koje su odredili ljudi u smislu prosječne preciznosti (P_{avg}), odziva (R_{avg}) i $F1_{avg}$ mjere.

4 Empirijski (eksperimentalni) podaci i analiza

U ovome poglavlju opisani su podaci koji su korišteni u eksperimentu kao i vrijednosti parametara u testiranim metodama (MAUI i RAKE).

4.1 Podaci

Za potrebe eksperimenta korišteno je 35 tekstova koji su preuzeti s internetskog portala *pogledaj.to* i

tvore skup podataka za **testiranje uspješnosti metoda** za automatsko ekstrahiranje ključnih riječi u tekstu. U nedostatku izvornih govornika talijanskog jezika, koji bi mogli anotirati ključne riječi na talijanskim tekstovima, eksperiment je proveden na tekstovima koji su originalno pisani na hrvatskom jeziku, a u eksperimentu su korišteni njihovi talijanski prijevodi koje su načinili profesionalni prevoditelji za talijanski jezik. Tekstovi su očišćeni od slika koje se popratno javljaju u tekstovima. Osim originalnog teksta iz svakog pojedinog članka, zadržani su naslovi i podnaslovi. Svi korišteni tekstovi pišu o temama vezanim za umjetnost i arhitekturu.

Tekstovi u svojoj originalnoj verziji u kakvoj su napisani i preuzeti s internetskog portala imaju pripadne oznake ključnih riječi (engl. *tags*) koje je definirao autor pojedinog teksta. Međutim, za svaki tekst je dodatno osigurano po 8 različitih skupova ključnih riječi koje su označili ljudi. Pripadne ključne riječi su ljudi originalno označili na hrvatskim tekstovima, a kasnije su ih prevoditelji zajedno sa samim tekstovima preveli na talijanski jezik. Na taj način pripremljeno je 35 tekstova pisanih na talijanskom jeziku s pripadajućim različitim skupovima anotacija (1 skupom ključnih riječi koje je odredio autor teksta i 8 različitih skupova koje su označili ljudi, dakle ukupno 9). Važno je napomenuti da se u svakom skupu anotacija nalazi minimalno 4 te maksimalno 8 ključnih riječi. Ključne riječi nisu uvijek izolirane riječi već i fraze sastavljene od dvije ili maksimalno tri riječi u nizu. Prilikom označavanja ključnih riječi ljudi su to činili neovisno jedni od drugih kako bi se osigurala neovisnost mišljenja u anotiranju ključnih riječi. Ukupan i prosječan broj označenih riječi u tekstovima vidljiv je u tablici 1.

Za testiranje nenadziranih metoda podatkovni skupovi za ovu vrstu eksperimenta variraju u broju tekstova između nekoliko desetaka (primjerice 20 tekstova u Wiki-20 (Medelyan, Witten, & Milne, 2008)) do nekoliko stotina ili tisuća tekstova sadržanih u podatkovnom skupu (primjerice 2000 tekstova u podatkovnom skupu Inspec (Hulth, 2003)). Razlog je što se nenadzirane metode, kao što je RAKE, mogu evaluirati i na manjem broju uzoraka. S druge strane, nadzirane metode, poput MAUI, zahtijevaju osim testova za testiranje i nešto veći broj tekstova u skupu za učenje.

Zbog navedenog, osim spomenutih 35 tekstova i 9 pripadnih skupova ručno označenih ključnih riječi, dodatno je pripremljeno još 85 tekstova preuzetih s portala *pogledaj.to*, koji imaju samo jedan skup označenih ključnih riječi koje je odredio sam autor pojedinog teksta. Takvi tekstovi u eksperimentu će služiti kao dodatni podaci u **učenju modela**, ne bi li se osigurao što veći broj podataka za učenje, a nastali su također prevođenjem s hrvatskog na talijanski jezik.

Dakle, skup podataka sadrži 120 dokumenata od kojih svaki dokument ima pridružen skup ključnih riječi koje je označio autor teksta. Dodatno, 35 tekstova ima još 8 dodatnih skupova ključnih riječi koje su označili ljudi. Zbog velikog broja označenih ključnih riječi, ti dokumenti pogodni su za testiranje uspješnosti izvođenja automatske ekstrakcije ključnih riječi.

