NLP: Wichtige und unwichtige Wörter

Christian Wartena

8. April 2021

1 Aufgabe

In dem Ordner *Firmen* liegen einige hundert Texte aus Wikipedia, die eine Firma beschreiben. Wir hätten gerne eine Funktion, die für jeden Text aus dieser Liste die wichtigsten Substantive anzeigen kann. In der vorigen Aufgabe haben wir einfach die häufigste Wörter genommen. Ein besseres Maß ist das tf.idf-Maß. Statt die Frequenz eines Wortes zu nutzen, multiplizieren Sie diese Zahl mit dem IDF-Wert (Inverse Dokumentfrequenz) eines Wortes. Der IDF-Wert eines Wortes w in einer Dokumentsammlung (Korpus) D wird wie folgt berechnet:

$$idf(w) = \log \frac{|D|}{df(w)}$$

wobei $df(w) = |\{d \in D \mid w \in d\}|$ die Anzahl der Dokumente in denen w vorkommt ist.

Vergleichen Sie die am höchsten gewichteten und die am niedrigsten gewichteten Wörter mit den Ergebnissen aus der vorigen Aufgabe.

2 Lernziele

Am Ende dieser Lerneinheit sollen Sie in der Lage sein:

1. Wörter aus einem Text nach dem IDF-Maß zu gewichten.

3 Hinweise

Für den zweiten Teil der Aufgabe ist es sinnvoll erst für alle Wörter zu zählen, in wie vielen Texten sie vorkommen. Hierfür kann man sehr gut wieder einen `Counter` nutzen. Diesen Counter kann man für jeden Text updaten mit einer Liste aller Substantive aus dem Text.