NLP Assignment 4

Dieses Assignment wird bewertet !

Die Zusammenarbeit unter Studierenden ist erwünscht, so lange sich diese auf die Diskussion von Konzepten oder Problemen mit python konzentrieren. Das Kopieren von Code ist nicht erlaubt, in diesem Falle werden alle Lösungen der beteiligten Parteien mit 0 Punkten bewertet; dazu werden alle Lösungen (manuell und automatisiert mit Plagiat-Checkern) auf Gruppenarbeit und Kopien untersucht.

In diesem Assignment geht es um den Ähnlichkeitsvergleich von Texten. Die Aufgabe besteht darin, mit verschiedenen Methoden aus den 10'000 Personenbeschreibungen „10k-people-raw.csv" die jeweils **10 ähnlichsten Personen** zu Michael Schumacher, Albert Einstein und Michael Jackson zu finden.

1.) Lesen Sie alle Personenbeschreibungen ein, normalisieren Sie die Texte und visualisieren Sie das Zipf'sche Gesetz für alle Wörter (ohne Satz- und Sonderzeichen).

2.) Bestimmen Sie mit dem Skalarprodukt der *TF-IDF* Vektoren und den Gewichten *Binary*/*Unary* (*Bag of Words* Modell, Folie 9 in Week_7.ppt) die oben erwähnten jeweils 10 ähnlichsten Personen zu Schumacher, Einstein und Jackson.

Diskutieren Sie jeweils die Resultate.

3.) Wiederholen Sie Aufgabe 2.), entfernen Sie aber Stop-Wörter („GermanST.txt").

4.) Wiederholen Sie Aufgabe 2.), verwenden Sie aber für die Dokumenten-Vektoren die *TF-IDF* Gewichte *Counts*/*Unary* (Folie 9, Week_7.ppt):

a) mit entfernten Stop-Wörter,

b) wie 4a), mit Stemming (`nltk.stem.snowball.SnowballStemmer('german')`),

c) wie 4b), aber mit *Cosine Similarity* (statt Skalarprodukt) als Ähnlichkeitsmetrik.

5.) Wiederholen Sie Aufgabe 2.) mit den *TF-IDF* Gewichte *Counts*/*IDF*, Stop-Wörter entfernt, mit Stemming und *Cosine Similarity* als Ähnlichkeitsmetrik.

6.) Wiederholen Sie Aufgabe 5.) mit der LSA (*Latent Semantic Analysis*) Methode. Die *Cosine Similarity* misst diesmal die Ähnlichkeit der Dokumenten-Vektoren im Konzept-Raum:

a) Verwenden Sie die $k = 100$ wichtigsten Hauptkomponenten,

b) die $k = 500$ wichtigsten Hauptkomponenten,

c) die $k = 1000$ wichtigsten Hauptkomponenten.

Abgabe : In Form eines `Jupyter Notebooks`
per Email an dominik.frefel@fhnw.ch bis 10. Juni 2018