TED TALKS Topic Modeling

Natural Language Processing Final Project Report

Hayat Aldhahri & Juri AlSayigh 11/11/21 NLP

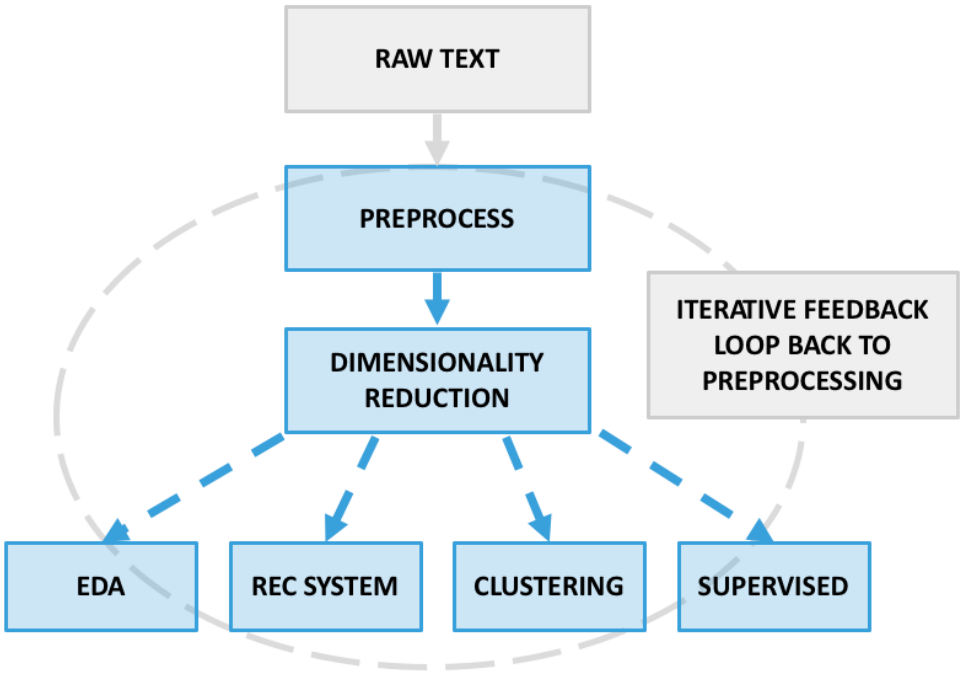
أكاديمية سدايا
SDAIA Academy

Overview

TED talks have been inspiring people since 1987 when it was founded by Richard Saulman as non-profit organization. Gathering speakers and experts from different fields such as technology, history, art, entertainment, and designs to share their ideas and thoughts for the public. Operating under the theme “ideas worth spreading”, TED talks videos and audio data are available online, uploaded in TED official website ever since it was first filmed in 1994.

In this project, we will be analyzing the dataset using machine learning algorithms to find the most inspiring topics and talks among the 2,550 talks.

As for the model development, we will be following the presented workflow for data analysis.



Data Description

There are two dataset used for this project.

- ted_main.csv
- transcripts.csv

The dataset contains audio-video information of TED talk uploaded in their official website. The recording are from 1994 to 2017. The dataset has been scraped from the official website and it is available as csv files in Kaggle.com for data analysis. The below table highlights are the features from the file “ted_main.csv” 17 features and 2,550 rows.

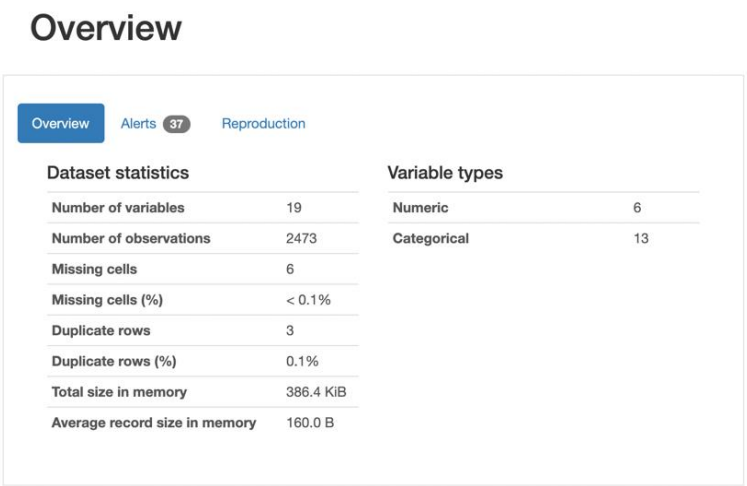
FEATURE	DESCRIPTION
COMMENTS	The number of comments for the topic
DESCRIPTION	Brief description on the topic presented
DURATION	The duration of the TED talk in seconds
EVENT	The event where the TED talk took place
FILM_DATE	The Unix timestamp of the filming The seconds count started on January 1st, 1970
LANGUAGES	The number of languages in which the TED talk is available
MAIN_SPEAKER	The main speaker of the TED talk The first named speaker of the talk
NAME	The official name of the TED talk (title and speaker)
NUM_SPEAKER	The number of speakers in the talk
PUBLISHED_DATE	The Unix timestamp for The published date of the TED talk
RATINGS	A stringified dictionary of the various ratings given to the talk (inspiring, fascinating, jaw dropping, etc.)
RELATED_TALKS	A list of dictionaries of recommended talks to watch next
SPEAKER_OCCUPATION	The occupation of the main speaker
TAGS	The themes associated with the talk
TITLE	The title of the talk
URL	The URL of the talk
VIEWS	The number of views on the talk

