Grundprimitive der Kategorisierung von Textdaten

Methodenvergleich am Beispiel von ZDF-Daten

Thema:

In dieser Arbeit werden drei grundprimitive Algorithmen zur Kategorisierung von Textdaten auf Inhalte der ZDF-Mediathek angewendet und verglichen. Ziel ist den Prozess der Kategorisierung nachvollziehbar darzustellen und das beste Modell für den domänenspezifischen Korpus zu finden. Zudem werden verschiedene Vorprozessierungen und Vektorisierungsmethoden getestet.

Datengrundlage:

Die ZDF-Mediathek ist in mehrere größere Rubriken aufgeteilt. Als Grundlage dienen Sendungen der Rubrik Dokumentation. Insgesamt besteht der Korpus aus **6058** Sendungen aufgeteilt auf **10** Sendungsreihen. Das Vokabular besteht aus **105.713** Types und insgesamt **1.427,596** Millionen Tokens. Die durch-schnittliche Textlänge pro Sendungen beträgt **235,74** Tokens.

Create
- Download der Daten aus der Content-Schnittstelle
- Speicherung der interessanten Schlüssel:Wert Paare
- Umwandlung in CSV-Format zur übersichtlicheren Speicherung

Process
- Entfernung von Außerreißern und fehlerhaften Daten
- Entfernung von Satzzeichen.
- Part-of-speech-Tagging, Stammformreduktion und Lemmatisierung der Daten

Vectorize
- Vektorisierung der Daten mit der einfachen Termfrequenz (TF), der One-Hot Enkodierung und der Termfrequenz – Inverse Dokumentenfrequenz (TF-IDF)

Classify
- Training eines Naiven Bayes Klassifikators (NB), einer Logistischen Regression (LOG) und einer Support Vektor Maschine (SVM)

Validate
- Cross-Validierung der verschiedenen Vorprozessierungen, Vektorisierung und Klassifikatoren.
- Finden der besten Parameter für jeden einzelnen Schritt

Ergebnisse:

Beim Klassifikationsprozess machen vor allem die zugrunde liegenden Daten den größten Unterschied. Auf der technischen Seite scheint der Klassifikator den größten Ausschlag zu geben. Die Logistische Regression mit TF-IDF Enkodierung liefert insgesamt die besten Ergebnisse. Großen Einfluss hat letztlich die Datengrundlage. Für viele Klassen scheinen nicht genug Daten vorhanden zu sein. Die trainierten Modelle passen sich zwar gut an die Daten an, da allgemein eine gute Precision erzielt wird. Dafür verfehlen sie ungesehene Daten richtig zu kategorisieren (allgemein niedrige Precision). Die Unterschiede zwischen der Logistischen Regression und den Support Vektor Maschinen ist generell aber als gering einzustufen.

	Precision			Recall			F1		
Klassifikator	NB	LOG	SVM	NB	LOG	SVM	NB	LOG	SVM
Vektorisierung	TF	TF IDF	TF	TF	TF IDF	TF	TF	TF IDF	TF
Micro							0,67	**0,78**	0,73
Macro	0,63	**0,81**	0,74	0,45	0,57	**0,62**	0,43	0,62	**0,66**
Weighted	0,65	**0,79**	0,73	0,67	**0,78**	0,73	0,63	**0,76**	0,72

Diese Poster ist Teil der Masterarbeit: **Grundprimitive der Kategorisierung von Textdaten - *Methodenvergleich am Beispiel von ZDF-Daten***. Eingereicht am Fachbereich 07 Geschichts- und Kulturwissenschaften der Johannes Gutenberg-Universität Mainz im Dezember 2020 von **Timo Schumacher**.