Fakultet elektrotehnike i računarstva

TermeX v1.0

Korisničke upute

Autor: Davor Delač

21. siječnja 2009.

Sadržaj

1 Uvod **2**

2 Instalacija za Microsoft Windows **2**

3 Upute za korištenje **3**
3.1 Sučelje 3
3.2 Rad s projektima 4
3.3 Obrada korpusa 5

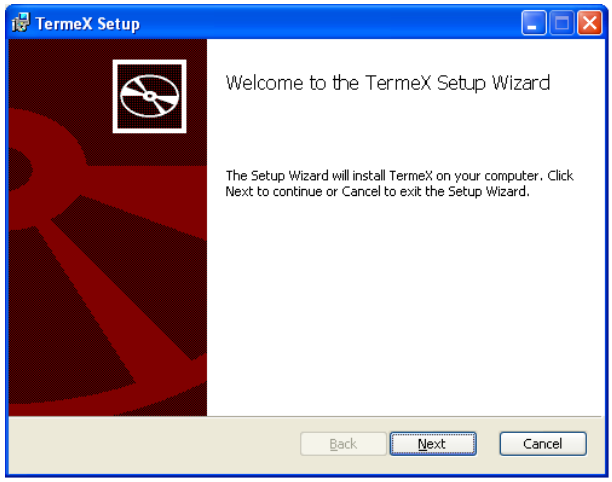
4 Dodatak: Mjere korištene u *TermeX-u* **7**

1 Uvod

TermeX je alat za automatsku ekstrakcju kolokacija i izradu terminoloških leksikona. Alat se temelji na ekstrakciji kolokacija korištenjem statističkih asocijacijskih mjera (AM). Omogućena je ekstrakcija kolokacija do dužine četiri riječi. U *TermeX*-u je implementirano 14 asocijacijskih mjera koje u kombinaciji s lematizacijom daju korisniku širok spektar mogućnosti pri konstrukciji terminološkog leksikona.

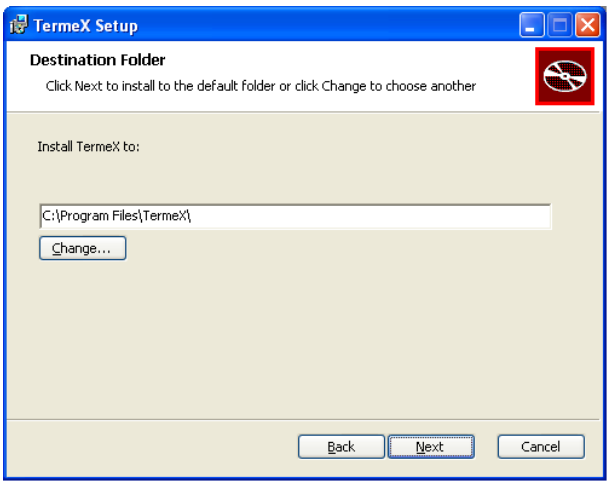
2 Instalacija za Microsoft Windows

TermeX ima korisničko sučelje razvijeno u programskom jeziku Java, stoga je potrebno prije instalacije na željeno računalo postaviti JRE koji se predhodno nabavi na stranicama `http://java.sun.com/javase/downloads/index.jsp`. Pošto je TermeX razvijen u Javi 1.6 najbolje bi bilo imati JRE 1.6 instaliran na računalu. Za instalaciju alata potrebno je pokrenuti instalacijski paket `TermeX.msi`. Nakon pokretanja korisniku se prikazuje prozor na slici 1.



Slika 1: Početak instalacije.

Odabirom opcije Next prelazi se na prozor prikazan slikom 2. U ovom prozoru korisnik odabire lokaciju na disku na koju će se *TermeX* instalirati. Daljnjim odabiranjem opcije Next završit će se instalacija.



Slika 2: Izbor lokacije na disku.

