

Genetičko programiranje susreće lingvistiku

Računalni postupci ekstrakcije kolokacija iz korpusa

Jan Šnajder

Zavod za elektroniku, mikroelektroniku, računalne i inteligentne sustave
Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

Zagrebački lingvistički krug
11. svibnja 2010.

- Obrada prirodnog jezika
- Računalna lingvistika
- Pretraživanje informacija (engl. *information retrieval*, IR)
- Dubinska analiza teksta (engl. *text mining*)

- 1 Kolokacije i višerječne jedinice
- 2 Ekstrakcija kolokacija iz korpusa
- 3 Vrednovanje ekstrakcije
- 4 Genetičko programiranja i ekstrakcija kolokacija
- 5 Alat TermeX
- 6 Zaključak

- 1 Kolokacije i višerječne jedinice
- 2 Ekstrakcija kolokacija iz korpusa
- 3 Vrednovanje ekstrakcije
- 4 Genetičko programiranja i ekstrakcija kolokacija
- 5 Alat TermeX
- 6 Zaključak

Višerječni izrazi (1)

Radna definicija (Evert, Baldwin):

Višerječni izraz (engl. *multiword expression*, MWE) jest kombinacija od dvije ili više (ne nužno slijednih) riječi čija semantička, sintaktička ili statistička obilježja nisu u potpunosti predvidiva, stoga takav izraz treba biti naveden u leksikonu.

- Npr. *tajna policija*, *morski pas*, *biti u banani*, *plava kosa*, *ad hoc*
- “Word with spaces”

Karakteristična svojstva (Manning & Schütze):

- 1 nekompozicionalnost (semantička netransparentnost)
- 2 nezamjenjivost (leksička okamenjenost)
- 3 nepromjenjivost (sintaktička rigidnost)

Višerječni izrazi (2)

MWE obuhvaćaju niz leksikaliziranih pojava:

- **Frazemi** (*rad na crno, crna ovca, biti u banani*)
- **Vlastita imena** (*Željko Rohatinski, Zlatarevo zlato*)
- **Stručno nazivlje** (*operacijski sustav, koronarna arterija*)
- **Leksičke kolokacije** (*morski pas, javna tajna, tajna policija*)
- **Ustaljene fraze i klišeji** (*plan i program, dobar dan*)
- ...

Važni za leksikografiju, prevođenje, učenje jezika, ...

U literaturi susrećemo razne definicije. Možemo ih razvrstati u tri grupe:

① Kolokacije = MWE

Dvije ili više riječi koje odgovaraju uobičajenom načinu da se nešto izrazi. (Manning & Schütze)

② **Frazeološki/leksikografski pristup:** Kolokacije \subset MWE

- ▶ kombinacija riječi koja je manje-više nekompozicionalna
- ▶ **leksičke kolokacije**

③ **Statistički pristup:** Kolokacije \leftrightarrow MWE

Niz od dvije ili više riječi koje se **supojavljaju** (statistički značajno) **češće** no što je to očekivano. (Firth, Sinclair)

Kolokacije – naša definicija

Naša radna definicija:

Niz od dvije ili više riječi koje imaju **tendenciju supojavlivanja** (zato što odgovaraju uobičajenom načinu da se nešto izrazi).

- Čestota supojavlivanja kao indikacija leksikalizacije
- Kolokacija je **empirijski pojam**, dok je MWE **teorijski pojam**
- U načelu se preklapaju, no:
 - ▶ neke kolokacije nisu MWE (*afera Brodosplit, kupovanje ispita*)
 - ▶ neke MWE možda (u danom korpusu) nisu kolokacije
- Prema tome:
 - ▶ Kolokacije \neq MWE
 - ▶ Kolokacije \neq leksičke kolokacije \subset MWE
- Interakcija između lingvistike, statistike i računalne lingvistike

Kolokacije – za što ih trebamo?

- Pretraživanje informacija – indeksiranje (Vechtomoiva i dr.; Wacholder & Songi, 2003), proširenje upita (Mandala i dr., 2000)
- Parsanje (Baldwin, 2004)
- Ekstrakcija informacija (Lin, 1998)
- Strojno prevođenje (Gerber & Yang, 1997)
- Generiranje prirodnog jezika (Smadja & McKeown, 1990)
- Razrješavanje višeznačnosti (Wu & Chang, 2004),
- Ekstrakcija terminologije (Goldman & Wehrli, 2001)
- Klasifikacija dokumenata (Scott & Matwin, 1999)
- ...

