Vorlesung "Computerlinguistische Techniken"

7. Übung (08.01.2016)

Wintersemester 2015/16 – Prof. Dr. Alexander Koller

Sie können alle NLTK-Klassen verwenden, die Ihnen nützlich erscheinen.

1 Distributionelle Semantik

a) Schreiben Sie ein Programm, das die Kookurrenz-Matrix für das Brown-Korpus berechnet. Nehmen Sie zunächst an, dass zwei Wort-Tokens zusammen auftreten, wenn eines in einem Fenster von zehn Tokens um das andere stand. Schreiben Sie eine Funktion zur Berechnung der Kosinus-Ähnlichkeit von zwei Wörtern.

b) Erweitern Sie Ihr Programm aus (a) so, dass Stoppwörter nicht zur Berechnung der Ähnlichkeit herangezogen werden. Stoppwörter sind Wörter aus geschlossenen Wortklassen, die häufig mit anderen Wörtern auftreten. Ein Anfangspunkt könnte die Liste `nltk.corpus.stopwords.words("english")` sein, die Sie aber um weitere Wörter und andere Tokens erweitern können.

c) Probieren Sie beide Versionen Ihres Programmes für einige Wortpaare aus und geben Sie die Ähnlichkeiten der Paare an. Hier sind einige Vorschläge für Wortpaare, die Sie nach Belieben erweitern sollten: state/country; fire/water; election/vote; good/evil; good/bad. Diskutieren Sie Ihre Ergebnisse.

Abgabe bis 19.01.2016, vor der Vorlesung, per Mail an `mgerdes@uni-potsdam.de`