Dokumentation
QUALM Modifikatoren

Im Ordner ‚Rechtschreibfehler_Kleinschreibweise_Abkuerzungen/code.und.messungen' befindet sich ein Python Programm für die Anwendung von QUALM-Indikatoren und Modifikatoren auf Korpora. Dieses enthält u. A. Methoden um Rechtschreibfehler zu verbessern und Texte in Kleinschreibweise zu transformieren.
Im Ordner scripts befinden sich außerdem zwei Skripte zur Auflösung von Abkürzungen in Textdaten: abbreviation_resolutions_and_saving_to_pickle und abbreviation_resolutions_csv. Ersteres liest Daten im conll Format (http://www.conll.org/) und gibt die Daten mit aufgelösten Abkürzungen als pickle aus (siehe http://www.nltk.org/howto/data.html). Letzteres arbeitet sowohl in der Eingabe als auch in der Ausgabe mit Datensätzen im csv-Format. Im Ordner corpora/abbreviation_lists befinden sich außerdem einige beispielhafte Abkürzungsressourcen. Hinweis: Da für Windows x64 keine PyEnchant
Version vorliegt, wurde dieses auskommentiert, kann aber je nach Umgebung wieder auskommentiert werden.
Im Ordner ‚code.und.messungen' befindet sich zusätzlich eine Version des Programms, welches CSV Dateien einlesen und analysieren kann. Hierzu muss die Eingabedatei auf ‚.csv' enden und es können weitere Parameter übergeben werden (siehe hierzu den Programmcode).
Die Korpora müssen zuerst heruntergeladen werden und im Ordner corpora abgelegt werden: siehe https://www.nltk.org/data.html
und [Gi11] Gimpel, Kevin; Schneider, Nathan; O'Connor, Brendan; Das, Dipanjan; Mills, Daniel; Eisenstein, Jacob; Heilman, Michael; Yogatama, Dani; Flanigan, Jeffrey; Smith, Noah A.: Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 42–47, 2011.

Im Ordner , Trainingsdatenselektion/code.und.messungen' befindet sich ein Java Programm für die Anwendung des QUALM- Modifikators ‚Wahl passender Trainingsdaten'. Für die Ausführung siehe das Readme.