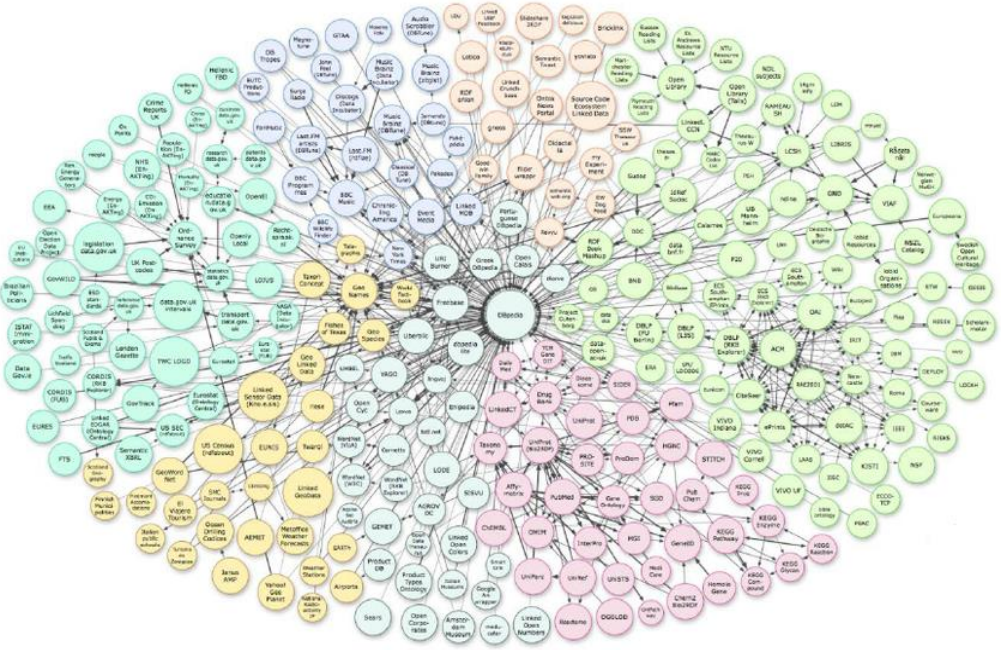
STRUKTURE PODATAKA I ALGORITMI

DRUGI DOMAĆI ZADATAK

(datum objave: 4. 12. 2022.)

Kako omogućiti računaru da razumije tekst? Kako modelovati ljudsko znanje? Ovakvi problemi predstavljaju polje intenzivnog naučno-istraživačkog rada. Često korištena ideja za modelovanje ljudskog znanja jeste leksička baza znanja koja se može predstaviti usmjerenim grafom. U najprostijem slučaju, čvorovi predstavljaju riječi iz određenog rječnika dok grane predstavljaju semantičke relacije između riječi kao što su: hiperonimi (modeluju *is-a* relacije), meronimi (modeluju *part-whole* relacije), sinonimi i slično. Primjer reprezentacije takvog grafa dat je na slici 1.



