**Zadatak:** Određivanje semantičke sličnosti tekstualnih dokumenata

**Općeniti opis zadatka – zajedničko svima**

**1. Kreiranje korpusa za učenje vektora riječi**

Za zadanu domenu *downoladati* između 100 i 500 članaka s Wikipedije koji se odnose na zadanu tematiku. Napomena: što više članaka imate, to će vektori riječi biti precizniji.

Za skidanje članaka i pohranjivanje teksta potrebno je napisati odgovarajuću skriptu. Za taj dio možete koristiti gotove Python pakete (npr. Python API wikipedia i sl; linkovi su u dokumentu s dodatnim uputama) koji omogućavaju skidanje članaka za neku zadanu kategoriju. (Možete koristiti bilo koji programski jezik/alat).

**2. Učenje vektora riječi**

Prikupljeni članci pohranjuju se u korpus dokumenata na kojem će se učiti vektori riječi. Za učenje vektora riječi koristi se jedan od odabranih modela (Word2Vec, GloVe ili FastText).

**3. Izračunavanje semantičke sličnosti**

Skup za testiranje mjerenja semantičke sličnosti treba sadržavati ukupno 10 tekstova koje ćete formirati na sljedeći način:

- Odabrati 7 tekstova iz korpusa koji ste pohranili u prvom dijelu zadatka (doc1, doc2, ..., doc7).
- Dodati još 3 dokumenta skroz druge tematike (doc8, doc9, doc10).

Za dokumente iz tog testnog skupa napravite izračune sličnosti, tako da se sličnost računa između svaka dva dokumenta i prizentira u tablici 10x10.

Budući da imate na raspolaganju samo vektore riječi, potrebno je na neki način izračunati vektor za cijeli dokument. Najjednostavniji pristup je da se računa vektor za cijeli dokument (tekst) tako da se uzme aritmetička sredina svih vektora riječi iz tog dokumenta. Potom se sličnost između 2 dokumenta računa kao kosinusna sličnost dva vektora (engl. cosine similarity). Ukoliko dobijete loše rezultate, možete pokušati implementirati i neku vlastitu ideju za određivanje sličnosti između 2 skupa vektora riječi (za dodatne bodove).

Dodatna napomena: Ukoliko neka riječ nije naučena kao vektor (npr. riječ iz preostala 3 dokumenta na kojima nije učeno), samo ju se preskoči.

Ono što treba očekivati od rezultata je veća sličnost između prvih 7 testnih dokumenata, a manju sličnost s preostala tri dokumenta koji nisu tematski povezani.

Kao rezultat eksperimenta potrebno je predati: sve skripte s kôdom, korpus dokumenata, posebno izdvojen skup za testiranje i tablicu s popisom sličnosti dokumenata.

**Popis zadataka s definiranom domenom i modelom koji se testira**

1. Domena: computer science, model Word2Vec
2. Domena: computer science, model GloVe
3. Domena: computer science, model FastText
4. Domena: programming, model Word2Vec
5. Domena: programming, model GloVe
6. Domena: programming, model FastText
7. Domena: artificial intelligence, model Word2Vec
8. Domena: artificial intelligence, model GloVe
9. Domena: artificial intelligence, model FastText
10. Domena: bioinformatics, model Word2Vec
11. Domena: bioinformatics, model GloVe
12. Domena: bioinformatics, model FastText