Praktischer Teil 14

Keyphrase Extraction

In diesem Praktikum extrahieren Sie automatisch Keyphrases aus einem Set von Dokumenten.

Das Programm evaluate_extractor.py extrahiert mit dem RAKE-Algorithmus Keyphrases aus allen Dokumenten im Verzeichnis corpus (.txt-Dateien). Anhand der manuell erfassten Listen (.key-Dateien) von Keyphrases wird der F-Score bestimmt.

Hinweis: Vor der ersten Ausführung von evaluate_extractor.py muss evtl. noch ein Sprachmodell nltk heruntergeladen werden. Dazu öffnen Sie eine python-Console und geben „import nltk“ und „nltk.download()“ ein. Im Fenster, das sich dann öffnet, müssen Sie im Tab „Models“ das Package „punkt“ auswählen und herunterladen.

Aufgabe 1:

Um ein Gefühl für die Daten zu bekommen, erfassen Sie von Hand aus 5 Dokumenten die für Sie relevant erscheinenden Keyphrases. Das Resultat wird im Praktikum in einem gemeinsamen Google Dokument erfasst.

Aufgabe 2:

a) Führen Sie das Programm evaluate_extractor.py aus. Der finale Score sollte 0.3% sein.

b) Wie verhält sich der Score, wenn Sie statt N=10 z.B. N=5 oder N=15 verwenden?

Aufgabe 3:

a) Das Programm verwendet die Stopwort-Liste in der Datei stopwords_simple.txt. Verwenden Sie stattdessen die Datei stopwords_nltk.txt. Wie verhält sich der Score? Was passiert, wenn Sie nur die ersten 500 Elemente der neuen Stopwort-Liste verwenden?

b) Versuchen Sie anhand Ihrer Beobachtungen eine bessere Stopwort-Liste für den Rake-Algorithmus zu finden oder zu generieren. Binden Sie diese ein. Welchen Score können Sie erreichen?

Aufgabe 4:

Visualisieren Sie die extrahierten Keyphrases.

Zusatzaufgabe: Erweitern und modifizieren Sie den RAKE-Algorithmus, um die extrahierten Keyphrases zu verbessern!