# Aufgabe 1

### Crawling und Tokenisierung

Laden Sie mit dem Unixprogramm *wget* als Crawler Seiten von der BBC-Webseite `www.bbc.com` herunter.

Schreiben Sie dann ein Programm, welches aus den heruntergeladenen Seiten die reinen Texte der Zeitungs-Artikel ohne Werbung, Navigationselemente etc. extrahiert.

Schreiben Sie außerdem ein Tokenisierer-Programm, welches den gesamten extrahierten Text in Tokens (Wörter, Satzzeichen, Klammern etc.) zerlegt, und dann jeden Satz mit Leerzeichen zwischen den Tokens in einer separater Zeile ausgibt.

Der Tokenisierer soll Abkürzungen korrekt behandeln. Eine Liste von englischen Abkürzungen finden Sie hier: `http://www.cis.uni-muenchen.de/~schmid/lehre/Experimente/data/abbreviations`

### Vorüberlegungen

- Welche Schritte sind bei der Tokenisierung sinnvoll?

Sie dürfen Python-Bibliotheken, die nicht zum Standard gehören, (in dieser und anderen Übungen) nur nach vorheriger Absprache verwenden.

Schicken Sie die beiden Programme an `schmid@cis.lmu.de`.

Überlegen Sie sich außerdem schon einmal, welche Sprache Sie in der nächsten Übungsaufgabe behandeln wollen, und welche morphologischen Phänomene Sie behandeln wollen.