4.2 Parametrizacija metoda

U ovome dijelu za svaku metodu bit će objašnjene osnovne postavke koje su korištene u eksperimentu. To su ponajprije postavke testiranja i ručno namještani parametri.

4.2.1 Postavke metode RAKE

Metoda RAKE **testirana** je na 35 dokumenata. Svakom dokumentu definiran je jedan skup ključnih riječi koji sadrži sve jedinstvene ključne riječi iz 9 različitih podskupova koje su označili ljudi na pojedinom dokumentu. Na taj način, riječi koje su se pojavljivale u više od jednog od 9 podskupova, u konačnoj listi ključnih riječi nalazile su se samo jednom.

Četiri su važna **parametra** u RAKE algoritmu koje je potrebno namjestiti prije korištenja. Prvi parametar je lista zaustavnih riječi. U ovom eksperimentu korištena je lista zaustavnih riječi koja je standardno definirana za talijanski jezik u NLTK (engl. *Natural Language Toolkit*) biblioteci ("NLTK 3.5b1 documentation," 2020) za korištenje u programskom jeziku Python (Bird, Klein, Loper, 2010). Ostali parametri su najmanja duljina riječi (postavljeno na 5 znakova), najveći broj riječi u frazi (postavljeno na 3) i broj pojavljivanja ključne riječi u tekstu (postavljeno na 4). Vrijednosti parametara podešene su kao u izvornom algoritmu.

U eksperimentu je korištena **implementacija** RAKE algoritma napisana u programskom jeziku Python te dostupna kao biblioteka **na izvoru** (PyPI, 2020). Implementirana je prema izvornom opisu tvorca algoritma (Berry, Kogan, 2010).

4.2.2 Postavke metode MAUI

Metoda MAUI **testirana** je na istih 35 dokumenata kao i metoda RAKE. Testiranje se izvodilo 9 puta. U svakoj fazi testiranja, učenje modela izvodilo se na ključnim riječima jednog od 8 timova, a deveti put na originalnim ključnim riječima koje je za pojedine tekstove definirao autor teksta. Na taj način je model učen na ključnim riječima jednog tima, a testiran na preostalih 8. Dodatno, kako bi skup za učenje bio što veći, u skup za učenje u svakoj fazi učenja modela,

dodano je još 85 tekstova koji imaju samo jedan skup označenih ključnih riječi i to od samog autora teksta.

Korišten programski kôd MAUI algoritma dostupan je na izvoru (O. Medelyan, 2020). Za korištenje algoritma, tekstovi moraju biti pohranjeni u zasebnim tekstualnim datotekama (.txt), a ključne riječi za učenje u posebnim datotekama (.key). Pojedinosti o tehničkoj izvedbi detaljnije su elaborirane u (O. Medelyan, 2009, 2020).

Parametri metode korišteni u eksperimentu jednaki su onima iz originalne implementacije algoritma. U slučaju kreiranja novog modela iz ulaznih dokumenata, algoritam svaki puta kreira nove *.key* datoteke sa ključnim riječima iz ulaznih tekstualnih datoteka, a u slučaju kada model već postoji kao i *.key* datoteke, tada algoritam evaluira automatski određene ključne riječi s već postojećim ključnim riječima.

5 Rezultati eksperimenta

Tablica 1 prikazuje usporedne rezultate postignute na zadatku ekstrakcije ključnih riječi metodama MAUI i RAKE. Gornji dio tablice (prva tri retka) odnosi se na rezultate postignute metodom MAUI, a donji dio (donja tri retka) na rezultate postignute metodom RAKE. U stupcima je usporedno prikazan ukupan broj kao i prosječan broj ključnih riječi u tekstu koje je ručno označio čovjek, koje je automatski odredio stroj (odnosno metoda) te broj riječi u njihovom presjeku. Iz rezultata se jasno iščitava da je metoda MAUI iz tekstova izvukla više ključnih riječi nego metoda RAKE (272:82). Drugim riječima, RAKE je prema brojnosti izvukao samo 30% riječi od ukupnog broja ključnih riječi koje je izvukla metoda MAUI. Štoviše, metoda MAUI je izvukla više riječi iz tekstova nego što je to inicijalno učinio čovjek (76 riječi više od čovjeka), dok je s druge strane RAKE predložio 114 riječi manje od čovjeka, što je zapravo 50% manje od broja riječi koje je predložio čovjek.