Kolokacije – pomoć pri ručnom indeksiranju

The screenshot shows the WinAIDE application window. The main text area contains a document snippet with several terms highlighted in green: "Pravilnik", "uzročnika zoonoza", "hranom", "Zakona o veterinarstvu", "ministar poljoprivrede", "šumarstva i vodnoga gospodarstva", "PRAVILNIK", "ZAKON ZA KONTROLU SALMONELA I DRUGIH ODREĐENIH UZROČNIKA ZOOZOZA KOJI SE PRENOSE HRANOM".

On the right side, there are two panels. The top panel, titled "Suggestions", shows a list of 2-gram suggestions with their frequencies. The bottom panel, titled "Suggestions", shows a list of descriptors with their reliability scores.

2-gram	Freq
uzročnik: zoonoze	42
program kontrole	26
nadležno tijelo	23
serotipova salmonela	16
javno zdravstvo	15
prehrane ljudi	11
uzimanje uzoraka	10
primarnu proizvodnju	8
rasplodna jata	7
Salmonella Enteritidis	7
Salmonella typhimurium	7
epidemiološka jedinica	6

Descriptor	Reliability
1 zoonosis : Id. 27737	100
1 contagious disease : Id. 1266	99
1 disease prevention : Id. 5216	98
1 health legislation : Id. 5821	97
1 veterinary inspection : Id. 5612	95
1 animal disease : Id. 792	94
1 food inspection : Id. 4688	94
1 public health : Id. 5259	71

eCADIS – Indeksiranje deskriptorima EUROVOC (Kolar i dr., 2005)

- 1 Kolokacije i višerječne jedinice
- 2 Ekstrakcija kolokacija iz korpusa**
- 3 Vrednovanje ekstrakcije
- 4 Genetičko programiranja i ekstrakcija kolokacija
- 5 Alat TermeX
- 6 Zaključak

Mjera leksičke asocijacije (1)

- **N-gram** – slijed od dvije ili više riječi (bigram, trigram, tetragram)
- Najjednostavniji pristup: ekstrakcija najfrekventnijih n-grama
- **Mjera leksičke asocijacije** (engl. *lexical association measure*, AM) n-gramu pridjeljuje vrijednost koja ukazuje na jakost sveza riječi unutar n-grama
- Mjeri afinitet jedne riječi naspram druge: pojavljuje li se n-gram u korpusu češće nego što je očekivano?

Primjer: Vjesnik (g. 1999–2009) – 56MW

$f(\text{morski}) = 12364$, $f(\text{pas}) = 5741$, $f(\text{morski pas}) = 322$

$f(\text{zelen}) = 8395$, $f(\text{zvijezda}) = 10672$, $f(\text{zelena zvijezda}) = 1$

Mjere leksičke asocijacije (2)

- 1 Mjera međusobne informacije

$$MI(x, y) = \log \frac{P(xy)}{P(x)P(y)}$$

- 2 Diceov koeficijent

$$Dice(x, y) = \frac{2f(xy)}{f(x) + f(y)}$$

- 3 χ^2 -test

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- 4 Log-vjerodostojnost (engl. *log-likelihood*)

$$G^2 = 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

Primjer ekstrakcije

bigram	AM
<i>premijerom Račanom</i>	26.5
<i>posljednje utakmice</i>	6.2
<i>riblje juhe</i>	31.0
<i>pripremnog razdoblja</i>	1.7
<i>papa Ivan</i>	22.1
<i>novi ministar</i>	13.3
<i>najmanje pet</i>	12.4
<i>cijene kruha</i>	6.2
<i>kvalitete vode</i>	11.0
<i>potrošačka košarica</i>	38.2
<i>šest kuna</i>	8.2
<i>premjerovi satovi</i>	43.3
<i>Andy Roddick</i>	50.2
<i>velika prigoda</i>	6.9