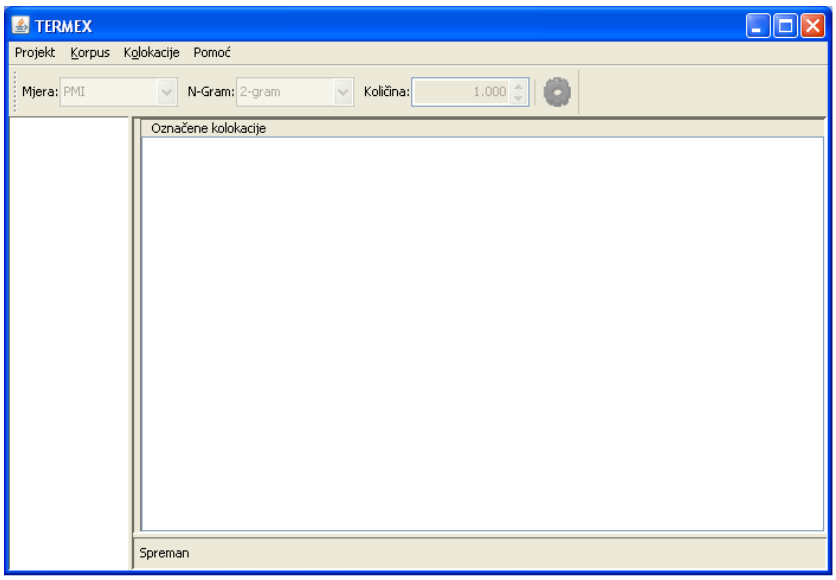
3 Upute za korištenje

TermeX ima jednostavno i intuitivno korisničko sučelje. Rad je podijeljen u projekte. Projekt predstavlja konstrukciju jednog terminološkog leksikona ostvarenu obradom proizvoljnog broja tekstovnih korpusa. *TermeX* nudi širok spektar funkcionalnosti opisan u sljedećim potpoglavljima.

3.1 Sučelje

Grafičko sučelje služi za jednostavni pristup funkcijama alata *TermeX*. Sučelje se sastoji od sljedećih šest dijelova:

1. *Pretraživač projekta* služi za prikaz trenutno otvorenih korpusa koji se obrađuju. Korisniku predstavlja popis otvorenih korpusa;
2. *Radna površina* služi za obradu korpusa. Sastoji se tablica koje predstavljaju rezultate upita i omogućuju odabir kolokacija;
3. *Popis izabranih kolokacija* sačinjavaju one kolokacija koje je korisnik izabrao za uključivanje u terminološki leksikon;
4. *Statusna traka* predstavlja trenutno stanje i napredak izvršavanja upita i obrade korpusa;
5. *Alatna traka* za izvršavanje upita omogućuje brzo i jednostavno izvršavanje upita nad korpusom;
6. *Izbornik*.



Slika 3: Grafičko sučelje.

Izbornik sadrži sljedeće četiri stavke:

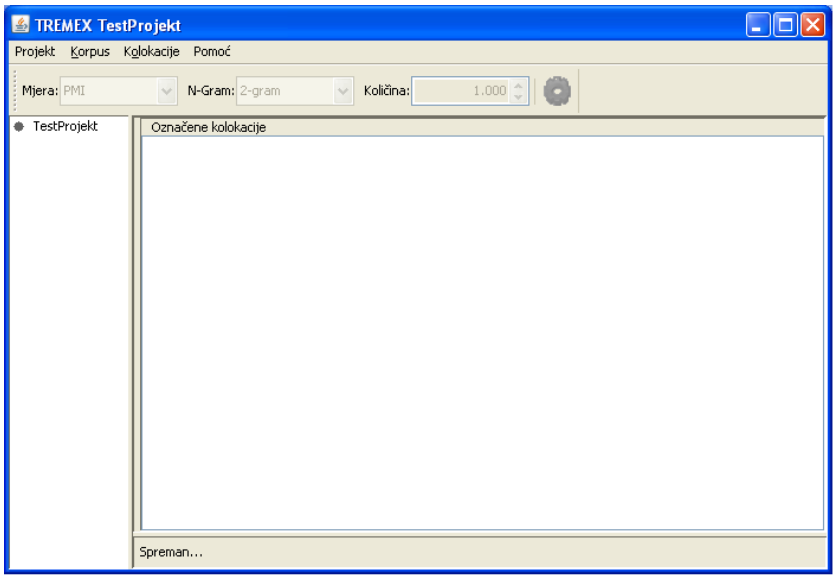
- Projekt – predstavlja niz funkcija za manipulaciju projektima;
- Korpus – nudi funkcije za otvaranje i zatvaranje korpusa unutar projekta;
- Kolokacije – niz funkcija za obradu korpusa;
- Pomoć.