Također, brojnost u presjeku riječi, odnosno riječima koje su zajedničke skupu kojeg je odredio čovjek i kojeg je izdvojila metoda, veća je u slučaju metode MAUI nego u slučaju RAKE. To bi značilo da je u tom kontekstu metoda MAUI uspješnija. Usprkos tome, valja uzeti u obzir i ukupan broj riječi koje je metoda izdvojila. U tom slučaju uspješnost metoda je točno 26,8% bez obzira koju metodu promatramo, MAUI ili RAKE. Stoga je uputno rezultate interpretirati i pomoću drugih mjera uspješnosti, kao što su standardne mjere iz strojnog učenja za vrednovanje uspješnosti dohvaćenih informacija poput odziva i preciznosti.

		UKUPAN BR. RIJEČI	**PROSJEČAN BR. RIJEČI**
MAUI	ČOVJEK	196	5.60
	STROJ	272	7.77
	PRESJEK	73	2.08
RAKE	ČOVJEK	196	5.60
	STROJ	82	2.34
	PRESJEK	22	0.62

Tablica 1. Usporedba rezultata ekstrahiranih ključnih riječi metodama MAUI i RAKE s ključnim riječima koje su označili ljudi (izraženo ukupnim i prosječnim brojem riječi) na 35 tekstova odabranih za testiranje

U tablici 2 prikazani su rezultati uspješnosti ekstrakcije ključnih riječi metodom MAUI (u prvom retku tablice) i metodom RAKE (u drugom retku tablice) u terminima preciznosti (*P*), odziva (*R*) i harmonijske sredine preciznosti i odziva – vrijednosti *F*1.

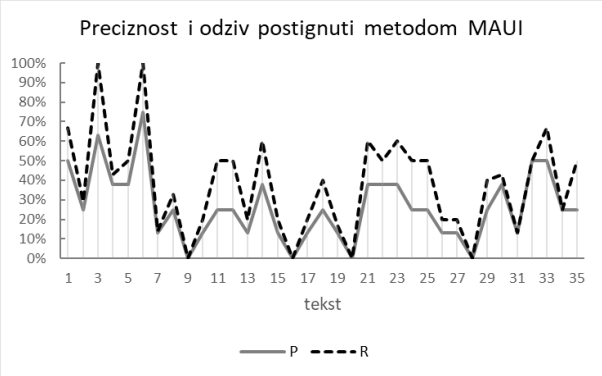
Izraženo mjerom preciznosti, uspješnost metoda je približno jednaka (razlika za 1%) dok odzivom metode veću uspješnost postiže MAUI (značajna prednost od 27%). U konačnici, prevaga uspješnosti u odzivu očituje se i u rezultatima *F*1 vrijednosti gdje MAUI naspram RAKE metode postiže prednost od približno 15%.

	P_{avg}	R_{avg}	$F1_{avg}$
MAUI	0.2709	0.3915	0.3194
RAKE	0.2611	0.1189	0.1620

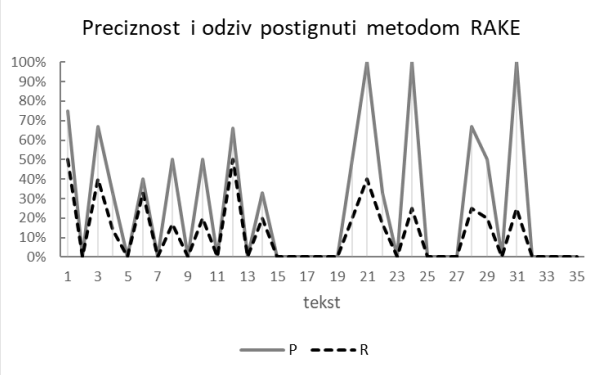
Tablica 2. Usporedba rezultata ekstrahiranih ključnih riječi metodama MAUI i RAKE, izraženi mjerama prosječne preciznosti (*P*), odziva (*R*) i *F*1 vrijednosti

U nastavku su rezultati eksperimenta ilustrirani grafikonima. Slika 1 ilustrira odnos postignute preciznosti (puna siva linija) i odziva (isprekidana crna linija) za pojedini dokument metodom MAUI. Uspješnost je prikazana postotnim vrijednostima na *y*-osi za pojedini tekst (1-35) na *x*-osi. Ista ilustracija prikazana je i za metodu RAKE na Slici 2.

