Zadatak: mjerenje semantičke sličnosti tekstova – dodatne upute

Mjerenje semantičke sličnosti tekstova ima važnu ulogu u različitim problemima iz područja računalne analize prirodnog jezika (Natural Language Processing, NLP) kao što su npr. pretraživanje informacija, klasifikacija dokumenata, razumijevanje prirodnog jezika, razlikovanje značenja riječi, otkrivanje parafraziranja, strojno prevođenje, sumarizacija teksta i dr. Također, nešto općenitiji zadatak, mjerenje sličnosti koncepata od značajne je važnosti i za neka druga područja u kojima je sličnost drugačije definirana, ali se koriste iste metode. Npr. predložene metode mogu se koristiti u području biotehnologije za određivanje sličnosti ontologije gena ili uspoređivanje proteina na temelju njihovih funkcija.

Semantička sličnost tekstova može se mjeriti na različite načine. Jedan od pristupa u mjerenju semantičke sličnosti je primjena metoda dubokog učenja (engl. Deep Learning, DL). Neki od poznatijih DL modela jesu npr. Word2Vec, FastText, GloVe. Ti modeli kao rezultat daju reprezentaciju riječi u obliku vektora, tzv. *word embeddings* na način da se vektori uče iz velikih korpusa tekstova. Tako se postiže da se riječi predstavljaju ovisno o kontekstu u kojem se pojavljuju; odnosno, riječi koje se pojavljuju u sličnom kontekstu, bit će blizu u vektorskom prostoru.

Međutim, kako bi se izračunala semantička sličnost rečenica, paragrafa ili cijelih tekstova, potrebno je uzeti u obzir sve riječi u tom tekstu. U tom slučaju se može računati prosječan vektor (centroid) svih riječi u rečenici/tekstu.

Uobičajeno je da se za mjerenje sličnosti vektora koristi mjera kosinusne sličnosti (engl. cosine similarity).

Korisni Linkovi za projektni zadatak

Linkovi za postupak screpanja sadržaja s Wikipedije:

Wikipedia-API 0.5.4 (https://pypi.org/project/Wikipedia-API/) ili

wikipedia 1.4.0 (https://pypi.org/project/wikipedia/)

Primjeri kôda:

https://wikipedia.readthedocs.io/en/latest/code.html#wikipedia.WikipediaPage.categories

https://wikipedia.readthedocs.io/en/latest/quickstart.html#quickstart

Word2Vec

https://pathmind.com/wiki/word2vec

https://radimrehurek.com/gensim/models/word2vec.html

GloVe

https://nlp.stanford.edu/projects/glove/

https://pypi.org/project/glove/

FastText

https://fasttext.cc/

https://pypi.org/project/fasttext/