Text Technology

Summer Semester 2025
University of Stuttgart

Date: 05-06-2025
Team Members:
1. Sanjay Dutta (Matriculation Number: 3802726)
2. Udyavara Vasundhara Shenoy (Matriculation Number: 3802768)
3. Rida Iftikhar (Matriculation Number: 3757664)

Project Title: NewsFocus: Bias Tracking and Emerging Topics across News Outlets

Project Description:

Step 1: Collect and fetch news articles from various news organisations based on a particular topic. Use APIs or scraping scripts to collect the following items-

- Title
- Author
- Publish Date
- Source
- Article_Content and many more…

Step 2: Process the collected data and perform various kinds of analysis, like determining possible bias, creating a summary, understanding the topic evolution, and trend analysis of the topic from the various news organisations reporting about it. Store the data in XML format (create an XML schema if required to validate).

Step 3: Allow user access to the processed data and run various queries. Store the data in NoSQL/ Knowledge graphs and perform queries like topic relevance within the specific time period, identify topic connections to one another, popular genre covered by a particular news organisation, etc. Use XSLT to generate HTML for easier human consumption.

Technology Stack:
Programming Language: Python, Django
XML Tech: XML + XML Schema (or RELAX NG), XSLT
Database: NoSQL (MongoDB)
Extension: spaCy-based opinion and entity detection, Neo4J, RAGs
Potential Tools: Langchain, LLM

Optional Extensions:

1. Integrate with RAGs to determine the *political bias* regarding the topic concerning the news organisation.

2. Form clustering based on topic type, like Sports, Politics, Music, Academics, etc

3. Use of RELAX NG