Übung zur Vorlesung "Computerlinguistik II / Sprachtechnologie"

Sommersemester 2018, Prof. Dr. Udo Hahn, Tobias Kolditz
Übungsblatt 6 vom 11.06.2018
Abgabe bis 20.06.2018 per E-Mail (PDF- & Python-Datei) an
tobias.kolditz@uni-jena.de

Aufgabe 1 : NLP-Ressourcen **6 pt**

Wie in der Vorlesung gezeigt, besteht das PennTreebank-Corpus aus Klartextdateien. Sätze sind in einer Klammerschreibweise angegeben, wie sie beispielsweise in den folgenden zwei Sätzen aus dem Corpus demonstriert wird:

```
( (S (S (NP-SBJ An enormous turtle)
         (VP has
                 (VP succeeded
                 (SBAR-LOC (WHADVP-1 where)
                           (S (NP-SBJ the government)
                              (VP has
                                  (VP failed
                                      (ADVP-LOC *T*-1)))))))))
     :
     (S (NP-SBJ He)
        (VP has
            (VP made
                (S (S-NOM-SBJ (NP-SBJ *)
                              (VP speaking
                                  (NP Filipino)))
                   (ADJP-PRD respectable)))))
     .))
```

Geben Sie für den Satz

> The turtle Pong Pagong is a character who stars in children's television show "Batibot".

analog zu den obigen Beispielsätzen einen Parse in Klammer-Form an. Verwenden sie für die Phrasenbezeichnungen einfache Kategorien, wie sie zuvor in Vorlesung und Übung verwendet wurden, z.B. S, NP, VP. Falls Sie es für nötig erachten, können Sie auch weitere Kategorien verwenden (s. http://languagelog.ldc.upenn.edu/myl/PennTreebank1995.pdf).

Aufgabe 2 : Algorithmen & Python **4 pt**

Schreiben Sie ein Programm, das eine Klartext-Datei mit einem PennTreebank-Parse einliest (s. Beispiel aus Aufgabe 1) und die Klammerung validiert (es muss die gleiche Anzahl öffnender und schließender Klammern vorkommen). Testen Sie Ihr Programm!

Hinweis: Werfen Sie einen Blick auf Abschnitt 7.2 des Python-Tutorials zum Lesen (und Schreiben) von Dateien:

https://docs.python.org/3/tutorial/inputoutput.html#reading-and-writing-files