Primjer ekstrakcije

bigram	AM
<i>Andy Roddick</i>	50.2
<i>premijerovi satovi</i>	43.3
<i>potrošačka košarica</i>	38.2
<i>riblje juhe</i>	31.0
<i>premijerom Račanom</i>	26.5
<i>papa Ivan</i>	22.1
<i>novi ministar</i>	13.3
<i>najmanje pet</i>	12.4
<i>kvalitete vode</i>	11.0
<i>šest kuna</i>	8.2
<i>velika prigoda</i>	6.9
<i>posljednje utakmice</i>	6.2
<i>cijene kruha</i>	6.2
<i>pripremnog razdoblja</i>	1.7

Postupak ekstrakcije

Postupak ukratko:

- 1 Opojavnichenje korpusa
- 2 Lematizacija pojavnica (flektivna morfološka normalizacija)
 - ▶ *potrošačkoj košarici* → *potrošački košarica*
- 3 Označavanje vrste riječi
 - ▶ *potrošački košarica* → AN
- 4 Prebrojavanje pojavnica
- 5 Prebrojavanje n-grama
- 6 Filtriranje n-grama prema uzorku vrsta riječi
 - ▶ AN, NN, ANN, AAN, NAN, NNN, ...
- 7 Izračunavanje mjere asocijacije za svaki preostali n-gram
- 8 Ekstrakcija visoko rangiranih n-grama

Proširenje na 3-grame, 4-grame, itd.

- Osnovne asocijacijske mjere primjenjive su samo na bigrame
- U literaturi su predložena razna proširenja za 3-grame i 4-grame
- Npr.:

$$MI(xyz) = \log \frac{P(xyz)}{P(x)P(y)P(z)}$$

- Npr. (Da Silva & Lopes, 1999; Tadić & Šojat, 2003):

$$MI(xyz) = \log \frac{MI(x, yz) + MI(xy, z)}{2}$$

- Heuristička asocijacijska mjera (Petrović i dr., 2006):

$$MI(xyz) = \begin{cases} 2 \log \frac{P(xyz)}{P(x)P(z)} & \text{ako } stop(y) \\ MI(xyz) & \text{inače} \end{cases}$$

- Posebno tretira n-grame sa “zaustavnim riječima” (prijedlozima i veznicima)
- Npr. *voda za piće, kruh i mlijeko*

Uzorci proširenja (1)

- **Uzorci proširenja** (Petrović i dr., 2010):
usustavljen pristup proširenju bigramskih asocijacijskih mjera na n-grame proizvoljnih duljina

$$G_1(g, w_1 \cdots w_n) = \frac{g(w_1, w_2 \cdots w_n) + g(w_1 \cdots w_{n-1}, w_n)}{2}$$

$$G_5(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_1 \cdots w_i, w_{i+1} \cdots w_n)$$

- 7 općenitih uzoraka proširenja

Uzorci proširenja (2)

- Dodatni uzorci za proširenje 3-grama i 4-grama u ovisnosti o položaju zaustavne riječi (engl. *stopword-sensitive patterns*)

$$H_3(g, w_1^4) = \begin{cases} \alpha_1 G_0^*(g, w_1^4, \{w_2, w_3\}) & \text{ako } stop(w_2) \wedge stop(w_3) \\ \alpha_2 G_0^*(g, w_1^4, \{w_2, w_3\}) & \text{ako } stop(w_2) \wedge \neg stop(w_3) \\ \alpha_3 G_0^*(g, w_1^4, \{w_1, w_3\}) & \text{ako } stop(w_3) \wedge \neg stop(w_2) \\ \alpha_4 G_0^*(g, w_1^4, \emptyset) & \text{inače} \end{cases}$$

- 1 Kolokacije i višerječne jedinice
- 2 Ekstrakcija kolokacija iz korpusa
- 3 Vrednovanje ekstrakcije**
- 4 Genetičko programiranja i ekstrakcija kolokacija
- 5 Alat TermeX
- 6 Zaključak

Načini vrednovanja

- Ne postoji usuglašeni način vrednovanja
- Mogućnosti:
 - 1 Usporedba ekstrahiranih kolokacija s popisom (leksikon, WordNet)
 - 2 Ručno ocjenjivanje ekstrahiranih kolokacija
 - 3 **Vrednovanje na uzorku**
- Evaluacijske mjere:

$$F_1 = \frac{2PR}{P + R} \quad AP = \frac{1}{N+} \sum_{r=1}^N P(r)rel(r)$$

Vrednovanje na uzorku (1)

