Ako koristite ovaj program, molim vas, citirajte: Filipović Đurđević, D., Đurđević, Đ. i Kostić, A. (2009). Vector based semantic analysis reveals absence of competition among related senses. *Psihologija,* 42(1), 95-106. doi:10.2298/PSI0901095F

PREGLED FUNKCIJA

VEKTORI_PRVOG_REDA	pravi vektore prvog reda za pojedinačno pojavljivanje svake reči, slučaj polisemije
VEKTORI_PRVOG_REDA	pravi standardne vektore prvog reda (jedna reč, jedan red)
VEKTORI_DRUGOG_REDA	pravi vektore drugog reda za svako pojavljivanje reči, slučaj polisemije
EKSTRAKCIJA_OKOLNIH_RECI	za svako pojavljivanje zadatih reči ispisuje prethodnih i sledećih n reči
KOSINUSNA_DISTANCA	za date vektore pravi kvadratnu, simetričnu matricu kosinusnih distance
NORMALIZACIJA_VEKTORA	svaki vektor prevodi u vektor jedinične dužine, radi po vrstama
SLUCAJNI_IZBOR	po slučajnom principu bira n vektora
STANDARDIZACIJA	za date vektore pravi vektore sa z skorovima, računato po kolonama, ili vrstama
SEMANTIČKO RAZDVAJANJE	Uf, moram da se setim… (Mislim da računa cos.dist između pojedinačnih vektora i koordinata centroida klastera koji su dobijeni u zasebnoj klaster analizi i za svaki vektor daje podatak o tome koji centroid mu je najbliži. Dakle, nešto kao pripisivanje pripadnosti klasteru, ali nisam sigurna, moram da proverim...)

NAPOMENA: Pod pojmom reč podrazumeva se površinski oblik reči, odnosno niz slova, string, ono što je uneto u spisak reči za koje se prave vektori. Ukoliko je potrebno napraviti vektore prvog reda za leme, onda se pored programa "vektori" koristi i program "vektori2" koji služi za spajanje vektora inflektivnih oblika u vektor za lemu o čijim inflektivnim oblicima se radi.

Izlazni fajl se formira tako što se redosled kolona čuva, odnosno ostaje isti kao redosled u fajlu koji sadrži kontekst reči, dok se redosled redova razlikuje od izvornog redosleda u fajlu koji sadrži mete. Taj redosled se formira tako što se reči ređaju onim redosledom kojim se pronađu u bazi. Ukoliko se primeni sažimanje reči u leme (vektori2), dobijeni redosled je kao u fajlu koji sadrži mapiranje reči na leme.

FONTOVI

Veoma je važno da sva slova budu napisana na način na koji su napisana u bazi Ebart. U slučaju baze koju sam ja sredila, radi se o Western European (Windows) kodnom rasporedu (ASCII, nije Unicode), a mapiranje je:

Slovo	Za Ebart	Slovo	Za Ebart
NJ	NJ	nj	nj
LJ	LJ	lj	lj
Č	È	č	è
Ć	Æ	ć	æ
Š	Š	š	š
Ž	Ž	ž	ž
Đ	Đ	đ	ð
DŽ	DŽ	dž	dž

Svi fajlovi treba da budu txt (Tab delimited).

U bazi EBART.novi.txt (u ovoj bazi su zamenjeni znaci za nasa slova i eliminisana su velika slova koja su ostala u staroj bazi) mapiranje izgleda ovako:

Slovo	Za Ebart	Slovo	Za Ebart
NJ	/	nj	nj
LJ	/	lj	lj
Č	/	č	~
Ć	/	ć	}
Š	/	š	{
Ž	/	ž	`
Đ	/	đ	|
DŽ	/	dž	d`

FORMIRANJE (STANDARDNIH) VEKTORA PRVOG REDA ZA LEME (ZBIRNO ZA RAZLIČITE INFLEKTIVNE OBLIKE)

