Metis NLP Module

Natural Language Processing on TED Talk Transcripts

Abstract

Social platforms have always played a major role in influencing our society. It is important to understand the direction of the content which is being exposed to the public. One such platform has been TED Talks which allows experts from different backgrounds to share ideas or personal experiences to the public. My aim was to dig deeper into these transcripts to understand the topics that were influencing most people and if these topics can be used to predict the views.

Design

Data was downloaded from Kaggle as a CSV file into pandas. Unnecessary columns were dropped and transcripts were cleaned using Regex and NLTK. For modeling I used TFIDF and NMF. 25 topics were clearly identified and named. These topics were used to fit into a linear regression hoping to predict the no.of views.

Data

Data contains about 2467 transcripts from 2006 to 2017 and talks were filtered with a minimum word count of 100.

Algorithms

- Data Cleaning
  - Regex and NLTK were used to perform preprocessing which included:
    - Removing unnecessary characters

        - Created custom stop words based on the initial results
- Modelling
    - TFIDF was used for vectorization and nmf for performing topic modelling.
- Visualizing Data
    - Word clouds
    - Matplotlib

Linear regression couldn't perform well on the data to predict the number of views. Looks like there was no linear relationship between the target and features. Finally, I performed sentiment analysis using Vader.

Future projects for this data could be to identify phrases instead of single words. Other regression models will also be used.

Tools

- Regex and NLTK for cleaning
- TFIDF and NMF were used for topic modeling
- Tableau for visualization

Communication

Presentation to the class