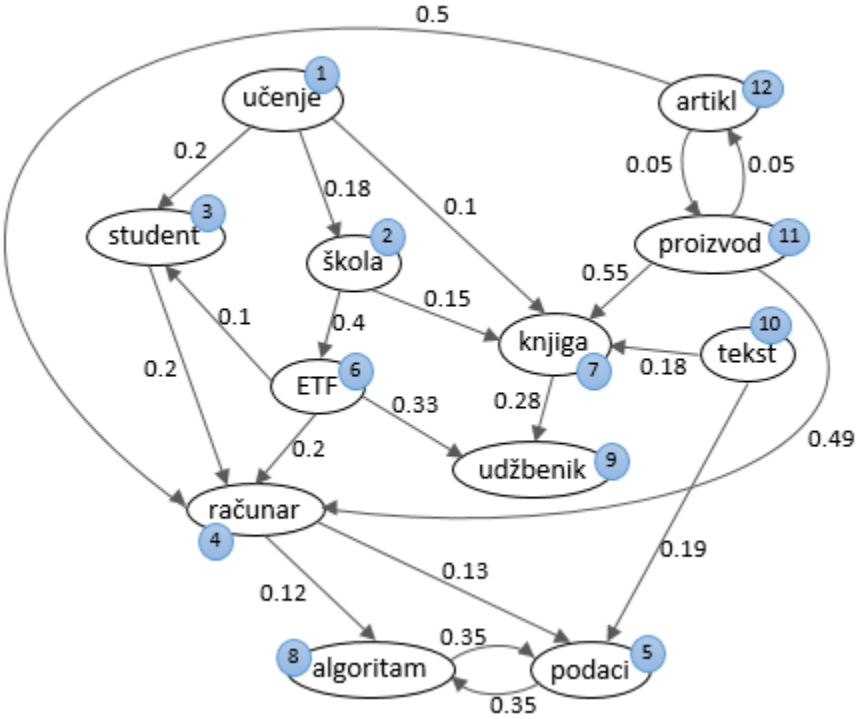
Slika 1. Primjer reprezentacije leksičke baze znanja

Zadatak: Neka je dat usmjeren težinski graf koji predstavlja leksičku bazu znanja. Težine pridružene granama predstavljaju udaljenost između riječi koja modeluje semantičku sličnost na način da riječi sa većom semantičkom sličnosti imaju manju udaljenost i obrnuto. Vrijednosti težina se nalaze u opsegu od 0 do 1. Implementirati rješenje koje će omogućiti:

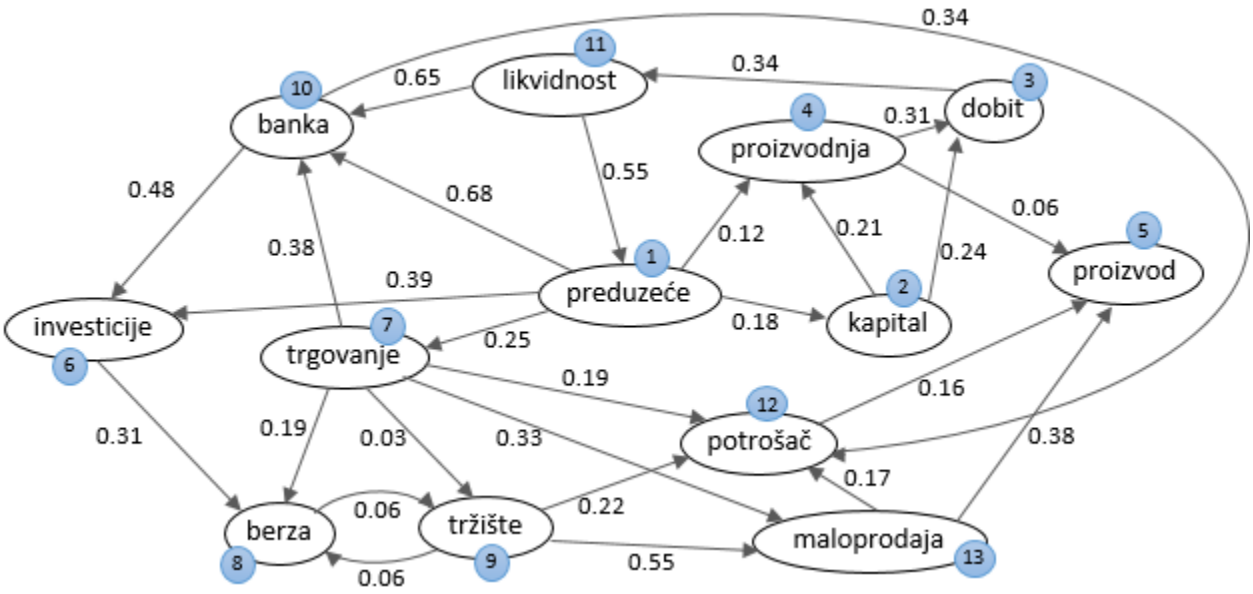
- Učitavanje grafa iz ulaznog fajla čiji se naziv učitava sa standardnog ulaza.
- Predstavljanje grafa odgovarajućom matricom susjednosti ili listom susjednosti (studentima se ostavlja da sami odaberu reprezentaciju grafa koja će bolje odgovarati njihovom rješenju),
- Ispisivanje obilaska grafa (BFS ili DFS načinom) od početnog čvora koji se zadaje putem standardnog ulaza.
- Pronalazak 5 najsličnijih riječi za svaku riječ u grafu. Sličnost riječi se definiše kao zbir sličnosti na putu između dvije riječi. Najsličnija riječ određenoj riječi se dobija pronalaskom riječi sa najkraćim rastojanjem od posmatrane riječi.
- Upis pronađenih riječi i odgovarajućih udaljenosti u izlazni fajl sortirano po neopadajućim vrijednostima udaljenosti. Ukoliko postoje dvije riječi sa istom sračunatom udaljenošću (sličnošću), a različitom udaljenošću u smislu broja grana od posmatrane riječi, pri upisu u fajl prvo se upisuje riječ sa manjom udaljenošću u smislu broja grana, potom riječ sa većom udaljenošću u smislu broja grana od posmatrane riječi.