1. KORAK – formiranje vektora prvog reda za pojedinačne oblike reči		
Program	**vektori.exe** koristiti sheet "**formiranje vektora**", podopciju "**vektori prvog reda (standardni)**"	
Ulazni fajlovi	Recnik	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalazi kontinuirani tekst, u našem slučaju radi se o Ebart medijskoj bazi, odnosno fajlu koji se zove ebart.txt.
	Spisak ciljnih reci	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalaze ciljne reči, tačnije svi inflektivni oblici svake ciljne leme, unete jedna ispod druge.
	Kontekst	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalaze, jedna ispod druge, kontekst reči, u našem slučaju radi se o hiljadu najfrekventnijih reči.
Izlazni fajlovi	Izlaz	Fajl u .txt formatu koji sadrži matricu dimenzija m x n, gde m predstavlja broj ciljnih reči, a n predstavlja broj kontekst reči. Pored toga, u prvoj koloni ovog fajla nalazi se ciljna reč na koju se odnose vrednosti vektora u datom redu. Svaki red predstavlja standardni vektor prvog reda za odgovarajuću ciljnu reč (elementi tog vektora su frekvencije zajedničkog pojavljivanja date ciljne reči i kontekst reči iz odgovarajuće kolone). Primetiti da ciljne reči nisu leme, već inflektivni oblici, odnosno oni tekstualni nizovi koji su uneti u spisak ciljnih reči.
Ostale specifikacije	Velicina prozora	Ovde je moguće definisati broj mesta levo i desno od ciljne reči na kojima će biti tražena kontekst reč. Standardna i istovremeno default vrednost je +/- 3.
	Kolona	Za svaki ulazni fajl moguće je definisati kolonu u kojoj se nalaze potrebne vrednosti. Default vrednost je 1, što znači da se potrebne vrednosti nalaze u prvoj koloni (pošto je prethodno preporučeno da se svi fajlovi sastoje iz jedne kolone). Dakle, ova vrednost ostaje 1.

2. KORAK – spajanje vektora pojedinačnih oblika u vektore za leme		
Program	**vektori2.exe**	
Ulazni fajlovi	Pojedinacni vektori	Fajl u .txt formatu koji je dobijen kao izlaz u prvom koraku, odnosno fajl koji sadrži standardne vektore prvog reda za inflektivne oblike ciljnih reči, čija prva kolona sadrži te oblike.
	Mapiranje	Fajl u .txt formatu koji sadrži dve kolone. Druga kolona identična je prvoj koloni fajla "Spisak ciljnih reči", a prva kolona sadrži podatak o lemi čiji oblik se nalazi u odgovarajućem redu u drugoj koloni. *Primer:* *BURA bura* *BURA bure* *BURA buri* *BURA buru* *BURA burom* *BURA burama*
Izlazni fajlovi	Zbirni vektori	Fajl u .txt formatu koji sadrži matricu dimenzija m x n, gde m predstavlja broj ciljnih **lema**, a n predstavlja broj kontekst reči. Pored toga, u prvoj koloni ovog fajla nalazi se ciljna lema na koju se odnose vrednosti vektora u datom redu. Svaki red predstavlja standardni vektor prvog reda za odgovarajuću ciljnu lemu (elementi tog vektora su frekvencije zajedničkog pojavljivanja date ciljne leme i kontekst reči iz odgovarajuće kolone). Dakle, ovoga puta radi se o vektorima za leme (zbirno za različite inflektivne oblike).