županije u iznosu
odluku Upravnog vijeća
terora pasa lotalica
Hrvatski filmski savez
veterani i kadeti
lutkarstvo Umjetničke akademije
sustav oružanih snaga
zemlje Srednjeg istoka
kriteriji za dobivanje
neposrednoj blizini mjesta
akcijskog plana vlade
beljskog pogona Poljoprivreda
drugoligaši i trećeligaši
drugih izvora financiranja
zgrada nije etažirana
vrata do vrata
sredstva za opremanje
dogradonačelnik Petar Mlinarić

umirovljenika Ivana Matiševa
tijelima lokalne uprave
lokaciju i gradnja
mjesto u sustavu
operirano slijepo crijevo
izbora u koaliciji
groblju u Vinkovcima
činjenje takvih djela
kuna za otkup
posao za državu
klub od ispadanja
obračuna potrošnje vode
Gradsko kazalište Joza
predsjednik Uprave HT-a
stožer osječkog prvoligaša
inozemni trgovački lanci
nadogradnja i prenamjena
subote ili nedjelje

Vrednovanje na uzorku (1)

županije u iznosu
odluku Upravnog vijeća
terora pasa lutalica
Hrvatski filmski savez
veterani i kadeti
lutkarstvo Umjetničke akademije
**sustav oružanih snaga
zemlje Srednjeg istoka**
kriteriji za dobivanje
neposrednoj blizini mjesta
akcijskog plana vlade
beljskog pogona Poljoprivreda
drugoligaši i trećeligaši
drugih izvora financiranja
zgrada nije etažirana
vrata do vrata
sredstva za opremanje
dogradonačelnik Petar Mlinarić

umirovljenika Ivana Matiševa
tijelima lokalne uprave
lokaciju i gradnja
mjesto u sustavu
operirano slijepo crijevo
izbora u koaliciji
groblju u Vinkovcima
činjenje takvih djela
kuna za otkup
posao za državu
klub od ispadanja
obračuna potrošnje vode
Gradsko kazalište Joza
predsjednik Uprave HT-a
stožer osječkog prvoligaša
inozemni trgovački lanci
nadogradnja i prenamjena
subote ili nedjelje

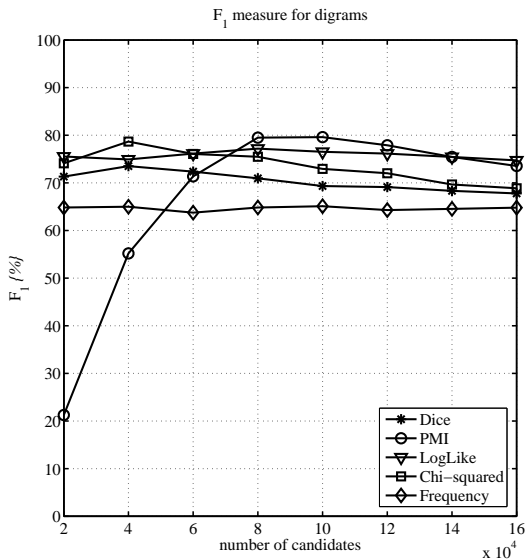
Vrednovanje na uzorku (2)

- Problem subjektivnosti
- Označavanje treba provesti **više označivača** te je potrebno odrediti **međusobno podudaranje** (engl. *inter-annotator agreement*, IAA)
- Npr. $\hat{\kappa}$ -koeficijent (Cohen, 1960):

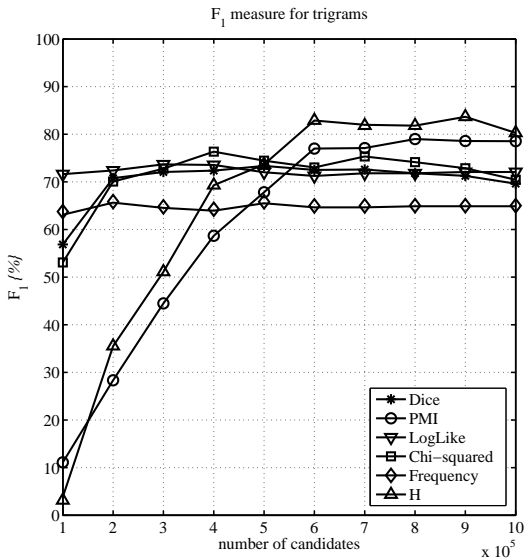
$$\hat{\kappa} = \frac{p_o - p_c}{1 - p_c} \quad \hat{\kappa} \in [-1, 1]$$