3.2 Rad s projektima

Projekt predstavlja obradu niza tekstovnih korpusa s ciljem stvaranja terminološkog leksikona. Izbornik nudi pristup funkcijama za rad s projektima:

- Novi Projekt – stvara novi projekt. Nakon odabira ove opcije u pretraživaču projekta pojavljuje se novi projekt pod nazivom `untitled`;
- Pohrani – pohranjuje izmjene izvršene nad projektom. Ako se projekt pohranjuje prvi put, funkcionalnost ove stavke ista je kao Pohrani kao...;
- Pohrani Kao... – određivanje datoteke za pohranu i pohranjivanje podataka o projektu u tu datoteku;
- Otvori – obnavlja podatke o projektu iz datoteke;
- Zatvori – zatvara projekt.

Za stvaranje projekta potrebno je odabrati opciju Novi Projekt. Ime projekta mjenja se tako da se pohrani kao datoteka s određenim nazivom. Kao rezultat, u pretraživaču projekta pojavljuje se ime projekta.



Slika 4: Novi projekt.

3.3 Obrada korpusa

Obrada korpusa započinje dodavanjem korpusa projektu. Korpus je tekstovna datoteka kodirana formatom UTF-8. Korpus se dodaje tako da se u izborniku Korpus izabere opcija Dodaj Korpus. Nakon odabira ove opcije otvara se prozor u kojemu se odabire tekstovna datoteka za obradu. Nakon odabira datoteke u statusnoj traci prikazuje se trenutno stanje obrade korpusa, a u pretraživaču projekta pojavljuje se ime korpusa. Kada je računalna obrada završena, korisnik može pristupiti ručnoj obradi dobivenih podataka.

Ručna obrada započinje postavljenjem upita nad korpusom. Upit predstavlja izračun jedne asocijacijske mjere (AM) te sortiranje dobivenih kolokacija prema vrijednosti te mjere. Prije postavljanja upita potrebno je odabrati korpus nad kojim se upit provodi. Odabir se čini dvostrukim klikom miša na željeni korpus, a promjena ikone pored korpusa ukazuje na to da je korpus odabran.

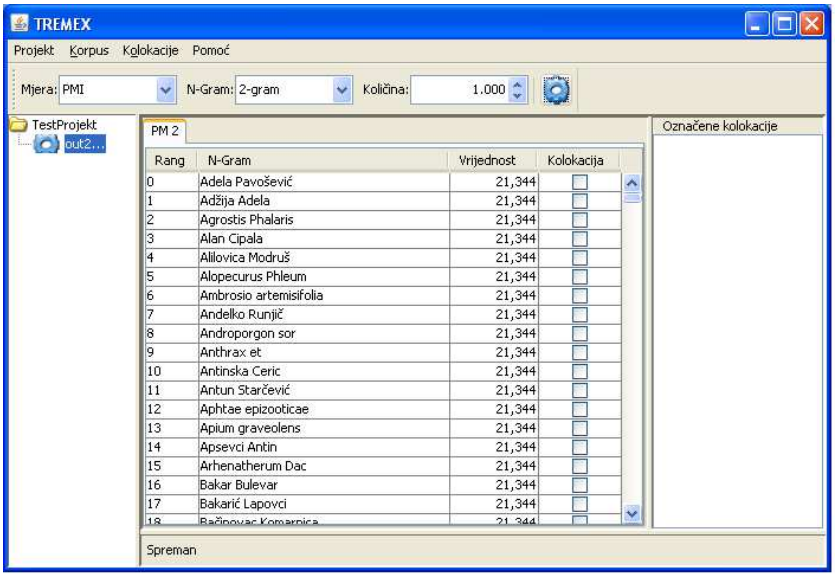
Upiti se izvršavaju pomoću alatne trake za izvršavanje upita. Alatna se traka sastoji od četiri dijela:

- Mjera – odabir asocijativne mjere (AM);

- N-gram – odabir duljine n-grama, odnosno broja riječi u traženim kolokacijama;
- Količina – broj najbolje rangiranih kolokacija prema traženoj mjeri;
- Gumb za izvršavanje upita.