<table>
<tr><th colspan="4">3. KORAK – računanje kosinusnih distanci</th></tr>
<tr><td>Program</td><td colspan="3">vektori.exe
koristiti sheet "Kosinusna distanca"</td></tr>
<tr><td>Ulazni fajlovi</td><td>Matrica</td><td colspan="2">Fajl u .txt formatu koji je dobijen kao izlaz u drugom koraku, dakle zbirni vektori, odnosno standardni vektori prvog reda za reči leme.</td></tr>
<tr><td>Izlazni fajlovi</td><td></td><td colspan="2">Kvadratna, simetrična matrica dimenzija m x m, gde m predstavlja broj ciljnih lema. Elementi matrice predstavljaju vrednost kosinusne distance između lema u odgovarajućem redu i koloni. Elementi na dijagonalama ove matrice treba da budu 1, jer se u tom slučaju lema poredi sama sa sobom, odnosno ista lema se nalazi u datom redu i datoj koloni.</td></tr>
<tr><td>Ostale specifikacije</td><td>Kolona</td><td colspan="2">Za svaki ulazni fajl moguće je definisati kolonu u kojoj počinju potrebne vrednosti. Default vrednost je 1, što bi značilo da potrebne vrednosti počinju već u prvoj koloni. Pošto je ulazni fajl za ovo računanje često izlazna matrica iz koraka 2, čiju prvu kolonu predstavljaju reči, vrednost kolone treba podesiti na 2, jer vrednosti vektora počinju u drugoj koloni.</td></tr>
<tr><th colspan="4">4. KORAK – selekcija cosinusnih distanci za odgovarajuće parove</th></tr>
<tr><td>Program</td><td colspan="3">Svaki koji može da posluži, npr. Excel</td></tr>
<tr><td>Ulazni fajl</td><td>Matrica kosinusnih distanci</td><td colspan="2">Fajl sa matricom koja je dobijena kao izlaz u trećem koraku.</td></tr>
<tr><td>Izlazni fajl</td><td>Vektor kosinusnih distanci</td><td colspan="2">Cilj je dobiti vektor sa vrednostima kosinusnih distanci između parova primova i meta, odnosno vektor koji će biti korišćen u analizi.
U ovom koraku se još uvek treba snaći kako ko zna i ume. Moja tehnika za sada podrazumeva kontrolisanje redosleda kojim su unete reči – uvek se drzim istog redosleda primova i meta, prvo primovi pa mete, tako da mogu lako da izaberem vrednosti u ekselu. Na primer:
U ovoj matrici kosinusnih distanci lako je uočiti princip po kojem se selektuju vrednosti koje su potrebe za analizu.
<table>
<tr><td></td><td>prim1</td><td>prim2</td><td>...</td><td>meta1</td><td>meta2</td></tr>
<tr><td>prim1</td><td>1</td><td>a</td><td></td><td>b</td><td>c</td></tr>
<tr><td>prim2</td><td>a</td><td>1</td><td></td><td>d</td><td>e</td></tr>
<tr><td>...</td><td></td><td></td><td>1</td><td></td><td></td></tr>
<tr><td>meta1</td><td>b</td><td>d</td><td></td><td>1</td><td>f</td></tr>
<tr><td>meta2</td><td>c</td><td>e</td><td></td><td>f</td><td>1</td></tr>
</table>
</td></tr>
</table>

FORMIRANJE VEKTORA DRUGOG REDA ZA POJEDINAČNA POJAVLJIVANJA REČI (FLEKTIVNIH OBLIKA)

1. KORAK – formiranje vektora prvog reda za pojedinačna pojavljivanja oblika reči		
Program	**vektori.exe** koristiti sheet "**formiranje vektora**", podopciju "**vektori prvog reda (pojedinačno pojavljivanje)**"	
Ulazni fajlovi	Recnik	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalazi kontinuirani tekst, u našem slučaju radi se o Ebart medijskoj bazi, odnosno fajlu koji se zove ebart.txt.
	Spisak ciljnih reci	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalaze ciljne reči, tačnije svi inflektivni oblici svake ciljne leme, unete jedna ispod druge.
	Kontekst	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalaze, jedna ispod druge, kontekst reči, u našem slučaju radi se o hiljadu najfrekventnijih reči.
Izlazni fajlovi	Izlaz	Fajl u .txt formatu koji sadrži matricu dimenzija m x n, gde m predstavlja broj pojedinačnih pojavljivanja ciljnih reči, a n predstavlja broj kontekst reči. Pored toga, u prvoj koloni ovog fajla nalazi se ciljna reč na koju se odnose vrednosti vektora u datom redu, a u drugoj koloni redni broj pojavljivanja nadjenog oblika. Svaki red predstavlja vektor prvog reda za pojedinačno pojavljivanje odgovarajuće ciljne reči (elementi tog vektora su nule i jedinice, gde nula znači da se data ciljna reč i kontekst reč nisu pojavile unutar prozora zajedno, a jedinica znači da su se unutar prozora zajedno našle data ciljna reč i kontekst reč iz odgovarajuće kolone). Primetiti da ciljne reči nisu leme, već inflektivni oblici, odnosno oni tekstualni nizovi koji su uneti u spisak ciljnih reči.
Ostale specifikacije	Velicina prozora	Ovde je moguće definisati broj mesta levo i desno od ciljne reči na kojima će biti tražena kontekst reč. Standardna i istovremeno default vrednost je +/- 3.
	Kolona	Za svaki ulazni fajl moguće je definisati kolonu u kojoj se nalaze potrebne vrednosti. Default vrednost je 1, što znači da se potrebne vrednosti nalaze u prvoj koloni (pošto je prethodno preporučeno da se svi fajlovi sastoje iz jedne kolone). Dakle, ova vrednost ostaje 1.

