NLP: Von Texten zu Wörtern

1 Aufgabe

In dem Ordner *Firmen* liegen einige hundert Texte aus Wikipedia, die eine Firma beschrei-ben. Wir hätten gerne eine Funktion, die für jeden Text aus dieser Liste die wichtigsten Substantive anzeigen kann.

Als Maß für die Wichtigkeit können Sie zunächst die Worthäufigkeit nehmen.

2 Lernziele

Am Ende dieser Lerneinheit sollen Sie in der Lage sein:

1. Ein Programm zu schreiben, das Texte in Sätze und Wörter aufteilt
2. Ein Programm zu schreiben, das Lemma und Wortart für Wörter in einem Text bestimmt.

Sie kennen außerdem die Konzepte: Lemma, Stamm, Wortart, Tokenisierung, Lemmatisierung und POS-Tagging.

3 Hinweise

Die Wikipedia-Texte lesen Sie am besten zeilenweise ein. Da Überschriften im Wiki-Text mit Gleichheitszeichen gekennzeichnet werden, ist es außerdem sinnvoll, diese Zeichen sowie Leerzeichen am Anfang und Ende der Zeile zu entfernen. Hierfür kann man die Funktion `strip` verwendet werden. Insgesamt haben wir dann:

```
datei = codecs.open(f,'r','utf8')
    for zeile in datei:
        zeile = zeile.strip('= \n')
```

Wenn Sie die Texte so einlesen, ist es sinnvoll, die Texte sofort alle zu tokenisieren und als Liste von Sätzen zu speichern, wobei ein Satz eine Liste von Wörtern ist.