Friedrich-Schiller-Universität Jena
Institut für Informatik
Martin Mundhenk
https://www.complexity.uni-jena.de/Python1

Aufgabenblatt 10, Abgabetermin 20.7.2020

Lösen Sie Aufgaben im Umfang von 15 Punkten.
Sie können Module aus der Vorlesung benutzen.

Aufgabe 54: Buchstabenhäufigkeiten **10 Punkte**

Schreiben Sie ein Programm, das eine Textdatei einliest und die Häufigkeit aller Buchstaben ausgibt. Klein- und Großbuchstaben werden nicht unterschieden. Beim Text `abbb abABb a1 () abbabb.\n abb.` ist das Ergebnis

```
a 7
b 12
```

Aufgabe 55: Auswertung der n-Gramm-Analyse **5 Punkte**

Analysieren Sie das Ergebnis der n-Gramm-Analyse. Schreiben Sie ein Programm, das

1. die Anzahl der n-Gramme und
2. die am häufigsten vorkommenden n-Gramme

bestimmt.

Aufgabe 56: Wort-Entropie **15 Punkte**

Die Entropie der Wörter eines Textes ist ein Maß für seinen Informationsgehalt. In einem Text aus n Wörtern, in dem k verschiedene Wörter $w_1, \ldots, w_k$ vorkommen, kann man diese Wort-Entropie durch folgende Formel ausdrücken:

$$p_1 \cdot \log \frac{1}{p_1} + p_2 \cdot \log \frac{1}{p_2} + \ldots + p_k \cdot \log \frac{1}{p_k}$$

Dabei ist p_i die relative Häufigkeit von Wort w_i im Text, also die Anzahl der Vorkommen von Wort i geteilt durch die Anzahl aller Vorkommen von Wörtern im Text. Zum Beispiel besteht der Text `ab ab ab abb aba. abb.` aus $n = 6$ Wörtern und $k = 3$ verschiedenen Wörtern $w_1 =$ `ab`, $w_2 =$ `abb` und $w_3 =$ `aba`. p_1 ist $\frac{3}{6}$, p_2 ist $\frac{2}{6}$ und p_3 ist $\frac{1}{6}$. Wenn man den Logarithmus zur Basis 2 nimmt, ist die Wort-Entropie ein Maß für den Informationsgehalt eines Wortes in Bit. Im o.g. Text ist sie 2, in meinem Python-Programm zum Berechnen der Wort-Entropie ist die Wort-Entropie 5.2, bei diesem Übungsblatt 6.8, in meiner „Schlagertextdatei“ und meiner „Fußballreportagendatei“ ist sie 8.3 beim *Faust* liegt sie bei 9.9 und *Ulysses* erreicht 10.5.
Schreiben Sie ein Programm, das die Wort-Entropie einer Textdatei berechnet.

Aufgabe 57: Konkordanz **15 Punkte**

Eine Konkordanz gibt an, wo ein Wort in einem Text vorkommt. Schreiben Sie eine Funktion, die eine Konkordanz für eine Textdatei erstellt. Die Konkordanz soll ein Dictionary mit den Wörtern als Schlüsseln sein. Die Werte sollen aus den Zeilennummern bestehen, in denen das Wort in der Textdatei vorkommt (die genaue Stelle in der Zeile lassen wir weg). Schreiben Sie eine Testfunktion, die wiederholt ein Wort einliest und es in der Konkordanz nachschlägt.