2. KORAK – formiranje standardnih vektora prvog reda za odabrane kontekst reči u kontektu istih tih kontekst reči		
Program	**vektori.exe** koristiti sheet "**formiranje vektora**", podopciju " **vektori prvog reda (standardni)**"	
Ulazni fajlovi	Recnik	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalazi kontinuirani tekst, u našem slučaju radi se o Ebart medijskoj bazi, odnosno fajlu koji se zove ebart.txt.
	Spisak ciljnih reci	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalaze, jedna ispod druge, kontekst reči, u našem slučaju radi se o hiljadu najfrekventnijih reči.
	Kontekst	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalaze, jedna ispod druge, kontekst reči, u našem slučaju radi se o hiljadu najfrekventnijih reči. U pitanju je isti fajl kao u prethodnom polju!
Izlazni fajlovi	Izlaz	Fajl u .txt formatu koji sadrži matricu dimenzija n x n, gde n predstavlja broj kontekst reči. Pored toga, u prvoj koloni ovog fajla nalazi se reč na koju se odnose vrednosti vektora u datom redu. Svaki red predstavlja standardni vektor prvog reda za odgovarajuću ciljnu reč (elementi tog vektora su frekvencije zajedničkog pojavljivanja date ciljne reči i kontekst reči iz odgovarajuće kolone). Primetiti da ciljne reči nisu leme, već inflektivni oblici, odnosno oni tekstualni nizovi koji su uneti u spisak ciljnih reči. Ne zaboraviti da su iste reči ovde bile i ciljne reči i kontekst reči.
Ostale specifikacije	Velicina prozora	Isto kao u prethodnom koraku
	Kolona	Isto kao u prethodnom koraku

3. KORAK – formiranje vektora drugog reda za pojedinačna pojavljivanja oblika reči		
Program	**vektori.exe** koristiti sheet "**formiranje vektora**", podopciju "**vektori drugog reda**" **Ukoliko su svi parametri za prethodna dva koraka istovremeno definisani, moguće je kliknuti na opciju** "**paketska obrada**" čime će ulazni fajlovi za treći korak biti automatski definisani. Tada je potrebno dodatno definisati izlazni fajl za treći korak. Na ovaj način, tri koraka se izvršavaju istovremeno.	
Ulazni fajlovi	Kontekst	Fajl u .txt formatu koji sadrži jednu kolonu u kojoj se nalaze, jedna ispod druge, kontekst reči, u našem slučaju radi se o hiljadu najfrekventnijih reči. U pitanju je isti fajl sa kontekst rečima, kao u prethodnim koracima!
	Matrica 1	Izlazni fajl iz prvog koraka.
	Matrica 2	Izlazni fajl iz drugog koraka.
Izlazni fajlovi	Izlaz	Fajl u .txt formatu koji sadrži matricu dimenzija m x n, gde m predstavlja broj pojedinačnih pojavljivanja ciljnih reči, a n predstavlja broj kontekst reči. Pored toga, u prvoj koloni ovog fajla nalazi se ciljna reč na koju se odnose vrednosti vektora u datom redu, a u drugoj koloni redni broj pojavljivanja nadjenog oblika. Svaki red predstavlja vektor drugog reda za pojedinačno pojavljivanje odgovarajuće ciljne reči (elementi tog vektora su sume odgovarajućih elemenata onih standardnih vektora prvog reda /matrica2/ onih kontekst reči koje su se pojavile u susedstvu date ciljne reči /matrica1/). Primetiti da ciljne reči nisu leme, već inflektivni oblici, odnosno oni tekstualni nizovi koji su uneti u spisak ciljnih reči.
Ostale specifikacije	Kolona	Isto kao u prethodnom koraku