Interpretacijom postignutih vrijednosti *P* i *R* krivulja na spomenutim grafikonima uočava se da je kod metode MAUI odziv u svih 35 slučajeva redovito veći ili jednak postignutoj preciznosti (Slika 1) za razliku od slučaja kod metode RAKE gdje je situacija obratna. Štoviše, na Slici 2 prikazan je i veći broj slučajeva gdje *P* i *R* krivulja bilježe nultu vrijednost postignute preciznosti i odziva (*P* i *R* krivulje priljubljene na *x*-os, npr. u slučaju tekstova 11, 13, 15, 16, 17, …) što znatno umanjuje prosječnu vrijednost *P*, *R* i *F*1 koju postiže metoda RAKE naspram metode MAUI.

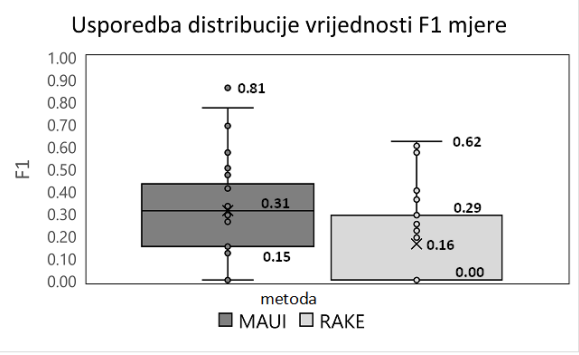


Slika 1. Prikaz odnosa postignute preciznosti i odziva metodom MAUI



Slika 2. Prikaz odnosa postignute preciznosti i odziva metodom RAKE

Raspodjela pojedinačnih vrijednosti postignute *F*1 mjere za ekstrahirane ključne riječi na pojedinom tekstu prikazana je dijagramom brkatih kutija (engl. *box-and-whisker plot*) na Slici 3. *F*1 mjera iskazana je na *y*-osi u postotnim vrijednostima. Lijeva kutija tamno sive boje prikazuje distribuciju postignutih vrijednosti *F*1 mjere primjenom metode MAUI, a desna kutija svijetlo sive boje primjenom metode RAKE.



Slika 3. Distribucija postignutih vrijednosti F1 mjere metodama MAUI i RAKE

Rasap vrijednosti u slučaju metode MAUI više nalikuje na standardnu zvonoliku distribuciju, nego što to nalikuje distribucija *F*1 vrijednosti koje su postignute metodom RAKE. Štoviše, distribucija postignuta metodom RAKE pomaknuta je ulijevo i njene vrijednosti donjeg kvartila priljubljene su na *x*-os (puno vrijednosti koje iznose 0) za razliku od MAUI gdje se donji kvartil podignuo na 15%. Samim time medijalna vrijednost kod RAKE metode je također 0, dok kod MAUI iznosi 31%. Konačno, medijalna vrijednost ukazuje na to da je centralna tendencija podataka takva da 50% izmjerenih *F*1 vrijednosti u slučaju MAUI iznosi više od 31%. Drugim riječima, ukazuje na veću uspješnost nego RAKE čiji je gornji kvartil ispod i doseže samo 29%.

Kod metode MAUI postoji i iznadprosječno visok (izolirani) rezultat od 81% koji predstavlja stršeću vrijednost s obzirom na populaciju (engl. *outlier*). On ne predstavlja netočno izmjeren podatak već primjer rijetke pojave visoke *F*1 vrijednosti u populaciji. Takav slučaj kod metode RAKE nije izmjeren. U konačnici, mjere centralne tendencije statistički opisuju skup vrijednosti *F*1 mjere koju postiže metoda MAUI pogodnijom, nego metoda RAKE.

