Praktikum Fortgeschrittene Methoden des Information Retrieval

Seltene gemeinsame Wörter – semantisch ähnliche Sätze

Gruppe Nr. 8

Mitglieder

Kevin Schramm - Matr. Nr. 3263473
Simon Hüning - Matr. Nr. 3739634

Aufgabeschritte

- Implementierung des feinen Satzähnlichkeitsskripts
    - Erstellung von Satzvektoren - SH
    - Konvertierung der Satzvektoren in Wortvektoren - SH
    - Erstellung in Satzpaaren - KS
    - Sortierung der Satzpaare - KS
    - Zählen gleicher Satzpaare - KS
- Evaluierung des Skriptes - SH & KS
    - Laufzeit
    - Speicherverbrauch
    - Wahl der Parameter
- Optimierung des Skriptes - SH & KS

Zeitplan

- Implementierung des Skriptes bis 02.12.2016
- Fertigstellung der Evaluierung bis 16.12.2016
- Optimierung des Skriptes bis 13.01.2017
- Fertigstellung Dokumentation, Anleitung bis zum 03.02.2017

Mögliche Herausforderungen

- Laufzeit - Es ist davon auszugehen, dass die Laufzeit bei steigender Anzahl von Sätzen stark ansteigt. Es gilt die Laufzeit möglichst gering zu halten.
- Speicherverbrauch - Bei steigender Anzahl von Sätzen und Wörter, die miteinander verglichen werden, ist es wahrscheinlich, dass auch der Arbeitsspeicherverbrauch in die Höhe geht.