# Mapping the news: Analyse von Zeitungsartikeln mit Natural Language Processing

Studiengang: BSc in Informatik | Vertiefung: Data Engineering
Betreuer: Prof. Dr. Erik Graf
Experte: Daniel Voisard

## Welche Auslandsthemen haben in den letzten 17 Jahren die Schweizer Tageszeitung „Der Bund“ beschäftigt und wie können Machine-Learning-Technologien helfen, diese Frage zu beantworten? Die Arbeit will exemplarisch aufzeigen, was in der wissenschaftlichen Medienanalyse mit den entsprechenden Tools des Natural Language Processing möglich ist.

### Vision

In den letzten Jahren wurde dank der Technologien im Big Data- und Machine Learning-Umfeld ein gigantisches Potenzial für die Auswertung grosser Datenmengen im Medienbereich eröffnet. Die zunehmende Digitalisierung von Medieninhalten, z.B. durch die Verfügbarkeit von Zeitungsartikeln in Online-Archiven, hat ebenfalls zu diesem Trend beigetragen.
Die Bachelorarbeit will dieses Potenzial nutzen und exemplarisch aufzeigen, was in der wissenschaftlichen Medienanalyse mit den entsprechenden Tools möglich ist. Hierbei werden sämtliche Auslandsseiten der Jahre 2006 bis 2022 aus der Berner Tageszeitung «Der Bund» mit Natural Language Processing auf ihre Themenschwerpunkte hin analysiert. Es handelt sich insgesamt um rund 5000 Ausgaben.

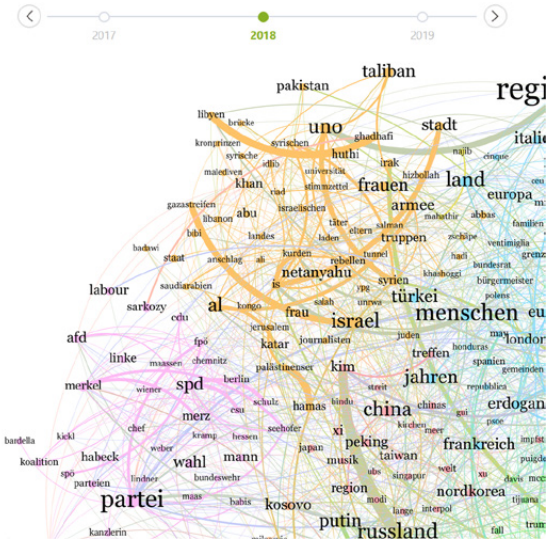
### Umsetzung

Für die Umsetzung des Projektes wurden die in einem früheren Projekt gewonnenen PDFs der Auslandsseiten mithilfe von Apache Tika paragraphenweise ausgelesen und anschliessend in einer MongoDB gespeichert. Für die Themenextraktion wurde BERTopic, eine auf Bidirectional Encoder Representations from Transformers (BERT) basierende Pipeline, genutzt. BERTopic ist auf das Extrahieren von Themen ausgelegt und beinhaltet im Einzelnen die folgenden Schritte:

- Daten in numerische Werte umwandeln
- Dimensionalität reduzieren
- Daten clustern
- Aufteilung der Themen in Tokens
- Cluster gegeneinander abgrenzen

### Ergebnis

Die so gewonnenen Themen wurden in der Graphdatenbank neo4j gespeichert, wobei die Themen als Knoten und die Erscheinungsdaten als Kanten definiert wurden. Anschliessend konnten die Graphen mit dem Visualisierungstool Gephi zu einer Themenlandkarte jahrweise zusammengefasst werden. Um die Lesbarkeit zu vereinfachen wurde das Louvainverfahren angewendet, das die Struktur von Graphnetzwerken in Cluster unterteilen kann. Dadurch konnten farblich gegeneinander abgegrenzte Themenbereiche sichtbar gemacht werden. Auf der Webseite mappingthenews.ch kann man mittels einer Zeitleiste durch die Karten der einzelnen Jahre navigieren. Auf diese Weise können spannende Veränderungen der Themenlandschaft im Zeitraum von 2006 bis 2022 beobachtet werden.



Rebecca Tabea Vogt
rebecca.vogt@gmx.ch