Dr. Maximilian Hadersbeck
Centrum für Informations- und Sprachverarbeitung
Ludwig-Maximilians-Universität

Abgabe: siehe Übungsseite

11. 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.

Aufgabe 11-1

Holen Sie die ersten 4 Bücher der Bibel von Projekt Gutenberg (http://gutenberg.spiegel.de/buch/5560/i für [i=1,..,4]) mit dem UNIX Befehl wget

Aufgabe 11-2

Verwenden sie lynx-dump um die Bücher, die in den .html Dateien gespeichert sind in eine Textdatei zu konvertieren.

Aufgabe 11-3

Fügen Sie alle Bücher zu einer Datei bibel.txt zusammen.

Aufgabe 11-4

Schreiben Sie eine Funktion, die eine Zeile bekommt und das längste Wort der Zeile zurückgibt.

Aufgabe 11-5

Schreiben Sie eine Funktion, die eine Zeile bekommt und die Anzahl der Wörter zurückgibt.

Aufgabe 11-6

Erzeugen Sie eine Frequenzliste aller Wörter aus der Datei bibel.txt.

a) Wieviele unterschiedliche Wörter kommen in der Datei vor?

b) Was sind die 10 häufigsten großgeschriebenen Wörter?

c) Was sind die 10 häufigsten kleingeschriebenen Wörter?

Aufgabe 11-7

Schreiben Sie ein Programm, das die ersten 3 Verse eines jeden Kapitels von jedem Buch ausgibt. Die Ausgabe soll folgende Form haben:

1. Buch Mose:

Kapitel 1

1. Am Anfang schuf Gott Himmel und Erde.

2. Und die Erde war wüst und leer, und es war finster auf der Tiefe; und der Geist Gottes schwebte auf dem Wasser.

3. Und Gott sprach: Es werde Licht! und es ward Licht.

Aufgabe 11-8

Finden sie heraus welcher Vater die meisten Söhne hat, indem sie mit einer geeigneten Regex die Vorkommen von "[Sohn], der Sohn [Vater]" finden und analysieren.

Aufgabe 11-9

Schreiben Sie eine Funktion, die eine Zeile bekommt und das kürzeste Wort der Zeile zurückgibt.