Osim deskriptivne statistike, eksperimentalni rezultati iskazani u terminima *F*1 mjere analizirani su i postupcima inferencijalne statistike, a sve kako bi se što snažnije potvrdila razlika u rezultatima ekstrakcije postignutih metodama RAKE i MAUI te samim time potvrdila ili opovrgnula statistička značajnost predominacije metode MAUI nad metodom RAKE.

U tom kontekstu, ispitujemo je li eksperimentalno utvrđena razlika uspješnosti dviju metoda, u terminima postignutih vrijednosti F1 mjere, statistički značajna. Kolmogorov-Smirnovim testom utvrđeno je da podaci nisu normalno distribuirani te se u nastavku provodi dvosmjerni Wilcoxonov test parova koji je neovisan o distribuciji podataka. Postavljena je nulta hipoteza, *H_0: razlika medijana dva uzorka jednaka je nuli*. Drugim riječima, tvrdimo da su medijani dvaju promatranih uzoraka identični. Uz razinu pogreške α=0,01=1%, dobiveni rezultati testa (Z=-3.0465, p=0.00228) su signifikantni za p<0,01 uz veliku veličinu efekta (r=0,53) te se nulta hipoteza odbacuje. Zaključujemo suprotno iskazu nulte hipoteze, da razlika medijana nije jednaka nuli već nekoj drugoj vrijednosti, a samim time da postoji statistički značajna razlika u rezultatima *F*1 mjere koji su postignuti RAKE metodom od onih koji su postignuti MAUI metodom. Zaključak donosimo sa sigurnošću od 99,77%.

6 Primjeri ekstrahiranih ključnih riječi

U tablici 3 može se vidjeti da postoje podudarnosti u pojedinim riječima između obje metode i riječi koje je čovjek odredio kao ključnima (npr. *Vučedol*). Međutim, postoje i djelomične podudarnosti u frazama koje su sastavljene od više riječi pa se samo

neke riječi iz fraze podudaraju, ali ne i cijela fraza (npr. *cultura di Vučedol*). Algoritam MAUI je u ovom primjeru teksta postigao više podudarnosti s ključnim riječima koje je odredio čovjek te je u blagoj prednosti ispred algoritma RAKE, što može biti posljedica toga što MAUI najprije uči na pripremljenim podacima, dok RAKE koristi nenadzirani pristup.

RIJEČ	ČOVJEK	MAUI	RAKE
Vučedol	+	+	+
capeggio	+	-	-
Museo	+	+	+
architettura	+	+	-
cultura di Vučedol	+	+	+/- cultura vučedol
Progetto	+	-	+
Vukovar	-	-	+
cultura	+	+	+
Ministero della cultura	+/- cultura	+	+/- cultura
Architetti	-	+	-
pubblico	-	+	-
Goran	-	+	-

Tablica 3. - Ključne riječi koje je odredio čovjek te ekstrahirane ključne riječi metodama MAUI i RAKE na primjeru teksta o gradu Vukovaru (+ riječ je označena, - nije označena, +/- djelomično označena)

Ekstrahiranje ključne riječi idućeg primjera prikazane su u tablici 4 te one također prema postignutim preklapanjima s riječima koje je označio čovjek, prednost daju algoritmu MAUI radije nego RAKE. Iako, postoje i primjeri u kojima se MAUI i RAKE preklapaju, čovjekove anotacije nisu u skladu s njima. Primjerice, MAUI i RAKE izdvajaju riječ *progetto*, a čovjek sličnu riječ *progetti*. Slično odstupanje morfološkom varijacijom prisutno je i u primjeru *premi* vs. *premio* te *pirano* vs. *piran*.

RIJEČ	ČOVJEK	MAUI	RAKE
conferenza	+	+	-
architetti	+	-	-
premio Piranesi	+	+	-
progetti	+	-	-
progetto	-	+	+
architettura	-	-	+
Pirano	-	+	+
giornate	-	-	+
Piran	-	-	+
premi	-	+	+
gli	-	+	-
facolta	-	+	-
facilita	-	-	+

Tablica 4. - Ključne riječi koje je odredio čovjek te ekstrahirane ključne riječi metodama MAUI i RAKE na primjeru teksta o nagradi *Piranesi* (+ riječ je označena, - nije označena, +/- djelomično označena)

U većini primjera algoritam RAKE blago zaostaje za MAUI u smislu podudaranja ekstrahiranih riječi s riječima koje je anotirao čovjek. Štoviše, u brojnim primjerima se uočava da algoritam MAUI ima izražajniju sposobnost ekstrahiranja ključnih riječi u formi koje sadrže više od jedne riječi (dvije ili više), dok RAKE to uspijeva samo povremeno, a najčešće predlaže ključne riječi koje sadrže samo jednu riječ.

