

Natural Language Processing

Project Description - Clustering of YouTube News Channel Videos based on Title & Description

1 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
...

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

2 Goal(s)

- **Clustering of data by topics** (i.e. topics like "Ukraine War", "Economy", "Politics", ...)

- Different methods of clustering

<div><div><div>* training based modellss</div><div><div>· active learning</div><div>· supervised learning</div></div></div></div>	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 modellss	

- **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)
 - ...