

---

Elektrotehnički fakultet u Beogradu  
Katedra za računarsku tehniku i informatiku

*Predmet:* Pronalaženje skrivenog znanja (MS1PSZ)

*Školska godina:* 2014/2015.

# Projekat za domaći rad

## - Projektni zadatak -

### *Napomena*

Pročitati tekst zadatka **u celini i pažljivo**, pre započinjanja realizacije ili traženja pomoći. Ukoliko u zadatku nešto nije dovoljno precizno definisano ili su postavljeni kontradiktorni zahtevi, student treba da uvede razumne pretpostavke, da ih temeljno obrazloži i da nastavi da izgrađuje preostali deo svog rešenja na temeljima uvedenih pretpostavki. Zahtevi su namerno nedovoljno detaljni, jer se od studenata očekuje kreativnost i profesionalni pristup u rešavanju praktičnih problema.

## Tekst zadatka:

Semantička sličnost predstavlja koncept dodeljivanja metrike skupovima izraza ili dokumenata zasnovane na sličnosti njihovog značenja. Ovaj koncept jedan je od ključnih za razumevanje prirodnih jezika, jer omogućava pravljenje smislenih poređenja i zaključivanja. Zbog toga određivanje semantičke sličnosti igra važnu ulogu u automatskoj kategorizaciji i sumarizaciji teksta, mašinskom prevodenju, pronalaženju informacija i drugim oblastima veštačke inteligencije. Problem semantičkog poređenja kratkih tekstova (*Short Text Semantic Similarity – STSS*) ima poseban značaj, jer su kratki tekstovi u širokoj upotrebi na Internetu, u formi natpisa i opisa proizvoda, anotacija slika i veb stranica, kratkih novinskih naslova i vesti, itd. Takođe, ovaj problem igra važnu ulogu u pitanjima vezanim za obrazovanje i učenje, kao što su automatsko testiranje i ocenjivanje zadataka.

Zadatak određivanja semantičke sličnosti para rečenica predstavlja dodeljivanje binarne ocene svakom paru, gde ocena jedan ukazuje da su rečenice slične, a nula obratno. Za određivanje semantičke sličnosti potrebno je, na osnovu analize teksta, prvo izračunati vrednosti unapred zatadih atributa. Ovi atributi koriste se za treniranje *data mining* (DM) modela, a zatim se vrši evaluacija performansi dobijenog modela pomoću validacionog skupa. Dostavljeni skup podataka unapred je podeljen na deo predviđen za treniranje (trening skup), koji iznosi 70% početnog skupa, i preostali deo za evaluaciju (verifikacioni skup). U nastavku je dat opis skupa podataka koji će se koristiti za treniranje i verifikaciju, kao i opis zahtevane realizacije.

## Dodatna literatura

- Furlan B., Batanović V., Nikolić B., "Semantic Similarity of Short Texts in Languages with a Deficient Natural Language Processing Support," *Decision Support Systems*, (ISSN)0167-9236, Vol. 55, Issue 3, pp. 710–719, June 2013. [PDF](#)
- Furlan B., Sivački V., Jovanović D., Nikolić B. "Comparable Evaluation of Contemporary Corpus-Based and Knowledge-Based Semantic Similarity Measures of Short Texts," *JITA*, vol. 1, no. 1, ISSN 2233-0194 (online), pp. 65-71, June 2011. [PDF](#)
- Jovanović D., Furlan B., Nikolić B., "A Software System for Measuring the Semantic Similarity of Short Texts," In *ETRAN*, Banja Vrućica (Teslić), R. Srpska, BiH, June 6-9, 2011. [PDF](#)

## Opis podataka

Za evaluaciju realizovanog sistema koristiti MSRRC korpus (*Microsoft Research Paraphrase Corpus*) [1]:

MSRRC je najveći korpus parafraza za engleski jezik koji se sastoji od 5801 para rečenica. Svaki par rečenica je ocenjen, tj. data je ocena semantičke sličnosti od strane dvoje sudija. Pri tom, dodeljivane ocene su binarne, gde ocena jedan ukazuje da su rečenice slične, a nula obratno. Slučajeve u kojima je došlo do neslaganja dodeljenih ocena rešavao je treći sudija. Od ukupno 5801 para rečenica, njih 3900 (67%), je proglašeno semantički sličnim, dok je ostatak predstavljao semantički različite parove. Konačno, korelacija u ocenama između troje sudija iznosi 83%, što predstavlja i gornju granicu preciznosti koju sistem može postići nad ovim korpusom.