Kao rezultat upita pojaviti će se tablica s podatcima o najbolje rangiranim kolokacijama. Tablica je načinjena od četiri stupca:

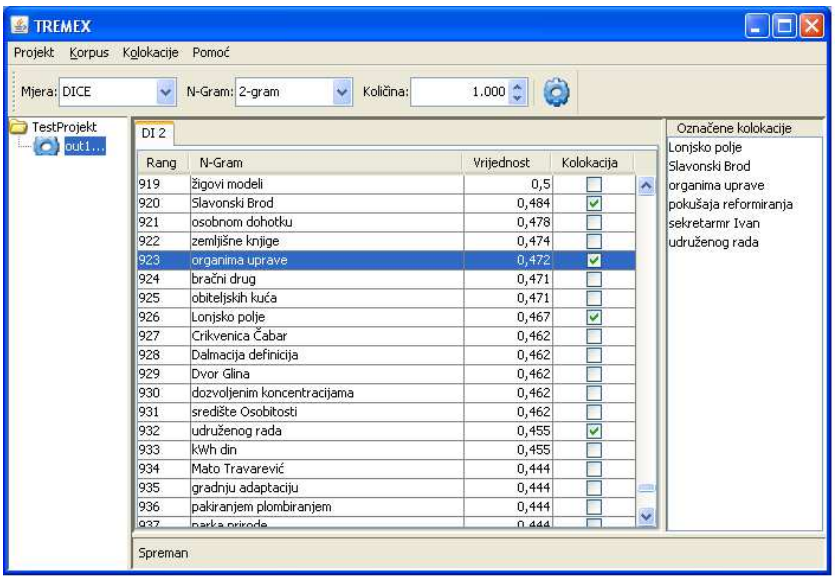
- Rang – rang dobiven primjenom asocijacijske mjere;
- N-Gram – N-gram za koji se računala asocijacijska mjera;
- Vrijednost – vrijednost asocijacijske mjere;
- Kolokacija – kućica u kojoj korisnik označava je li dotični n-gram kolokacija.



Slika 5: Rezultat izvođenja upita.

Kolokacije koje želi preuzeti u svoj leksikon korisnik odabire postavljanjem oznake u kućicu uz zapis u tablici ili pritiskom tipke `Space` dok je dotični zapis tablice odabran. Odabrana kolokacija pojavljuje se u popisu Označene kolokacije te se označava u svim ostalim tablicama za sve druge korpuse.

Pri obradi korpusa korisniku su na raspolaganju sljedeće funkcije iz izbornika Kolokacije:



Slika 6: Obrada korpusa.

- Obnovi listu selekcija – iz formatirane tekstovne datoteke dodaje odabrane kolokacije. Predpostavlja se da je svaka kolokacija u datoteci navedena u zasebnome retku;
- Pohrani listu selekcija – pohranjuje odabrane kolokacije u formatiranu tekstovnu datoteku (svaka kolokacija zapisuje se u zaseban redak);
- Odaberi sve – odabire sve n-grame iz trenutne tablice kao kolokacije;
- Poništi sve – poništava odabir za sve n-grame iz trenutne tablice;
- Pogledaj kontekst – otvara prozor s konkordancijama za trenutno odabrani n-gram. Prozor se sastoji od tablice u kojoj su za dotični n-gram ispisani lijevi i desni kontekst iz korpusa te odmak n-grama od početka korpusa (odmak je izražen u bajtovima).

4 Dodatak: Mjere korištene u *TermeX-u*

$$G_0(I, w_1 \cdots w_n) = \log_2 \frac{P(w_1 \cdots w_n)}{\prod_{i=1}^{n} P(w_i)}, \tag{1}$$

$$G_0(DICE, w_1 \cdots w_n) = \frac{nf(w_1 \cdots w_n)}{\sum_{i=1}^{n} f(w_i)}, \tag{2}$$

$$G_1(g, w_1 \cdots w_n) = \frac{g(w_1, w_2 \cdots w_n) + g(w_1 \cdots w_{n-1}, w_n)}{2}, \tag{3}$$

