Natural Language Processing

Problem Formulation - Clustering News Outlet Output

Paul & Zoe & Max

1 Motivation

- Overview of content on news channels broadcasting in English on YouTube.
- Comparing the focus of the topic areas of the different news stations with each other.

2 Goal(s)

- **Clustering of data by topics** (i.e. topics like "Ukraine War", "Economy", "Politics", ...)
  - Different methods of clustering
    - training based modells
      - active learning
      - supervised learning

      If needed, since the data is raw, there is no training data. A training data-set would have to be created. Also, in light of newly emerging topics, this is most likely not the best approach.
    - none training based modells
- **Comparison of clusters within single news station**
  - Distance between clusters
  - Size and popularity of clusters
  - Cluster development over time
  - Relation between emotionality of title and view count
  - Sentiments of news channel on identified topics
  - ...
- **Comparing clusters of different news channels**

- Differences and overlaps of clusters
- Comparison of similar clusters (i.e. matching categories)
- ...

3 Data

We have collected the **metadata of uploaded videos** on the following Youtube news channels (DW-News, CNN-News, BBC-News, Al-Jazeera-English, Fox-News and CCTV-Video-News-Agency) since 25-03-2023. The collected information of the videos has the following attributes.

Title	Views	Video-Length	Description	Time-Upload	Time-Crawling	Channel
...	...	...	...	...	...	...

The data is retrieved every 15 minutes. In the appendix you can find a part of such a data set.

4 Models/methods/algorithms

5 Tasks of each team member with description

- Data processing (Zoe)
- Create model overview (Max/Paul)
- Problem formulation (Paul)
- Power Point presentation (Zoe)
- Model application & tweaking (all)

6 Your plan and schedule for the project steps

- Data processing -
- Model selection -
- Model application -
- Evaluation -

7 Potential problems

- No sufficiently large data set.
- Incomplete video description.
- During scraping deleted commas and semicolons.