Ukupan korpus je slučajnim izborom podeljen na dva dela. Prvi deo sadrži približno 70% (4076 parova) i predstavlja treniranje skup, dok drugi deo sadrži preostalih 1725 parova i namenjen je verifikaciji.

## Zadaci

Projekat koji izrađuje svaki student sastoji se iz zadataka opisanih u nastavku. Štampane materijale pripremiti prema uputstvima datim u zadacima, a sve zajedno na sledeći način:

1. Na naslovnoj strani jasno napisati naziv predmeta, prezime i ime studenta, broj indeksa i adresu e-pošte.
2. Sve zajedno čvrsto povezati u jednu celinu, tako da se listovi ne mogu rasipati (najbolje spiralom).

### Zadatak 1 – priprema podataka (20 poena)

Nad ulaznim tekstualnim fajlovima izvršiti uklanjanje specijalnih karaktera iz rečenica, a zatim izvršiti izračunavanje vrednosti atributa za određivanje sličnosti opisanih u nastavku.

**Atributi za određivanje sličnosti** - Upotrebom *Stanford CoreNLP* [2] za svaki par rečenica generisati listu atributa na osnovu koje će se dalje vršiti ocenjivanje njihove ukupne sličnosti. Atributi se mogu podeliti u tri grupe:

1. atributi zasnovani na vrstama reči (plitko parsiranje);
2. atributi zasnovani na relacijama među rečima (duboko parsiranje);
3. opšti atributi.

Na kraju, reči koje nisu prepoznate od strane alata označiti sa *UNKNOWN* i izostaviti iz dalje analize. Takođe, ovo uraditi i za kardinalne brojeve (CD).

#### Atributi zasnovani na vrstama reči

Za svaku vrstu reči koju kao rezultat daje *Stanford POS tagger* (tabela 1) određuju se:

- a) Razlika broja reči između prve i druge rečenice za datu vrstu reči. Ovi atributi označavaju se prefiksom *diff\_Tag\_* praćenim oznakom vrste reči. Npr. atribut *diff\_Tag\_NNP* označava razliku u broju ličnih imenica jednine. Da bi ovi atributi bili neosetljivi na promenu redosleda rečenica u paru, računa se njihova apsolutna vrednost.
- b) Semantička sličnost skupova reči date vrste iz jedne i druge rečenice. Ovi atributi označavaju se prefiksom *semSim\_Tag\_* praćenim oznakom vrste reči. Npr. atribut *semSim\_Tag\_NN* označava semantičku sličnost skupova imenica jednine. S obzirom da se upotrebom navedenih alata može odrediti semantička sličnost samo između imenica i glagola, ovi atributi se određuju za ove vrste reči.
- c) Leksička sličnost skupova reči date vrste iz jedne i druge rečenice. Ovi atributi označavaju se prefiksom *lexSim\_Tag\_* praćenim oznakom vrste reči. Npr. atribut *lexSim\_Tag\_JJ* označava leksičku sličnost skupova atributa.

Pored opisanih, dodati su specijalni atributi koji grupišu imenice i glagole. Razlika u broju svih imenica (opštih, ličnih, jednine, množine) obeležava se *diffNouns*, dok se semantička i leksička sličnost obeležavaju sa *semSimNouns* i *lexSimNouns* respektivno. Glagoli svih oblika na isti način su grupisani i daju attribute *diffVerbs*, *semSimVerbs* i *lexSimVerbs*.

Oznaka	Vrsta reči	Oznaka	Vrsta reči
CC	Veznik	PRP\$	Prisvojna zamenica
CD	Kardinalni broj	RB	Prilog
DT	Determinator	RBR	Prilog – komparativ
EX	Egzistencijalno „there“	RBS	Prilog – superlativ
FW	Strana reč	RP	Rečca
IN	Predlog	SYM	Simbol
JJ	Pridev	TO	Predlog „to“
JJR	Pridev – komparativ	UH	Uzvik
JJS	Pridev – superlativ	VB	Glagol – osnovni oblik
LS	Označavač stavke u listi	VBD	Glagol – prošlo vreme
MD	Modalni glagol	VBG	Glagol – gerundiv ili prezent particip
NN	Imenica – jednina ili zbirna	VBN	Glagol – past particip
NNS	Imenica – množina	VBP	Glagol – prezent ne-trećeg lica jednine
NNP	Lična imenica – jednina	VBZ	Glagol – prezent trećeg lica jednine
NNPS	Lična imenica – množina	WDT	Wh-determinator (npr. „which“)
PDT	Predeterminator	WP\$	Prisvojna wh-zamenica
POS	Prisvojni završetak („’s“)	WRB	Wh-prilog (npr. „how“, „where“, „why“)
PRP	Lična zamenica		