- Podudaranje se smatra zadovoljavajućim ako $\hat{\kappa} \geq 0.67$ (diskutabilno!)
- (Petrović i dr., 2010): $\hat{\kappa} = 0.729 \pm 0.056$ za trigramme i
 $\hat{\kappa} = 0.726 \pm 0.062$ za tetragrame
- (Delač i dr., 2009): $\hat{\kappa}$ ovisi o vrsti MWE (najmanje podudaranje kod terminoloških izraza, a najveće kod vlastitih imena)

Rezultati – bigrami

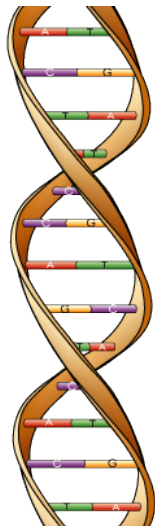


Rezultati – trigrami



- 1 Kolokacije i višerječne jedinice
- 2 Ekstrakcija kolokacija iz korpusa
- 3 Vrednovanje ekstrakcije
- 4 Genetičko programiranja i ekstrakcija kolokacija**
- 5 Alat TermeX
- 6 Zaključak

Genetičko programiranje



- **Genetički algoritam** – računalni postupak koji oponaša biološku evoluciju u svrhu rješavanja složenih optimizacijskih problema
 - ▶ populacija rješenja
 - ▶ funkcija dobrote
 - ▶ selekcija
 - ▶ križanje
 - ▶ mutacija
- Stohastičko pretraživanje prostora rješenja
- **Genetičko programiranje** – uporaba genetičkih algoritama za evoluiranje računalnih programa
 - ▶ računalni programi prikazani stablastom podatkovnom strukturom

Evoluiranje mjere leksičke asocijacije

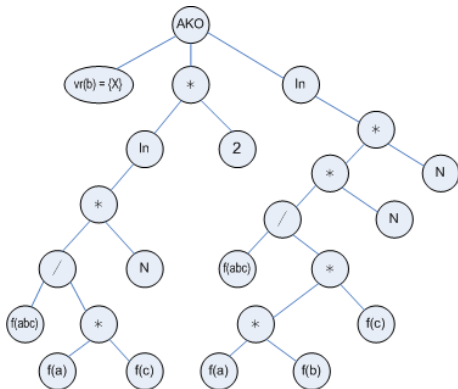
Ideja: genetičkim programiranjem **evoluirati mjeru leksičke asocijacije** (Šnajder i dr., 2008)

- Listovi stabla:

- ▶ konstante
- ▶ frekvencije n-grama
- ▶ vrsta riječi

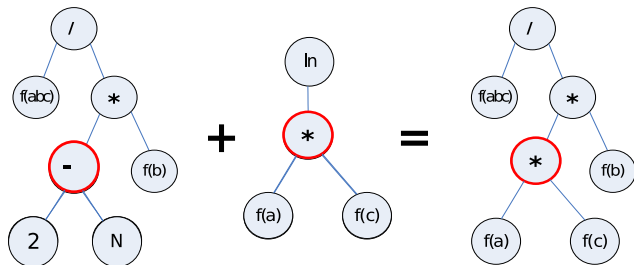
- Unutarnji čvorovi:

- ▶ aritmetički operatori:
+, -, log
- ▶ operator grananja:
“*i*-ta riječ je vrste T”



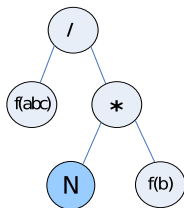
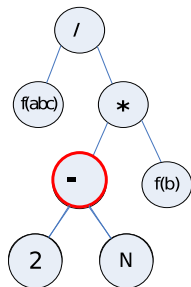
Križanje mjera

