Übung 5: Formate

1. Basics

   a) Kopieren Sie sich die Datei `/home/public/vorkurs_ss19/starwars.xml` in Ihr Vorkurs-Verzeichnis.

   b) Öffnen Sie die Datei mit einem Text-Editor und fügen Sie die Daten von Prinzessin Leia hinzu. Leias Heimatplanet ist Alderaan. Speichern Sie die Datei.

   c) Lesen Sie die XML-Datei nun mit einem DOM-Parser ein. Schreiben Sie nun ein kurzes Python-Skript, das Name und Heimatplanet der Personen ausgibt. Die Ausgabe sollte so aussehen:

   ```
   Luke Skywalker: Tatooine
   Han Solo: Corellia
   Prinzessin Leia: Alderaan
   ```

   d) Ändern Sie das Skript noch einmal, so dass nur Jedis ausgegeben werden, also nur Personen bei denen das tag `<jedi />` definiert ist.

   e) Erstellen Sie ein Dokument `starwars.json` und versuchen Sie, die Daten in `starwars.xml` in JSON-Format zu übertragen. Was ist problematisch?

   f) Verwenden Sie das `json`-Modul im Python-Interpreter, um `starwars.json` einzulesen. Falls die Datei nicht eingelesen werden kann, korrigieren Sie sie.

Zusatzaufgabe Probieren Sie, Aufgaben 1c und 1d mit einem SAX-Parser zu lösen. Was ist einfacher/schwieriger?

2. BNC-Preprocessing

   a) Der British National Corpus ist ein großes Korpus. Wir arbeiten, um die Rechner nicht unnötig zu beanspruchen, daher mit jeweils nur einer Datei. Schauen Sie auf die Uhr und merken Sie sich die letzte Stelle der aktuellen Minutenzahl. Wenn das eine 9 ist, nehmen Sie stattdessen den Buchstaben P. Nun kopieren Sie sich aus `/resources/corpora/monolingual/annotated/bnc/Texts/A/A0/` die Datei, die sich ergibt, wenn sie an A0 die gemerkte Zahl oder den Buchstaben anhängen (sie sollten dann einen Dateinamen wie z.B. A05.xml haben) in Ihr Vorkurs-Verzeichnis.

   b) Wir wollen nun aus dieser Datei den eigentlichen Text des Korpus extrahieren. Erweitern Sie dazu Ihr Python-Skript (Sie brauchen sich eigentlich nur an dem Element `w` zu orientieren, das die Wörter repräsentiert). Entscheiden Sie sich für einen DOM- oder SAX-Parser.

   c) Geben Sie nun statt den Wörtern die Lemmata aus. Die Lemmata stecken im Attribut „hw".