Programmierkurs Python I – WS 2010/11

Übung 5

---

1 Wörter zählen (3 + 1 Punkte)

Auf der Homepage (unter „Beispiellösungen“) findest du die Datei `A00-pos-lines`, ein Ausschnitt aus dem British National Corpus. In der Datei sind alle Wörter mit Kategorieinformationen (POS-Tags) versehen, in der Form `wort$tag`. Pro Zeile gibt es genau ein Wort-POS-Paar. „Wörter“ können Leerzeichen enthalten (Mehrwortausdrücke). POS-Tags enthalten keine Leerzeichen. Satzzeichen zählen als ein Wort (und sind getagt).

Schreibe ein Programm, das die Datei einliest (entweder von der Homepage oder von Eurem Dateisystem), und zählt, welches Wort wie oft vorkommt. Als Ausgabe solltet Ihr eine Datei schreiben, das für jedes Wort seine Häufigkeit auflistet und wie oft es mit welchem POS-tag vorkommt, etwa so (hypothetische Zahlen):

```
name   17      NN1:12  VVT:5
```

Bonusaufgabe: Auf der Homepage findest Du außerdem die Datei `A00-pos`, die im Original-BNC-Format gedruckt ist. Sie unterscheidet sich von der anderen Datei nur darin, dass mehrere Wort-Pos-Paare auf einer Zeile stehen können. Erweitere Dein Programm so, dass es auch diese Datei wie oben erklärt verarbeiten kann.

2 Der Goldkäfer (6 Punkte)

„Der Goldkäfer“ ist eine Geschichte von E. A. Poe. Es geht unter anderem darum, wie jemand einen verschlüsselten Text anhand von Buchstabenhäufigkeiten entschlüsselt.

Die Annahme ist, dass im verschlüsselten Text jedes Zeichen einfach durch ein anderes ersetzt wurde. Die Taktik zum entschlüsseln besteht nun darin, einen hinreichend umfangreichen Beispieltext (Referenztext) zu betrachten und zu ermitteln, welche Zeichen wie häufig vorkommen. Dann bestimmt man die entsprechenden Buchstabenhäufigkeiten im verschlüsselten Text, und übersetzt das häufigste Zeichen im Text mit dem häufigsten Zeichen im Referenztext, das zweithäufigste Zeichen mit dem zweithäufigsten aus dem Referenztext, usw.

Implementiere einen Algorithmus, der diese Art der Dechiffrierung benutzt.

Zum Entschlüsseln haben wir einen Text verschlüsselt, (brown-sample-enc.txt), den ihr von der Homepage laden könnt. Satzzeichen sind nicht verändert worden. Um die Buchstabenhäufigkeiten zu bestimmen, könnt ihr den Text brown.txt als Referenz benutzen.

3 with (2 Punkte)

Ersetze folgendes Code-Fragment durch äquivalenten Code ohne das `with`-Statement:

```
with open('datei.txt') as f:
  for line in f:
    sys.stdout.write(line)
```

Abgabe bis Donnerstag, 2.12.09, 14:00 per Mail an
`regneri@coli.uni-sb.de`
`stth@coli.uni-sb.de`