- **Razmjena genetičkog materijala** – *eksploatacija rješenja*
- Dvije mjere (roditelji) se kombiniraju i daju novu, treću mjeru (dijete)
- Razmjena dvaju slučajno odabranih podstabala roditelja

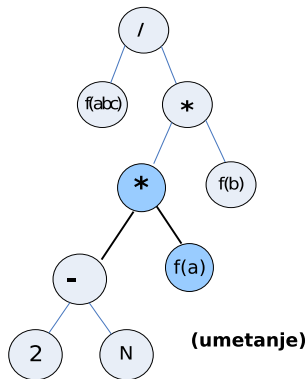


Mutacija mjera

- **Unošenje novog genetičkog materijala** – *eksploracija rješenja*
- Brisanje slučajno odabranog čvora (25% vjerojatnosti)
- Umetanje čvora na slučajno odabranoj poziciji (75% vjerojatnosti)



(brisanje)

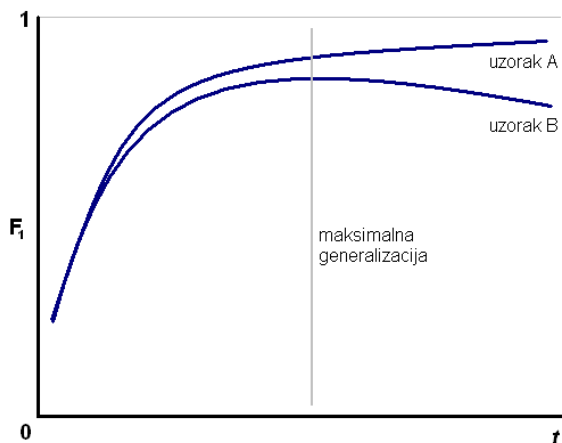


(umetanje)

- Dobrota mjere leksičke asocijacije izražena je mjerom F_1
 - ▶ mjereno na uzorku od **100 pozitivnih/100 negativnih** primjera
 - ▶ n-grami poredani prema mjeri leksičke asocijacije
 - ▶ n prvorangiranih smatraju se pozitivnim, $n = [1, 200]$
 - ▶ najbolja vrijednost F_1 uzeta kao mjera dobrote
- **Parsimonijski pritisak**
 - ▶ sprječava neograničeni rast stabla

$$fitness(j) = F_1(j) + \eta \frac{L_{max} - L(j)}{L_{max}}$$

Zaustavljanje evolucije



- dostizanje k iteracija bez poboljšanja dobrote na **uzorku za provjeru**

Eksperiment (1)

- **Cilj:** evoluirati 3-gramske mjere asocijacije za ekstrakciju kolokacija iz korpusa pravnih propisa
- **Korpus:** 7008 dokumenata Narodnih novina
 - ▶ opojavničen, lematiziran, označene vrste riječi
- Ručno označeni uzorci n-grama
 - ▶ **vrednovanje:** 100 pozitivnih/100 negativnih
 - ▶ **provjera:** 100 pozitivnih/100 negativnih
 - ▶ **pozitivni:** leksičke kolokacije, terminološki izrazi, vlastita imena

Eksperiment (2)

Početna populacija:

- 50 – 50,000 slučajno generiranih rješenja (stabala)
- poznate mjere imaju 50% vjerojatnosti da budu uključene:
 - ▶ međusobna informacija
 - ▶ Diceov koeficijent
 - ▶ heuristička mjera H (Petrović i dr., 2006):

$$H(a, b, c) = \begin{cases} 2 \log \frac{f(abc)}{f(a)f(c)} & \text{ako } stop(b), \\ \log \frac{f(abc)}{f(a)f(b)f(c)} & \text{inače.} \end{cases}$$

Parametri:

- troturnirska selekcija, parsimonija η : $[0, 0.05]$
- vjerojatnost mutacije: $[0.0001, 0.3]$,
- 800 pokretanja sa slučajno odabranim vrijednostima parametara