7 Zaključak

U ovom radu opisano je automatsko ekstrahiranje ključnih riječi iz teksta standardnim računalnim postupcima. Detaljno su opisane dvije standardne metode MAUI i RAKE. Također, ispitana je efikasnost standardne nadzirane metode MAUI te standardne nenadzirane metode RAKE. Eksperimentalnim rezultatima pokazana je uspješnost ekstrakcije obje metode u terminima preciznosti, odziva i *F*1 mjere i to na tekstovima koji su pisani na talijanskom jeziku. Na temelju postavljenih istraživačkih ciljeva te provedene analize i usporedbe eksperimentalnih rezultata ekstrakcije, zaključuje se da:

1. je standardnim metodama RAKE i MAUI moguće uspješno ekstrahirati ključne riječi iz tekstova pisanih na talijanskom jeziku,
2. nadzirana metoda MAUI postiže bolje rezultate nego metoda RAKE na talijanskom skupu podataka mjereno sa standardnom *F*1 mjerom te
3. su eksperimentalno dobiveni rezultati statistički relevantni.

Osim navedenih osnovnih doprinosa, važno je napomenuti kako je za potrebe eksperimenta načinjen podatkovni skup kojeg tvore prikupljeni tekstovi na talijanskom jeziku uključujući i anotacije ključnih riječi. Također valja napomenuti da je ovo prva usporedna studija nadzirane i nenadzirane metode za automatsku ekstrakciju ključnih riječi na talijanskim tekstovima. Predstavljeni rezultati važni su za daljnja istraživanja sofisticiranijih metoda za ekstrakciju ključnih riječi jer će se njihova efikasnost moći usporediti s uspješnošću standardnih i dobro poznatih metoda. U budućem radu, usporedna analiza metoda na talijanskim tekstovima planira se proširiti dodavanjem drugih metoda koje se temelje na ekstrakciji ključnih riječi iz grafa poput *Text Rank* (Mihalcea, Tarau, 2004) i *SBKE* metode (Beliga i sur., 2016).

Literatura

Abilhoa, W. D., & De Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, *240*, 308–325. https://doi.org/10.1016/j.amc.2014.04.090

Beliga, S. (2019). *Keyword Extraction Based on Structural Properties of Language Complex Networks*. PhD Thesis. University of Rijeka, Rijeka. Retrieved from https://www.bib.irb.hr/1026311

Beliga, S., Kitanović, O., Stanković, R., & Martinčić-Ipšić, S. (2018). Keyword Extraction from Parallel Abstracts of Scientific Publications. In J. Szymański & Y. Velegrakis (Eds.), *Lecture Notes in Computer Science* (Vol. 10546 LNCS, pp. 44–55). Cham: Springer Verlag. https://doi.org/10.1007/978-3-319-74497-1_5

Beliga, S., & Martinčić-Ipšić, S. (2017). Network-enabled keyword extraction for under-resourced languages. In A. Calì, D. Gorgan, & M. Ugarte (Eds.), *Lecture Notes in Computer Science* (Vol. 10151 LNCS, pp. 124–135). Cham: Springer Verlag. https://doi.org/10.1007/978-3-319-53640-8_11

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2014). Toward Selectivity Based Keyword Extraction for Croatian News. In P. Rupino da Cunha, N. T. Nguyen, O. Boucelma, B. Cautis, & Y. Velegrakis (Eds.), *Surfacing the Deep and the Social Web (SDSW 2014)* (p. 14). CEUR Proc. vol. 1310. Retrieved from http://ceur-ws.org/Vol-1310/

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An Overview of Graph-Based Keyword Extraction Methods and Approaches. *Journal of Information and Organizational Sciences*, *39*(1), 1–20. Retrieved from https://jios.foi.hr/index.php/jios/article/view/938

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2016). Selectivity-Based Keyword Extraction Method. *International Journal on Semantic Web and Information Systems*, *12*(3), 1–26. https://doi.org/10.4018/IJSWIS.2016070101

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2018). Keyword Extraction Based on Selectivity and Generalized Selectivity. In M. D. Lytras, N. Aljohani, E. Damiani, & K. T. Chui (Eds.), *Innovations, Developments, and Applications of Semantic Web and Information Systems* (pp. 170–204). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-5225-5042-6.ch007

Berry, M. W., & Kogan, J. (2010). *Text mining: applications and theory*. John Wiley & Sons.

