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

Abgabe: siehe Übungsseite

8. 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.

Benutzen Sie für die folgenden Aufgaben: Verbrechen und Strafe von F. Dostojewski [Übersetzung von 1924, von zeon.org] ,(http://www.cis.uni-muenchen.de/kurse/max/einfprog/texte/Verbrechen_Strafe_1924.txt)

Aufgabe 8-1

Erzeugen Sie die Frequenzliste wie in Übung 7 und geben Sie das häufigste Wort aus und die Anzahl wie oft das Wort im Text vorkommt.

Aufgabe 8-2

Erzeugen Sie die Frequenzliste wie in Übung 7 und geben Sie ein Wort aus, das nur ein einziges Mal im Text vorkommt. Ihr Programm kann auch bei verschieden Durchläufen, unterschiedliche Wörter mit der Häufigkeit 1 ausgeben.

Aufgabe 8-3

Verwenden Sie für die nächsten Aufgaben eine Textdatei, die sie aus einer .html Datei extrahieren. Dazu folgende Hinweise:

a) Mit dem UNIX Befehl wget : **`wget "http://de.wikipedia.org/wiki/Ludwig_Wittgenstein" -O wittgenstein.html`** speichern Sie von der Kommandozeile aus automatisch den Inhalt einer www-Seite in einer lokalen Datei. Mit dem obigen Befehl speichern Sie z.B. den Inhalt der Wikipedia Seite über Ludwig Wittgenstein in der Datei wittgenstein.html.

b) Mit dem UNIX Befehl lynx können Sie den Text aus der html Datei extrahieren und in einer Textdatei speichern.

```
lynx -dump -assume_charset=UTF-8 -hiddenlinks=ignore -nolist -verbose wittgenstein.html
> wittgenstein.txt
```

Aufgabe 8-4

Spalten Sie alle Zeilen der Datei wittgenstein.txt in Wörter auf und erzeugen Sie eine Frequenzliste der Wörter. Speichern Sie alle Wörter, die mindestens zehn Mail in der Datei wittgenstein.txt vorkommen in der Liste woerter_10. Geben Sie am Ende die Liste aus.

Aufgabe 8-5

Erweitern Sie das Programm, sodass es alle großgeschriebenen Wörter aus der Frequenzliste findet und die Wörter zusammen mit Ihrer Häufigkeit in der Datei "grosse_woerter.txt" speichert.

Aufgabe 8-6

Wie lauten die regulären Ausdrücke, die folgende Wörter in der Frequenzliste finden (geben Sie nur den regulären Ausdruck an!). Testen Sie Ihre Ausdrücke mit einem Programm. a) alle Wörter mit genau 3 Buchstaben

b) alle Wörter in denen ein Vokal vorkommt.

c) alle Wörter mit Umlauten

d) alle Wörter die mit 'net' enden.

e) alle Wörter die drei bis fünf Buchstaben lang sind und mit einem Vokal beginnen.

f) alle Wörter die mit einem Großbuchstaben beginnen und mit dem Suffix 'ung' enden.

g) alle Wörter die nur aus Großbuchstaben bestehen und maximal 8 Buchstaben lang sind.

h) alle Wörter in denen kein Vokal und kein Umlaut vorkommt.

Aufgabe 8-7

Schreiben Sie ein Programm, das die Anzahl aller kleingeschriebenen Wörter in der Datei wittgenstein.txt ermittelt. Geben Sie die Anzahl aus.

Aufgabe 8-8

Geben Sie alle Zeilen der Datei aus, die mit einem Großbuchstaben beginnen.

Aufgabe 8-9

Geben Sie alle Zeilen der Datei aus, in denen "der", "die" oder "das" vorkommt.

Aufgabe 8-10

Geben Sie alle Wörter der Datei aus, die von einem Satzzeichen gefolgt sind.

Aufgabe 8-11

Schreiben Sie ein Programm, das ein Wort einliest und nachsieht, ob das von rückwärts gelesene Wort als Teil eines Wortes in einer großen Datei vorkommt.

http://www.cis.uni-muenchen.de/kurse/max/einfprog/texte/sz_1Mio.txt.gz

Beispiel: "LOVE" =umgedreht=> "EVOL" =suche.... in der Datei kommt vor in=> Revolution =siehe>= "R.EVOL.ution"