Text-Mining Praktikumsbericht

Sebastian Gottwald, Simon Bordewisch

4. Februar 2016

1 Einleitung

Im Rahmen des Moduls Text-Mining hatten wir die Aufgabe einen Klassifikator für den Stanford Named Entity Recognizer (Stanford NER) zu erstellen. Als Datengrundlage standen uns ein, von Studenten erstelltes, Programm zur Verfügung, welches eine kategorisierte Liste von Titel der deutschen Wikipedia erzeugt. Unsere Aufgabe war es, diese Daten aufzubereiten und anhand von Wikipediaartikeln aus einem XML-Dump, Trainingsdaten für den Stanford NER Klassifikator zu erstellen, welcher Personen-, Organisations- und Ortsnamen erkennt

Der Stanford NER ist eine Java-Implementation eines Named Entity Recognizers (NER). Ein NER markiert Wort-Sequenzenn in einem Text welche bestimmte Kategorien repräsentieren (z.B. Personen, Orte, Organisationen oder auch Gene und Proteine).

2 Methodik und Vorgehen

2.1 Aufarbeitung der Vergleichsdaten

Wir hatten insgesamt vier verschiedene Programme des Vorjahrespraktikas zur Verfügung. Wir entschieden uns das Programm zu nutzen, welches auf den Titeln der Wikipediaartikel arbeitet. Damit bekamen wir eine umfangreiche und nahezu fehlerfreie Liste von Organisationen sowie Personen- und Ortsnamen. Wir mussten zunächst diese Liste aufarbeiten, sodass sie ein für uns nutzbares Format besitzt.

2.2 Extraktion der Wikipedia Artikel

Als Datengrundlage zur erstellung der Trainingsdaten des Klassifikators, diente uns der aktuelle Wikipedia Dump "dewiki-latest-pages-articles.xml.bz2". Zum extrahieren der Daten nutzten wir StAX-API da diese sich für Große Datenmengen eignet. Beim parsen de Wikipediadumps haben wir überprüft um welchen Normdatentyp es sich bei dem jeweiligen Artikel handelte, da Personenartikel die Bezeichnung "Typ=p", Ortsartikel die Bezeichung "Typ=g" und Körperschaftsartikel die Bezeichnung "Typ=k" in ihren Normdaten besitzen. Zur Bereinigung der extrahierten Artikel verwendeten wir das Python-Script "WikiExtraktor.py" (Quelle: https://github.com/attardi/wikiextractor), welches den Klartext der Wikipediaartikel im Ordner "Ergebnisse/AA/wiki_00" abspeichert.

2.3 Suche anhand von Vergleichsdaten

2.4 Erstellung des Klassifikators

3 Probleme bei der Lösung der Aufgabe

4 Ergebnisse