

Übung 8: Tree Tagger

1. In dieser Aufgabe wollen wir einen Text mit dem TreeTagger lemmatisieren und die Nomen-Lemmata extrahieren.
 - a) Wechseln Sie auf den Rechner ella, gehen Sie in Ihr **Vorkurs-Verzeichnis** und kopieren Sie sich die Datei `/home/public/vorkurs_ss19/Darth_Vader.txt`. Es handelt sich dabei um den ersten Absatz des Wikipedia-Artikels über Darth Vader.
 - b) Führen Sie mit **source** das setup-Skript für den TreeTagger aus.
 - c) Benutzen Sie eine Pipe, um den Text durch den Tree-Tagger zu schicken. Da es ein englischer Text ist, sollten Sie **tree-tagger-english** verwenden.
 - d) Filtern Sie aus der Ausgabe des TreeTaggers nun einige Wörter heraus, so dass nur noch Nomen (POS-tag: NN oder NNS) übrigbleiben.
 - e) Wie Sie nun sehen können, wird für einige Nomen kein Lemma gefunden. TreeTagger gibt an der Stelle ein **<unknown>** aus. Werfen Sie nun aus der Ausgabe alle Zeilen heraus, in denen ein unknown vorkommt.
 - f) Entfernen Sie bitte die ersten beiden Spalten, so dass nur noch das Lemma auf jeder Zeile steht.

2. In der nächsten Aufgabe sollen Sie das tree-tagger-Programm direkt benutzen. Es heißt **tree-tagger**. Die wichtigste Option, die das Programm bekommt, ist das statistische Modell. Es steckt in einer sog. par-Datei (für englisch: `/resources/processors/tagger/tree-tagger/lib/english.par`).
 - a) Rufen Sie tree-tagger zunächst nur mit der par-Datei auf und füttern Sie ihn über eine Pipe mit einzelnen Wörtern Ihrer Wahl. Das tree-tagger-Programm kann nur einzelne Wörter verarbeiten.
 - b) Spielen Sie mit den Optionen herum, die tree-tagger anbietet.
 - c) Da Sie ja mittlerweile wissen, wie man Texte so umformatiert, dass jedes Wort auf einer einzelnen Zeile steht – tun Sie das mit obigem Text und füttern Sie ihn direkt in das tree-tagger-Programm (*nicht* in das shell-Skript **tree-tagger-english** sondern das Programm **tree-tagger**).
 - d) Erstellen Sie eine Lexikon-Datei, in der Sie die unbekanntenen Wörter manuell taggen (es gibt eine Beispiel-Lexikon-Datei: `lib/english-lexicon.txt`). Einige unbekanntene Wörter: Episode, Anakin, prequel, ...
 - e) Taggen Sie nun noch einmal **Darth_Vader.txt**, aber übergeben Sie dem tree-tagger Ihr manuell erstelltes Lexikon. Sie sollten jetzt keine **<unkown>** mehr sehen.