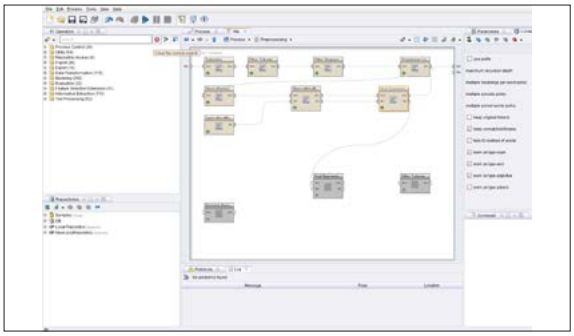
Quentin
Willimann

Diplomand	Quentin Willimann
Examinator	Prof. Hansjörg Huser
Experte	Stefan Zettel, Ascentiv AG, Zürich
Themengebiet	Software

Textmining

Textklassifikation und Generierung von automatischen Zusammenfassungen



Preprocessing in Rapidminer

Ausgangslage: Text Mining hat die Aufgabe, interessantes und nicht triviales Wissen aus unstrukturierten, bzw. schwachstrukturierten Texten zu extrahieren. Dazu müssen mehrere Fachrichtungen in Betracht gezogen werden: Informations Retrieval, Data Mining, Maschinelles Lernen, Statistik und Computerlinguistik. In dieser Bachelorarbeit sollen anhand von zwei Teilproblemen (Classification und automatic Summarization) die Prozessschritte von Textmining eruiert werden.

Vorgehen/Technologien: Für diese Bachelorarbeit werden zwei verschiedene Tools eingesetzt. Zum einen Rapidminer für die Entwicklung eines Klassifizierungsmodells und zum anderen Python für das gezielte Lernen und Bearbeiten der automatischen Zusammenfassung. Bei beiden Teilaufgaben geht es in erster Linie darum, den ganzen Ablauf zu modellieren. Danach wird versucht, iterativ die Resultate zu optimieren. Die erarbeiteten Schritte sind unter anderem folgende:

- Preprocessing,
- Postprocessing,
- Clustering,
- Classification,
- Evaluation.

Für die Klassifikation werden mehrere Verfahren (Support-Vektor-Maschine, Naïve Bayes, ID3, …) angewandt und miteinander verglichen. Nebst den herkömmlichen Algorithmen wird auch versucht, die Resultate mittels linguistischer Methoden zu verbessern. Dabei werden Prozesse verwendet, welche die Wörter in Typen unterscheiden (Substantive, Verben, Adjektive, …) oder die Wörter als Synonyme erkennen und zusammenfassen.

Ergebnis: Klassifikation: Da die gegebenen Dokumente eher kurz sind, ist eine genaue Vorhersage, zu welcher Klasse ein Text gehört, nicht ganz einfach. Nach dem Lernen des Klassifikationsmodells anhand von 1000 Texten pro Klasse wird versucht, eine kleinere Menge an Dokumenten vorherzusagen. Dabei erreichte meine Prozesskette eine Genauigkeit von rund 92,5% für zwei Klassen, 91,3% für drei Klassen und 83,5% für vier Klassen. Problematisch sind vor allem Klassen, die nahe beieinander liegen und sich je nachdem sogar überschneiden. Automatische Zusammenfassung: Auch bei dieser Problemstellung sind wieder die im Vorfeld erarbeiteten Prozessschritte involviert. Jedoch liegt hier das Augenmerk mehr beim Herausfinden der Themen, welche ein Dokument umfassen. Ein Erlernen und Clustern basiert hier auf noch kürzeren Abschnitten. Aus einem wissenschaftlichen Text werden Sätze selektiert, welche den Text möglichst gut zusammenfassen sollen. Dieser generierte Text wird dann mit Hilfe von Cosine Similarity mit dem verfügbaren Abstract verglichen.