$$G_2(g, w_1 \cdots w_n) = \frac{g(w_1 \cdots w_{\lfloor n/2 \rfloor}, w_{\lceil n/2 \rceil} \cdots w_n) + g(w_1 \cdots w_{\lfloor n/2+1 \rfloor}, w_{\lceil n/2+1 \rceil} \cdots w_n)}{2}, \tag{4}$$

$$G_3(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_i, w_{i+1}), \tag{5}$$

$$G_4(g, w_1 \cdots w_n) = g(w_1 \cdots w_{n-1}, w_2 \cdots w_n), \tag{6}$$

$$G_5(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_1 \cdots w_i, w_{i+1} \cdots w_n), \tag{7}$$

$$G_6(g, w_1 \cdots w_n) = G_0\Big(g, (w_1w_2, w_2w_3, \ldots, w_{n-1}w_n)\Big), \tag{8}$$

$$G^* : \Omega \times W^+ \times S \to \mathbb{R}, \quad S \subset W. \tag{9}$$

$$G_0^*(I, w_1w_2w_3, \{w_2\}) = \log_2 \frac{P(w_1w_2w_3)}{P(w_1)P(w_3)}, \tag{10}$$

$$H(g, w_1w_2w_3) = \begin{cases} \alpha_1 G_0^*(g, w_1w_2w_3, \{w_2\}) & \text{if } \mathit{stop}(w_2) \\ \alpha_2 G_{4,6}^*(g, w_1w_2w_3, \varnothing) & \text{otherwise,} \end{cases} \tag{11}$$

$$\alpha_1 = \frac{1}{\max\limits_{\{w_1w_2w_3 \in W^3 | \mathit{stop}(w_2)\}} G_0^*(g, w_1w_2w_3, \{w_2\})}, \tag{12}$$

$$\alpha_2 = \frac{1}{\max\limits_{\{w_1w_2w_3 \in W^3 | \mathit{stop}(w_2)\}} G_{4,6}^*(g, w_1w_2w_3, \varnothing)}\ . \tag{13}$$

$$H_1(g, w_1w_2w_3w_4) = \begin{cases} \alpha_1 G_0^*(g, w_1w_2w_3w_4, \{w_2, w_3\}) & \text{if } \mathit{stop}(w_2) \wedge \mathit{stop}(w_3) \\ \alpha_2 G_0^*(g, w_1w_2w_3w_4, \{w_2\}) & \text{if } \mathit{stop}(w_2) \wedge \neg \mathit{stop}(w_3) \\ \alpha_3 G_0^*(g, w_1w_2w_3w_4, \{w_3\}) & \text{if } \mathit{stop}(w_3) \wedge \neg \mathit{stop}(w_2) \\ \alpha_4 G_6^*(g, w_1w_2w_3w_4, \varnothing) & \text{otherwise,} \end{cases} \tag{14}$$

$$H_2(g, w_1w_2w_3w_4) = \begin{cases} \alpha_1 G_0^*(g, w_1w_2w_3w_4, \{w_2, w_3\}) & \text{if } stop(w_2) \wedge stop(w_3) \\ \alpha_2 G_0^*(g, w_1w_2w_3w_4, \{w_1, w_2\}) & \text{if } stop(w_2) \wedge \neg stop(w_3) \\ \alpha_3 G_0^*(g, w_1w_2w_3w_4, \{w_3, w_4\}) & \text{if } stop(w_3) \wedge \neg stop(w_2) \\ \alpha_4 G_6^*(g, w_1w_2w_3w_4, \varnothing) & \text{otherwise,} \end{cases} \tag{15}$$

$$H_3(g, w_1w_2w_3w_4) = \begin{cases} \alpha_1 G_0^*(g, w_1w_2w_3w_4, \{w_2, w_3\}) & \text{if } stop(w_2) \wedge stop(w_3) \\ \alpha_2 G_0^*(g, w_1w_2w_3w_4, \{w_2, w_3\}) & \text{if } stop(w_2) \wedge \neg stop(w_3) \\ \alpha_3 G_0^*(g, w_1w_2w_3w_4, \{w_1, w_3\}) & \text{if } stop(w_3) \wedge \neg stop(w_2) \\ \alpha_4 G_6^*(g, w_1w_2w_3w_4, \varnothing) & \text{otherwise,} \end{cases} \tag{16}$$