Tabela 1 – vrste reči i njihove oznake

### Atributi zasnovani na relacijama među rečima

Za svaku relaciju koju kao rezultat daje *Stanford parser* odrediti:

- Razliku broja datih relacija između prve i druge rečenice. Ovi atributi označavaju se prefiksom *diff\_Dep\_* praćenim skraćenim nazivom relacije. Npr. atribut *diff\_Dep\_dobj* označava razliku u broju direktnih objekata. Da bi ovi atributi bili neosetljivi na promenu redosleda rečenica u paru, računa se njihova apsolutna vrednost.
- Semantičku sličnost dobijenih relacija iste vrste iz jedne i druge rečenice. Ovi atributi označavaju se prefiksom *semSim\_Dep\_* praćenim skraćenim nazivom relacije. Npr. atribut *semSim\_Dep\_dobj* označava izračunatu semantičku sličnost skupova direktnih objekata.
- Leksičku sličnost skupova datih relacija iz jedne i druge rečenice. Ovi atributi označavaju se prefiksom *lexSim\_Dep\_* praćenim skraćenim nazivom relacije. Npr. atribut *lexSim\_Dep\_dobj* označava izračunatu leksičku sličnost skupova direktnih objekata.

### Opšti atributi

Opšti atributi nisu zasnovani na rezultatima parsiranja, već na skupovima svih reči u rečenicama. Prvi atribut je apsolutna razlika ukupnog broja reči između prve i druge rečenice i naziva se *diff\_All*. Drugi atribut je ukupna leksička sličnost *overallLexsim* koja uzima u obzir sve reči iz jedne i druge rečenice i računa njihovu leksičku sličnost na način opisan u nastavku bez uzimanja u obzir kojoj vrsti reči pripada određena reč.

## Semantička i leksička sličnost

**Atributi zasnovani na vrstama reči (plitko parsiranje)** - Za određivanje vrednosti atributa koji predstavljaju semantičku i leksičku sličnost skupova reči pojedinih vrsta koristi se sledeći algoritam (pojednostavljena verzija algoritma [3]):

1. Rečenice  $r_1$  i  $r_2$ , koje sačinjavaju jedan par čija se sličnost određuje, formiraju dva skupa reči, odnosno dva skupa lema reči koje ove rečenice sadrže. Za svaku vrstu reči  $t$ , iz ova dva skupa izdvajaju se dva podskupa koji sadrže samo ove vrste reči. Na dalje to su skupovi  $st_1$  i  $st_2$ :

$$st_1 = \{w_{11}, w_{12}, \dots, w_{1n}\}$$

$$st_2 = \{w_{21}, w_{22}, \dots, w_{2m}\}$$

$n$  i  $m$  predstavljaju broj reči date vrste  $t$  u rečenicama  $r_1$  i  $r_2$  respektivno. Unija ova dva skupa smešta se u novi:

$$st = \{w_1, w_2, \dots, w_r\}$$

gde je  $r$  broj reči u skupu  $st$ .

2. Formira se vektor  $v_1$  dužine  $r$ , u kome je vrednost svakog  $i$ -tog člana jednaka maksimalnoj semantičkoj ili leksičkoj sličnosti između  $i$ -tog člana niza  $st$  i svih članova niza  $st_1$ .

$$v_1 = \{sim_{max}(w_1, st_1), sim_{max}(w_2, st_1), \dots, sim_{max}(w_r, st_1)\}$$

vrednost sličnosti  $sim$  računa se u zavisnosti od toga da li je u pitanju semantičko ili leksičko poređenje.

