Ausarbeitung zur Projektarbeit Intelligente Systeme

Klassifizierung von Fließtexten

Antonia Berg und Kai Köster

17. Januar 2016

# Einleitung

Das vorliegende Projekt beschäftigt sich mit der Klassifikation von Texten, wobei die Klassifikationstechnik "Naive Bayes-Klassifikation" genutzt wird. Beispielhaft wurde die Klassifizierung anhand von Film- und Nachrichtentexten getestet, für die zu Beginn Test- und Trainingsdaten vorlagen. Zur Klassifikation können bestimmte Merkmale der Texte herangezogen werden, wobei durch die Trainingsdaten die Referenzwerte erstellt werden.

# Vorgehensweise und Implementierung

Unser Programm analysiert die angegebenen Merkmale der bereitgestellten Texte. Um möglichst einfach neue Merkmale hinzufügen zu können, haben wir die Klasse Criterion erstellt in welcher der Kriterientyp, der Inhalt und letztlich die Ergebnisse gespeichert werden.

In der Klasse TextClassification werden die Texte analysiert und klassifiziert. Dafür werden die erforderlichen Daten gespeichert, zum Beispiel die Gesamtanzahl der Wörter und Sätze pro Text.

In der Trainingsphase werden dann die Referenzwerte für die einzelnen Kriterien ermittelt. Dafür gibt es zwei Funktionen für jede Kategorie eine. (siehe UML-Diagramm im Anhang)

In der Test- und Evaluationsphase werden die Werte berechnet und dann mit den Referenzwerten der einzelnen Kriterien aus der Trainingsphase verglichen. Danach wird die Treffergenauigkeit errechnet.

Zum leichteren Auslesen der Texte aus den Ordnern, erstellten wir die Klasse TxtFolderReader, welche Ordner mit .txt-Dateien auslesen kann und diese als Liste von Strings zurückgibt.

# Ergebnisse

Die gegebenen Texte sind von uns unter folgenenden Merkmalen analysiert worden: der durchschnittlichen Satzlänge in dem Text, sowie der Häufigkeit von „ , “ und „ ? “.

Dadurch erreichten wir eine Genauigkeit der Resultate im Test von 83,3 % für Filmbeschreibungen und 93,3% für „Zeit“-Artikel, wodurch wir eine Gesamtgenauigkeit von 88,3% erhielten.

Für den Datensatz der Evaluation erreichten wir 76,4% für Filme und 96,2% für die Artikel. Daraus ergibt sich eine Trefferquote von 86%.

# Fazit

Durch weitere Merkmale könnte man die Treffergenauigkeit die Treffergenauigkeit erhöhen. Merkmale lassen sich in dem vorliegenden Projekt einfach hinzufügen. Derzeit muss bei jedem Lauf des Programmes die Trainingsphase durchlaufen werden, zukünftig wäre es ratsam die Werte der Kriterien zu speichern, zum Beispiel in Form einer .csv-Datei, um die neuen Texte schneller klassifizieren zu können.