Dr. Maximilian Hadersbeck
Leonie Weißweiler

Ludwig-Maximilians-Universität
Centrum für Informations- und Sprachverarbeitung

Abgabe: 2.2.2017

13. Übung zur Vorlesung

Einführung in die Programmierung für Computerlinguisten

Bitte achten sie bei ihren Lösungen darauf, dass die Groß- und Kleinschreibung der Nutzereingaben keine Rolle spielen darf - die Eingabe "Spam and Eggs" soll also das gleiche Ergebnis liefern wie die Eingabe "spam and eggs". Ausnahmen sind im Angabentext eindeutig gekennzeichnet.

Es geht um die Erkennung der Sprache, die in einer Datei vorliegt. Die Computerlinguistische Theorie sagt, dass die Häufigkeit der Buchstabenfolgen der Länge 2 und 3 in einem Text für jede Sprache spezifisch sind. Testen Sie das bei deutschen und englischen Texten. z.B.

Text: “heute ist montag”

Buchstabenfolgen der Länge 2 sind: ‘he’ ‘eu’ ‘ut’ ‘te’ ‘e’ usw.

Buchstabenfolgen de Länge 3 sind: ‘heu’ ‘eut’ ‘ute’ ‘te’ ‘e i’ ‘is’ usw.

Verwenden sie für die folgenden Aufgaben zwei Auszüge aus dem Europarl Corpus, die sie auf der Übungswebsite herrunterladen können. Diese beinhalten den identischen Text einmal in deutscher, und einmal in englischer Sprache.

Aufgabe 13-1

Schreiben Sie eine Funktion, die eine Frequenzliste aller Buchstabenfolgen der Länge 2 entwickelt.

Tipp: verwenden sie Slicing, weil das Modul re keine überlappenden Matches finden kann.

Aufgabe 13-2

Schreiben Sie eine Funktion, die eine Frequenzliste aller Buchstabenfolgen der Länge 3 entwickelt.

Aufgabe 13-3

Schreiben Sie eine Funktion, die ein dictionary übergeben bekommt und die die 10 häufigsten Werte des dictionary ausdruckt.

Aufgabe 13-4

Schreiben Sie ein Hauptprogramm, das die beiden Dateien einliest und jeweils für beide Sprachen die häufigsten 10 Buchstabenfolgen der Länge 2 und der Länge 3 ausgibt. Betrachten Sie das Ergebnis.