Rješenje testirati na grafovima *graf1* (slika 2) i *graf2* (slika 3). *Graf1* je dat u ulaznom fajlu **graf1.txt** na *Moodle* platformi za elektronsko učenje. Smatrati da se u okviru ulaznog fajla graf reprezentuje matricom susjednosti po sljedećem formatu: u prvom redu fajla se nalazi broj *n* koji predstavlja broj čvorova u grafu, u drugom redu se nalazi niz riječi (sortiran po identifikatorima čvorova datim na slikama 2 i 3) a zatim slijedi *n* redova koji sadrže samu matricu susjednosti. *Graf2* je dat u ulaznom fajlu **graf2.txt** na *Moodle* platformi po istom formatu.

Format izlaznog fajla **rezultat.txt** treba biti **riječ1 [slicnaRijec1:udaljenost1 ... slicnaRijec5:udaljenost5]**. Moguće je da za neke čvorove ne postoji 5 dostižnih čvorova (ispisuje se onoliko sličnih riječi koliko je dostižno) ili ne bude uopšte dostižnih čvorova (u ovom slučaju ispis je **riječ1 []**). Na primjer, ako krećemo od traženja najsličnijih riječi za *graf1* od riječi ucenje, prvi red u fajlu treba biti: ucenje [knjiga:0.10 skola:0.18 student:0.20 udzbenik:0.38 racunar:0.40] itd.



Slika 2. Graf koji predstavlja dio leksičke baze opšteg znanja



Slika 3. Graf koji predstavlja dio leksičke baze za finansijski domen

Napomene:

U okviru domaćeg zadatka je dozvoljeno korišćenje gotovih linearnih struktura podataka, kao što su strukture podataka iz STL biblioteke jezika C++. Ukoliko se koriste gotove implementacije iz drugih izvora, potrebno je jasno naznačiti izvor u komentaru iznad svake takve funkcije.

Studenti su, pored rješenja domaćeg zadatka u vidu izvornog koda i izvršnog fajla, obavezni da predaju izvještaj sa opisom rješenja i logovima na dva primjera izvršavanja (*graf1* i *graf2*). Logovi mogu biti priloženi kao screenshot ekrana nakon izvršavanja programa ili kao tekstualni fajl, obuhvatajući sve faze simulacije. Domaći zadatak se predaje kao jedna arhiva imenovana po principu **Ime-Prezime-BrojIndeksa** na Moodle link za predaju. **U skladu sa već opisanim u propozicijama predmeta, zadatke je potrebno raditi samostalno i zadaci će biti podvrgnuti detekciji sličnosti.**

Rok za predaju domaćeg zadatka: 18. 12. 2022. godine do 16.00h. Raspored odbrane projektnog zadatka biće objavljen nakon završenog roka za predaju.