Rezultati

- **20%** mjera ima dobrotu $F_1 > 80\%$
 - ▶ 23% ako se poznate mjere uključe u početnu populaciju, 13% ako ih se ne uključi
- Ukupno najbolja: $F_1 = 88.4\%$ s 205 čvorova

$$\begin{aligned}
& f(abc) f(a) f(c) * / f(abc) f(ab) f(c) - f(c) f(bc) f(b) \\
& -f(abc) + / + / N * f(b) + * \ln f(c) f(b) * * N f(a) * \\
& f(abc) f(a) f(abc) f(a) f(c) * / f(bc) * f(bc) f(b) + / \\
& f(a) N IF(vr(b)=\{X\}) * (-14.426000) f(b) + / N * f(bc) f(b) \\
& -(2.000000) * \ln \ln / f(a) f(c) * (2.000000) * \ln \ln / N * \\
& \ln * / f(bc) * f(bc) f(b) + / N * (-14.426000) f(b) + / N * \\
& f(abc) N f(a) * f(a) f(abc) f(a) f(c) * / f(bc) * f(abc) \\
& f(b) + / N * (-14.426000) f(b) + / N * f(b) f(c) * \ln \ln / \\
& f(abc) f(a) f(c) * / f(c) * \ln \ln (2.000000) * \ln \ln / N * \\
& / N * / N * \ln f(c) * / f(a) f(b) + * \ln \ln f(abc) f(abc) \\
& f(a) f(a) N IF(vr(b)=\{X\}) (-14.426000) f(b) + * / N * / N * \\
& \ln f(c) * / f(a) f(b) + * \ln \ln * \ln \ln / f(abc) f(a) f(c) \\
& * / f(a) f(b) + * \ln \ln (2.000000) * \ln \ln / N * \ln \ln \\
& IF(vr(c)=\{X\}) N * IF(vr(b)=\{X\})
\end{aligned}$$

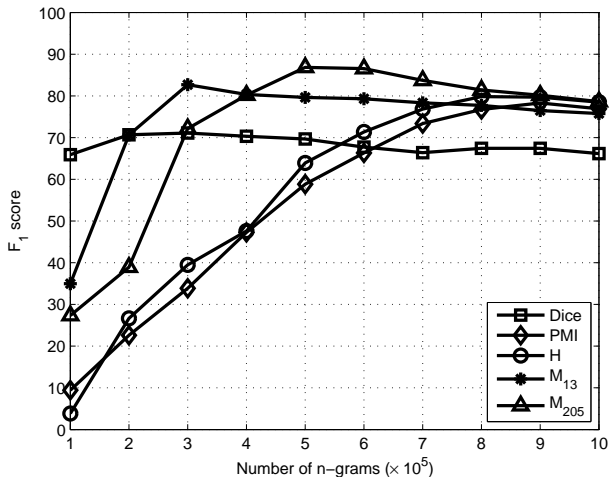
Rezultati (1)

- **20%** mjera ima dobrotu $F_1 > 80\%$
 - ▶ 23% ako se poznate mjere uključe u početnu populaciju, 13% ako ih se ne uključi
- Ukupno najbolja: $F_1 = 88.4\%$ s 205 čvorova
- Jedno od optimalnih rješenja s manje od 30 čvorova:

$$M_{13}(a, b, c) = \begin{cases} -0.423 \frac{f(a)f(c)}{f^2(abc)} & \text{ako } stop(b) \\ 1 - \frac{f(b)}{f(abc)} & \text{inače} \end{cases}$$

- ▶ mjera M_{13} je evolvirala a da mjera H nije bila uključena u početnu populaciju!
- ▶ 96/100 visoko rangiranih mjera slične su strukture
- potvrđuje tezu da trigrame sa zaustavnim riječima treba tretirati drugačije (Petrović i dr., 2006)

Rezultati (2)



- Mjere M_{13} i M_{205} nadmašuju preostale tri mjere leksičke asocijacije

- 1 Kolokacije i višerječne jedinice
- 2 Ekstrakcija kolokacija iz korpusa
- 3 Vrednovanje ekstrakcije
- 4 Genetičko programiranja i ekstrakcija kolokacija
- 5 Alat TermeX**
- 6 Zaključak

TermeX – Terminology Extraction Tool

The screenshot shows the TREME X application window titled "TREMEX: test3.txt". The interface includes a menu bar with "Projekt", "Klokacije", and "Pomoć". Below the menu bar, there are input fields for "Mjera:" (set to "PMI"), "Klokacija:" (set to "2-gram"), and "Broj:" (set to "1.000"). A gear icon is visible to the right of these fields. The main area contains a table with the following columns: "Rang", "NGram", "Vrijednost", and "Klokacija". The table lists 18 n-grams, with the 5th, 8th, and 13th rows highlighted in blue. To the right of the table is a panel titled "Označene kolokacije:" containing a list of terms: "Neven viđen", "Robert Posavec", "dečko pokloni", and "glasićem Matijine". At the bottom of the window, the text "Spreman..." is visible.