3. Na isti način na koji je kreiran vektor  $v_1$ , kreira se vektor  $v_2$  u kome će vrednost svakog  $i$ -tog člana biti maksimalna semantička ili leksička sličnost između  $i$ -tog člana niza  $st$  i svih članova niza  $st_2$ .
4. Kosinusna sličnost vektora  $v_1$  i  $v_2$  predstavlja rezultujuću sličnost između skupova reči  $st_1$  i  $st_2$ :

$$sim(st_1, st_2) = \frac{v_1 \cdot v_2}{\|v_1\| * \|v_2\|}$$

**Atributi zasnovani na relacijama među rečima (duboko parsiranje)** - Za određivanje semantičke i leksičke sličnosti skupova pojedinih relacija koje su iste vrste  $r$ , vrednost sličnosti dve relacije  $rel_1$  i  $rel_2$  računati na sledeći način:

$$sim_{dep}(rel_1, rel_2) = sim_{word}(g_1, g_2) \cdot 2^{sim_{word}(d_1, d_2) - 1}$$

gde su  $g_1$  i  $d_1$  vodeća i zavisna reč u relaciji  $rel_1$ , a  $g_2$  i  $d_2$  vodeća i zavisna reč u relaciji  $rel_2$ . Funkcija  $sim_{word}$  predstavlja jedan od pristupa opisanih u nastavku za računanje semantičke, odnosno leksičke sličnosti, između dve reči.

1. Algoritam (tj. pojednostavljena verzija algoritma [4]) polazi od skupova relacija  $sr_1$  i  $sr_2$  određenog tipa  $r$ :

$$sr_1 = \{rel_{11}, rel_{12}, \dots, rel_{1n}\}$$

$$sr_2 = \{rel_{21}, rel_{22}, \dots, rel_{2m}\}$$

i konstruiše matricu  $n \times m$  gde su  $n$  i  $m$  broj relacija datog tipa u rečenicama.

$$M = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{bmatrix}$$

Elementi matrice predstavljaju sličnosti pojedinačnih relacija:

$$\alpha_{ij} = sim_{dep}(rel_{1i}, rel_{2j}), \text{ gde je } 0 \leq \alpha_{ij} \leq 1.$$

2. Nakon konstrukcije matrice pronalazi se maksimalan element  $\alpha_{ij}$  i dodaje se u listu  $\beta$ , a svim elementima matrice iz vrste  $i$  i kolone  $j$  dodeljuje se vrednost 0. Operacija se ponavlja sve dok u matrici ima elemenata većih od 0. Na kraju se dobija lista  $\beta$  sa maksimalnim elementima po vrstama i kolonama.
3. Za dobijanje rezultujuće sličnosti koristi se sledeća jednačina:

$$sim(sr_1, sr_2) = \frac{(\sum_{i=1}^{|\beta|} \beta_i) * (m + n)}{2mn}$$

### Leksička sličnost para reči

Za određivanje leksičke sličnosti koristi se algoritam *Normalized Maximal Consecutive Longest Common Subsequence starting at any character n (NMCLCSn)*, implementiran pomoću metode *longestCommonContiguousSubstring* dostupne unutar *Stanford CoreNLP* paketa, na sledeći način:

```
import edu.stanford.nlp.util.StringUtils;
/**
 * Normalized Maximal Consecutive Longest Common Subsequence starting at
 * character n
 * http://www.site.uottawa.ca/~mdislam/publications/tkdd.pdf
 *
 * @param s1
 * @param s2
 * @return
 */
public static double NMCLCSn(String s1, String s2) {
    if(s1==null || s2==null || s1.length()==0 || s2.length()==0) return 0;
    return Math.pow(StringUtils.longestCommonContiguousSubstring(s1, s2),2) /
        (s1.length() * s2.length());
}
```

### Semantička sličnost para reči

Za određivanje semantičke sličnosti para reči koristiti algoritam *lin* [5] dostupan u alatu *Java WordNet::Similarity* [6]. Pogledati `TestExamples.java` kao primer upotrebe.

Za vrste reči za koje nije podržano određivanje semantičke sličnosti odrediti samo leksičku sličnost.

### Zadatak 2 – treniranje modela (20 poena)

Učitati u bazu podataka skupove podataka za treniranje i validaciju. Realizovati tri DM model upotrebom sledećih algoritma: (1) stabla odlučivanja (*Decision Trees*) i (2) dva proizvoljna DM algoritama.

Izvršiti evaluaciju dobijenih modela nad validacionim skupom na sledeći način. Rezultati validacije predstavljaju binarne vrednosti koje se mogu predstaviti na način prikazan u tabeli 2.

