Scraping, Embedding and Clustering articles from Swiss online news outlets

Student

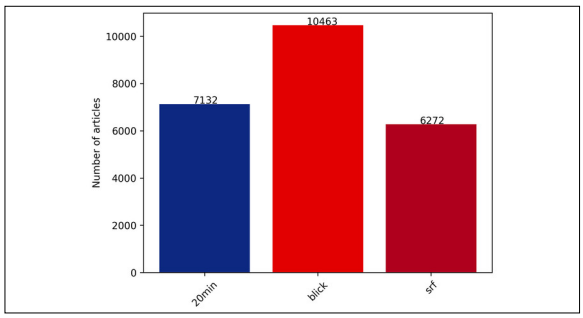
Moritz Schiesser

Initial Situation: The landscape of the Swiss online news media is diverse and complex.
The most popular news outlets, especially when offered for free and supported by advertising, are subject to the constraints of profitability and thus are forced to produce content that is appealing to the majority of the readership.
Further, the editors wield massive power on the public opinion, therefore influencing the political future of the country.

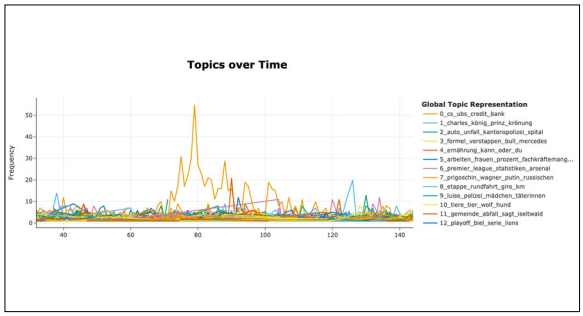
Definition of Task: In preparation for sentiment analysis in a future project, a initial system for scraping and structuring the data is created, and a workflow to embed and cluster the data using various clustering algorithms is described.
A variety of embedding models, among them transformer based models, are combined with simpler and modern clustering techniques through the use of BERTopic, a library specifically designed to detect topics from text documents. The results are compared through defined metrics, where especially their performance on clustering is of interest. Using established metrics such as the Silhouette score, the Davies-Bouldin Index and the Calinski-Harabasz Index, and a new metric that measures the diversity of the labels assigned of the clustering, dozens of configurations are tested, explored and discussed.

Conclusion: The comparisons show that large modern transformer based models, specifically trained or fine-tuned on German data, such as `deepset/gbert-large` used with modern and established clustering algorithms such as `HDBSCAN` outperform simpler techniques like `TF-IDF` and `KMeans` when applied to the dataset mined and refined with this project.

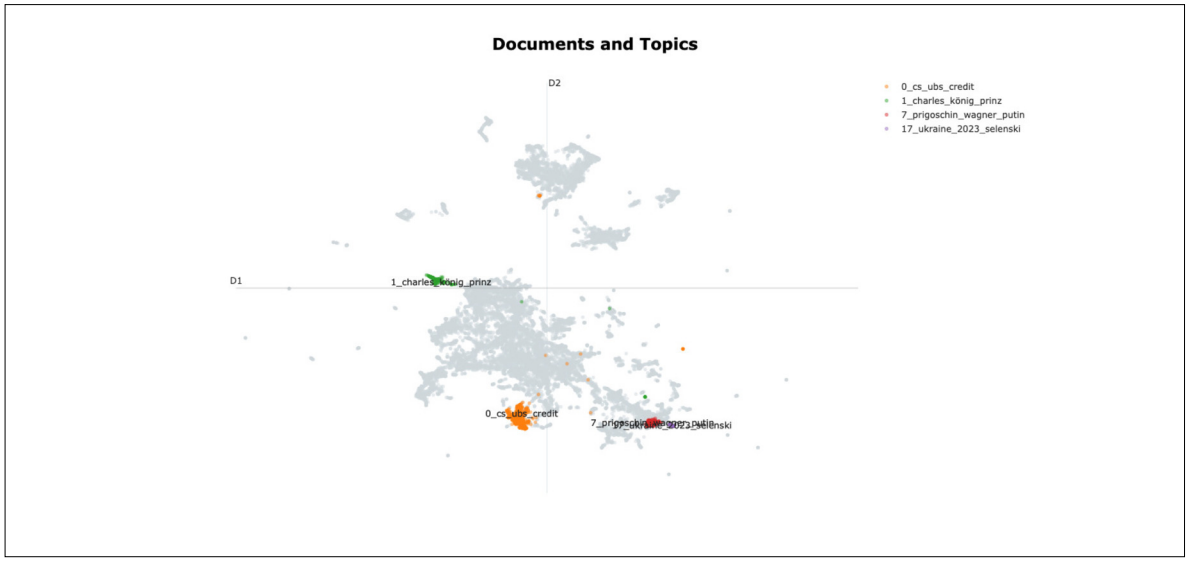
The number of scraped articles per news-outlet.
Own presentment



New articles per topic over time. Nicely visible is the peak of the reporting on the collapse of the Credit Suisse 2023.
Own presentment



Visualization of select clusters in a 2D Space.
Own presentment



Advisor
Prof. Dr. Mitra Purandare

Subject Area
Data Science