Rang	NGram	Vrijednost	Klokacija
0	Kombinacija apaurina	14,402	<input type="checkbox"/>
1	Matijine sestre	14,402	<input type="checkbox"/>
2	Neven viđen	14,402	<input checked="" type="checkbox"/>
3	Ogromna kuća	14,402	<input type="checkbox"/>
4	Posavec Životopis	14,402	<input type="checkbox"/>
5	Robert Posavec	14,402	<input checked="" type="checkbox"/>
6	Stazama orgazma	14,402	<input type="checkbox"/>
7	Tržišni dodatak	14,402	<input type="checkbox"/>
8	dečko pokloni	14,402	<input checked="" type="checkbox"/>
9	dnevnoj sobi	14,402	<input type="checkbox"/>
10	dodatak životopisu	14,402	<input type="checkbox"/>
11	enciklopedije natuknicu	14,402	<input type="checkbox"/>
12	genijalnim izlascima	14,402	<input type="checkbox"/>
13	glasićem Matijine	14,402	<input checked="" type="checkbox"/>
14	hard core.	14,402	<input type="checkbox"/>
15	ispisano GARI	14,402	<input type="checkbox"/>
16	išarana grafitima	14,402	<input type="checkbox"/>
17	jednakoj mjeri	14,402	<input type="checkbox"/>
18	jednaku reakciji	14,402	<input type="checkbox"/>

(Delač i dr., 2009) <http://ktlab.fer.hr/termex/>

- 1 Kolokacije i višerječne jedinice
- 2 Ekstrakcija kolokacija iz korpusa
- 3 Vrednovanje ekstrakcije
- 4 Genetičko programiranja i ekstrakcija kolokacija
- 5 Alat TermeX
- 6 Zaključak

- **Višerječni izrazi** bitni su za NLP/IR/TM
- Višerječne izraze moguće je identificirati na temelju statističke analize supojavljanja – **kolokacije**
- **Mjere leksičke asocijacije** mogu se upotrijebiti za ekstrahiranje kolokacija iz korpusa
- Za izraze s više od dvije riječi potrebna su **proširenja** uobičajenih mjera asocijacije
- Vrednovanje postupka u obzir mora uzeti **subjektivnost**
- **Genetičko programiranje** može se upotrijebiti za evoluiranje novih mjera asocijacije
- Nastavak istraživanja: kolokacije u TM/IR, sintaktičke značajke, semantički modeli kolokacija, . . .

Reference

(Potpuniji popis referenci može pronaći u navedenim člancima)

- Delač, D., Krleža, Z., Dalbelo Bašić, B., Šnajder, J., Šarić, F. TermeX: A Tool for Collocation Extraction. (2009) *Lecture Notes in Computer Science (Computational Linguistics and Intelligent Text Processing)*. 5449, 149–157.
- Kolar, M., Vukmirović, I., Dalbelo Bašić, B., Šnajder, J. (2005). Computer-Aided Document Indexing System. *Journal of Computing and Information Technology*, 13(4), 299–305.
- Petrović, S., Šnajder, J., Dalbelo Bašić, B. (2010). Extending lexical association measures for collocation extraction. *Computer Speech and Language* 24(2), 383–394.
- Petrović, S., Šnajder, J., Dalbelo Bašić, B., Kolar, M. (2006). Comparison of Collocation Extraction Measures for Document Indexing. *Journal of Computing and Information Technology*, 14(4), 321–327.
- Šnajder, J., Dalbelo Bašić, B., Petrović, S., Sikirić, I. (2008). Evolving New Lexical Association Measures Using Genetic Programming. *Proceedings of ACL-08: HLT, Short Papers*, 181–184.