Projektplan: Text Mining - Sentiment Analyse

Fragestellung: Ist es möglich, mithilfe der Sentiment Analyse festzustellen, ob Hauptcharaktere in Büchern gut oder böse sind?

Datenquelle: Harry Potter Romane in Englisch

Tools: Python, Nltk, Calibre, Javascript, D3, SentiStrength

Vorbereitung:

- Download der Bücher und Bereinigen der Texte mit Calibre (Umwandlung in txt-Files, Vorwort/Danksagung etc. entfernen)
- Download SentiStrength
- Lernen, wie man mit SentiStrength arbeitet
- Download D3
- Lernen, wie man mit D3 arbeitet

Programm:

• Text aufbereiten: erst Sätze, dann einzelne Wörter trennen, PoS-Tagging, Named Entity Recognition = Extrahieren der wichtigen Charaktere (Quelle: http://ilias.uni-duesseldorf.de/ilias_0500xy/ilias/goto.php?target=file_446596_download&client_id=UniRZ)

• Häufigkeit der Charaktere zählen und nur die wichtigsten aufnehmen (also eine untere Grenze einführen)

• Sentiment Analyse der Wörter in deren Umgebung: SentiStrength nimmt einen txt-File entgegen, wobei jede Zeile einer Sentiment Analyse unterzogen wird und am Ende der Zeile das Ergebnis geschrieben wird. Man kann auch Keywords angeben, in deren Umgebung die Analyse gemacht wird, was ich mit den Charakternamen testen werde.

• Jedem Charakter einen Ranking-Wert zuordnen (negativ – „böse“, positiv – „gut“)

• Umwandlung in Datenaustauschformat (JSon) als Schnittstelle für JS

• Visualisierung der Ergebnisse mit D3 (wie das genau geht, muss ich mir noch aneignen)

Visualisierung:

- Charaktere als Kugeln – je größer, desto wichtiger der Charakter
- Auf der linken Seite stehen die Charaktere, welche ein negatives Ranking haben, in der Mitte neutrale und rechts diejenigen, die ein positives Ranking haben
- Hintergrund in den Farben rot und grün (böse – gut)
- optional: Der Position auf der y-Achse eine Bedeutung geben
- optional: Beziehungen zwischen Charakteren mit Linien verdeutlichen