Übung zur Vorlesung "Computerlinguistik II / Sprachtechnologie"

Sommersemester 2018, Prof. Dr. Udo Hahn, Tobias Kolditz
Übungsblatt 5 vom 05.06.2018
Abgabe bis 13.06.2018 per E-Mail (PDF-Datei) an
tobias.kolditz@uni-jena.de

Aufgabe 1 : Chunking – Regex **3 pt**

Schreiben Sie einen regulären Ausdruck, der auf Grundlage der von Ihnen zugewiesenen POS-Tags (jeweils durch Leerzeichen getrennt) Chunks im Text aus Übungsblatt 1 erkennt. Nehmen Sie dabei folgende NP-Chunks als gegeben an:

- die letzten Sekunden (Tags: „ART ADJA N“, siehe Blatt 1, Aufgabe 2)
- des Bundespräsidenten Horst Köhler
- widersprüchliche Gefühle
- er
- Hand
- seiner Frau
- dem Ausgang
- des Schlosses Bellevue
- etwas Anrührendes
- dem Mann

Aufgabe 2 : Korpusrecherche **3 pt**

a) Digitales Wörterbuch der deutschen Sprache (DWDS) 1 pt

„Babo“ wurde 2013 von einer Jury des Langenscheidt-Verlags zum Jugendwort des Jahres gewählt. Suchen Sie über eine Korpusabfrage des DWDS (`https://www.dwds.de/r`) den ersten Beleg für dieses Wort im Unterkorpus *Die ZEIT (1946–2018)*. Wann und in welchem Kontext tauchte das Wort in der Zeit erstmals auf?

b) Deutsches Textarchiv (DTA) 2 pt

Überfliegen Sie die Dokumentation zur Abfragesyntax des Deutschen Textarchivs:

`http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe`

Schreiben Sie eine Suchanfrage, die alle Belege für „Saulus“ und „Paulus“ in einem Satz findet. Wie viele Treffer liefert die Suche im Korpus (`http://www.deutschestextarchiv.de/`)?

Aufgabe 3 : Annotationsformate **4pt**

Gegeben sei der folgende Ausschnitt aus einem XML Dokument:

```
<unit id="d12.u21">
    <text>
        Bei Li-Fraumeni-Krebs, bei dem das p53-Gen defekt ist, arbeitet das
        p53-Protein nicht richtig, und die Krebszellen können weiter wachsen.
    </text>
```

```
    <entity id="d12.u21.e1" grp="DISO" offset="16" len="5">Krebs</entity>
    <entity id="d12.u21.e2" grp="CHEM" offset="35" len="3">p53</entity>
    <entity id="d12.u21.e3" grp="CHEM" offset="68" len="11">p53-Protein</entity>
</unit>
```

1. Welche Art Annotationen sind vorhanden?
2. Handelt es sich hier um ein in-line oder ein stand-off Format?
3. Wandeln Sie den Text in den anderen Formatstyp um.
4. Welche Probleme könnten bei dieser Art Umwandlung entstehen?