Ishod verifikacije (predviđena sličnost)	Stvarna sličnost	
	DA ( <i>true</i> )	NE ( <i>false</i> )
DA ( <i>true</i> )	TP ( <i>true positive</i> )	FP ( <i>false positive</i> )
NE ( <i>false</i> )	FN ( <i>false negative</i> )	TN ( <i>true negative</i> )

Tabela 2 – predstavljanje rezultata testiranja

Izraz „stvarno pozitivni“ (*True Positives – TP*) odnosi se na one parove rečenica koji predstavljaju parafraze i pravilno su označeni kao takvi od strane algoritma. „Stvarno negativni“ (*True Negatives – TN*) predstavljaju neslične parove rečenica koje je algoritam pravilno prepoznao. „Lažno pozitivni“ (*False Positives – FP*) čine parovi semantički različitih rečenica, koji su pogrešno označeni kao parafraze. Konačno „lažno negativni“ (*False Negatives – FN*) su parovi rečenica koji jesu parafraze, ali su pogrešno ocenjeni kao semantički različiti. Na osnovu ovih vrednosti odrediti vrednosti parametara tačnosti (*Accuracy – A*), preciznosti (*precision – P*), osetljivosti (*recall – R*) i F-meru (*F*). Ove mere intenzivno se koriste u teoriji pretraživanja informacija i računaju se na sledeći način:

$$A = \frac{TP + TN}{TP + FP + FN + TN}$$

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

U kontekstu STSS, tačnost predstavlja odnos broja pravilno ocenjenih parova i ukupnog broja parova rečenica u korpusu nad kojim se vrši evaluacija. Preciznost se može shvatiti kao odnos broja pravilno identifikovanih parova parafraza i ukupnog broja parova označenih parafrazama od strane algoritma. Osetljivost predstavlja odnos između tačno identifikovanih parova parafraza i stvarnog broja parova parafraza u korpusu. F-mera se računa kao harmonijska sredina preciznosti i osetljivosti.

### Zadatak 3 – prošireni model (10 poena)

Predložiti dodatne atribute koji će poboljšati učinak dobijenih DM modela. Detaljno dokumentovati ove atribute i prikazati dobijene vrednosti parametara A, P, R i F.

#### Proizvodi

Na prvi deo ispita potrebno je doneti kompletno urađen projektni zadatak.

Za usmenu odbranu uraditi i pripremiti sledeće:

- U elektronskoj formi (CD/DVD) :
  1. Implementirati softverski sistem upotrebom navedenih NLP alata i Microsoft SQL Server baze podataka (ili opciono drugih alata sa istim funkcionalnostima). Potrebno je doneti kompletan projekat u elektronskoj formi.
  2. Na odbrani će biti dostavljen nezavistan validacioni skup podataka nad kojim treba demonstrirati rad realizovanog sistema. Demonstracija treba da omogući zadavanje proizvoljnih ulaznih podataka, tj. putanje do fajla čiji format odgovara formatu podataka dostavljenom za treniranje modela.
- U štampanoj formi predati kompletnu dokumentaciju. Detaljno dokumentovati dati softverski sistem opisujući svaki korak po uputstvima datim u zadatku. Takođe, uvrstiti sve tražene rezultate i naglasiti ključne delove sistema.

#### Reference

- [1] "Microsoft Research Paraphrase Corpus," 2005. [Online]. Available: <http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/>
- [2] The Stanford Natural Language Processing Group, "Stanford CoreNLP." [Online]. Available: <http://nlp.stanford.edu/software/corenlp.shtml>
- [3] L. Li, Y. Zhou, B. Yuan, J. Wang, and X. Hu, "Sentence similarity measurement based on shallow parsing," in *Fuzzy Systems and Knowledge Discovery*, 2009, pp. 487–491.
- [4] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Trans. Knowl. Discov. Data*, vol. 2, no. 2, pp. 1–25, Jul. 2008.
- [5] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet:: Similarity: measuring the relatedness of concepts," in *HLT-NAACL*, 2004.
- [6] D. Hope, "Java WordNet::Similarity." [Online]. Available: [http://www.sussex.ac.uk/Users/drh21/Java WordNet Similarity beta version 11.01.source.zip](http://www.sussex.ac.uk/Users/drh21/Java%20WordNet%20Similarity%20beta%20version%2011.01.source.zip)