

Themenspezifische Gruppierung deutscher Online-Zeitungen mit Natural Language Processing

Bachelorarbeit

Georg Donner
Matrikel-Nummer 553821

Betreuer	Prof. Dr. Gefei Zhang
Erstprüfer	Prof. Dr. Gefei Zhang
Zweitprüfer	Prof. Dr. Barne Kleinen

Inhaltsverzeichnis

1	Einleitung	2
	Literaturverzeichnis	3
	Abbildungsverzeichnis	4

1 Einleitung

Natural Language Processing ist ein großes Feld, welches besonders in der letzten Zeit im Zuge der Digitalisierung viel an Aufmerksamkeit und Wichtigkeit gewonnen hat. Es ermöglicht uns Informationen schneller zu finden, Systeme durch gesprochene Sprache zu steuern oder ganze Texte zu generieren. Eine weitere Aufgabe ist es, eine große Menge an Texten in Kategorien einzuteilen, um die Daten auf eine gewünschte Teilmenge für eine spezifischere Suche oder Analyse zu reduzieren. Die Kategorisierung der Dokumente nach ihrem Inhalt ist hier der häufigste Anwendungsfall, es ist aber auch möglich Texte nach ihrem generellen Genre oder Schreibstil zu vergleichen.

Diese Arbeit wird am Beispiel deutscher Online-Zeitungen untersuchen, welche Möglichkeiten es gibt Texte unabhängig von ihrem Inhalt zu vergleichen. Dabei werden lexikalische und syntaktische Merkmale, aber auch die Verwendung inhaltlich irrelevanter Wörter als Features verwendet. Es wird überprüft, inwiefern die Artikel gruppiert werden können und Rückschlüsse auf Unterschiede im Schreibstil ganzer Zeitungen statt nur einzelner Artikel zulassen.

Des Weiteren wird untersucht, ob und wie sich der Schreibstil einer Zeitung je nach Thema, wie z.B. Politik und Sport, unterscheidet.

Literaturverzeichnis

Abbildungsverzeichnis

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Ort, Datum

Unterschrift