Bird, S., Klein, E., & Loper, E. (2010). *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. Retrieved from https://www.nltk.org/book_1ed/

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* - (Vol. 10, pp. 216–223). Morristown, NJ, USA: Association for Computational Linguistics (ACL). https://doi.org/10.3115/1119355.1119383

Kozłowski, M., & Kozłowski, M. M. (2014). PKE: a novel Polish keywords extraction method. *Pomiary Automatyka Kontrola*, *60*(5), 305-308.

McBride, B. (2001). Jena: Implementing the RDF Model and Syntax Specification. In *Proceedings of the Second International Conference on Semantic Web - Volume 40* (pp. 23–28). Aachen, DEU: CEUR-WS.org.

Medelyan, O. (2009). *Human-competitive automatic topic indexing*. PhD Thesis. The University of Waikato, Hamilton, New Zealand.

Medelyan, O. (2020, January 10). Maui-indexer - Google Code Archive . Retrieved from https://code.google.com/archive/p/maui-indexer/

Medelyan, Olena, Witten, I. H., & Milne, D. (2008). Topic Indexing with Wikipedia. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WIKIAI 2008)* (pp. 19–24).

Merrouni, Z. A., Frikh, B., & Ouhbi, B. (2016). Automatic keyphrase extraction: An overview of the state of the art. In *Colloquium in Information Science and Technology, CIST* (Vol. 0, pp. 306–313). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CIST.2016.7805062

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts. In Lin D. & Wu D. (Eds.), *Proceedings of EMNLP 2004* (pp. 404–411). Barcelona, Spain: Association for Computational Linguistics.

Milne, D., & Witten, I. H. (2013). An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, *194*, 222–239. https://doi.org/10.1016/j.artint.2012.06.007

Natural Language Toolkit — NLTK 3.5b1 documentation. (2020, January 5). Retrieved March 21, 2020, from https://www.nltk.org/

PyPI. (2020, January 5). rake-nltk. Retrieved March 21, 2020, from https://pypi.org/project/rake-nltk/

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010).

Automatic Keyword Extraction from Individual Documents. In *Text Mining: Applications and Theory* (pp. 1–20). Chichester, UK: John Wiley and Sons. https://doi.org/10.1002/9780470689646.ch1

Tonkin, E., & Tourte, G. J. L. (2014). *Working with text : tools, techniques and approaches for text mining* (1st ed.). Chandos Publishing.

W3C - Standards. (n.d.). Retrieved March 11, 2020, from https://www.w3.org/standards/

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining : practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: Practical Automatic Keyphrase Extraction. In *Proc. ACM Conf. on Digital Libraries* (pp. 1–23). Berkeley, CA, US: ACM. Retrieved from http://www.nzdl.org/

Automatic keyword extraction from text with standard computer procedures

Abstract

Automatic keyword extraction takes a great interest as a research issue in the field of natural language processing and information retrieval. Although numerous methods for keyword extraction task have been developed, their effectiveness depends on many factors such as the approach used in method development, the domain to which they are adapted, the type of language or tasks for which they are constructed, etc., and still, there is a room for progress and improvements. In this paper, two existing methods are explained and reconstructed - RAKE and MAUI, which are the standard representatives of the unsupervised and supervised group of keyword extraction methods. It was experimentally tested whether the methods could successfully extract keywords from texts written in Italian, which had not been tested so far. For the experimental purposes, Italian texts were collected and annotated with keywords. The effectiveness of the MAUI method proved to be more promising than the RAKE method, which was confirmed earlier in the keyword extraction experiment from texts written in English.

Keywords: *automatic keyword extraction; standard methods; RAKE; MAUI; Italian language.*