NLP – Project Proposal

Question/need:

- How did the topics in the first half of 2022 covered by German newspaper Frankfurter Allgemeine Zeitung (FAZ) change over time?
- Which topics of online newspaper articles affect the amount of traffic and conversions generated by each article?
- The question may (1) help editorial boards to more deliberately control topics for published stories; (2) help journalists gain a better understanding on which factors correlate to audience interest.

Data Description:

- I will scrape 5-month article data from https://www.faz.net/faz-live/ which contains an overview of all articles published by FAZ and attached information such as article url, paid/free or publishing time.
- Based on this, I will scrape each article url for article specific information such as full body text, teaser text, header as well as article meta data such as source (news agency, print daily article, etc.), ressort or author.
- I will gain information on target variables „Visits“ and “Conversions” for each article from my employer, FAZ.
- An individual unit of analysis will be a certain article with its body text and features.
- Based on the corpus of published articles’ texts I will conduct topic modeling to monitor coverage of topics over time and generate features for two regression models predicting visits and orders
- Target variable for my regression model will be (1) the number of Visits for each article and (2) conversions of each (“last viewed”) article.
- My features include article text, article source (news agency, daily print newspaper), paid/free status, length of the teaser text (no. of words), length of the article itself, existence of digital pages a reader has to click on to read entire article, article ressort (finance, politics, etc.), approximated reading time displayed at the top of article, publishing time (morning, noon, evening, night), publishing part of week and author.
- If time allows I will compare topics covered by FAZ to topics covered by its main competitor Sueddeutsche Zeitung (SZ) based on teaser texts of articles published in the first three weeks of June.

Tools:

- For webscraping I will use request and beautifulsoup in combination with CURL inserter for converting cookie information from network traffic console from curl to Python to enable scraping of paid articles.
- I will use text preprocessing libraries NLTK, spaCy and genism in combination with scikit-learn and statsmodels for linear regression
- Additionally, I will use pandas and numpy for data preparation as well as matplotlib and seaborn for visualization.
- If time allows I will build an interactive website UI for journalists using Flask.

MVP Goal:

- MVP contains line plots to display coverage of the main topics over time for the first half of 2022.