The transcripts.csv file has the following feature:

- Transcripts: The official English transcript of the talk.
- URL: URL of the talk.

Algorithm

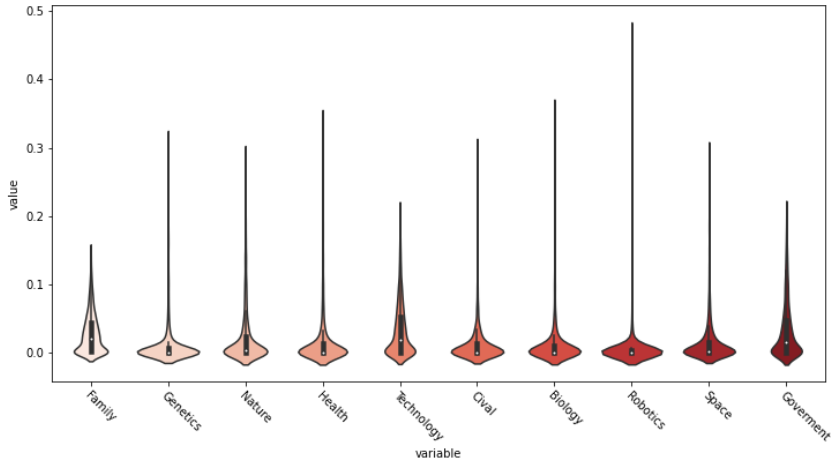
- Data Pre-Processing:
  - Ensure the data has no nulls/missing values.
  - Ensure there are no duplicate values.
- Joining the two datasets (ted_main.csv) and (transcripts.csv).
- Performed Pandas Profile Report for EDA.



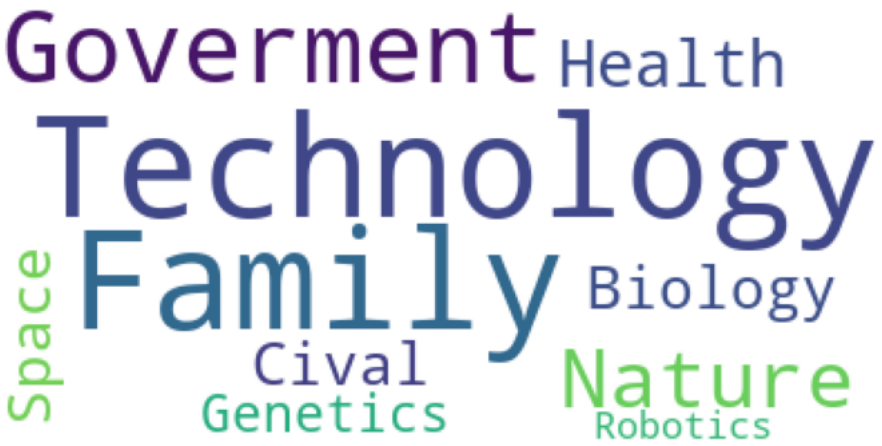
- Using the transcript column for text prepressing:
    - Change the text to lower case.
    - Remove everything between brackets .
    - Remove stop words.
    - Remove frequent words.
    - Remove rare words.
    - Apply lemmatizing.
    - Attempt to correct spelling but the transcript was already correct.

- Return the transcript column to the Data frame after text prepressing.
- As for the topic modeling part: Our objective for the NLP to extract and categorize the ideas and thoughts in the transcripts into specific topics. we started using LDA with 'tfidf' for the model. The results from the LDA model show that there are words that are not related to each other and repeated words. As a result, we try to use NMF to optimize the topic molding in our final evaluation and we found with NMF, the topic modeling results are much better in terms of a smaller number of repeated words and that the words can be classified into specific topics. For instance, the first list (girl, men, mother, father and god) can be categorized into one topic "Family".
- Depending on the words of each topic we named the topics accordingly as the following [Family, Genetics, Nature, Health, Technology, Civil, Biology, Robotics, Space and Government].
- Reflect the Topic labels in the data frame.

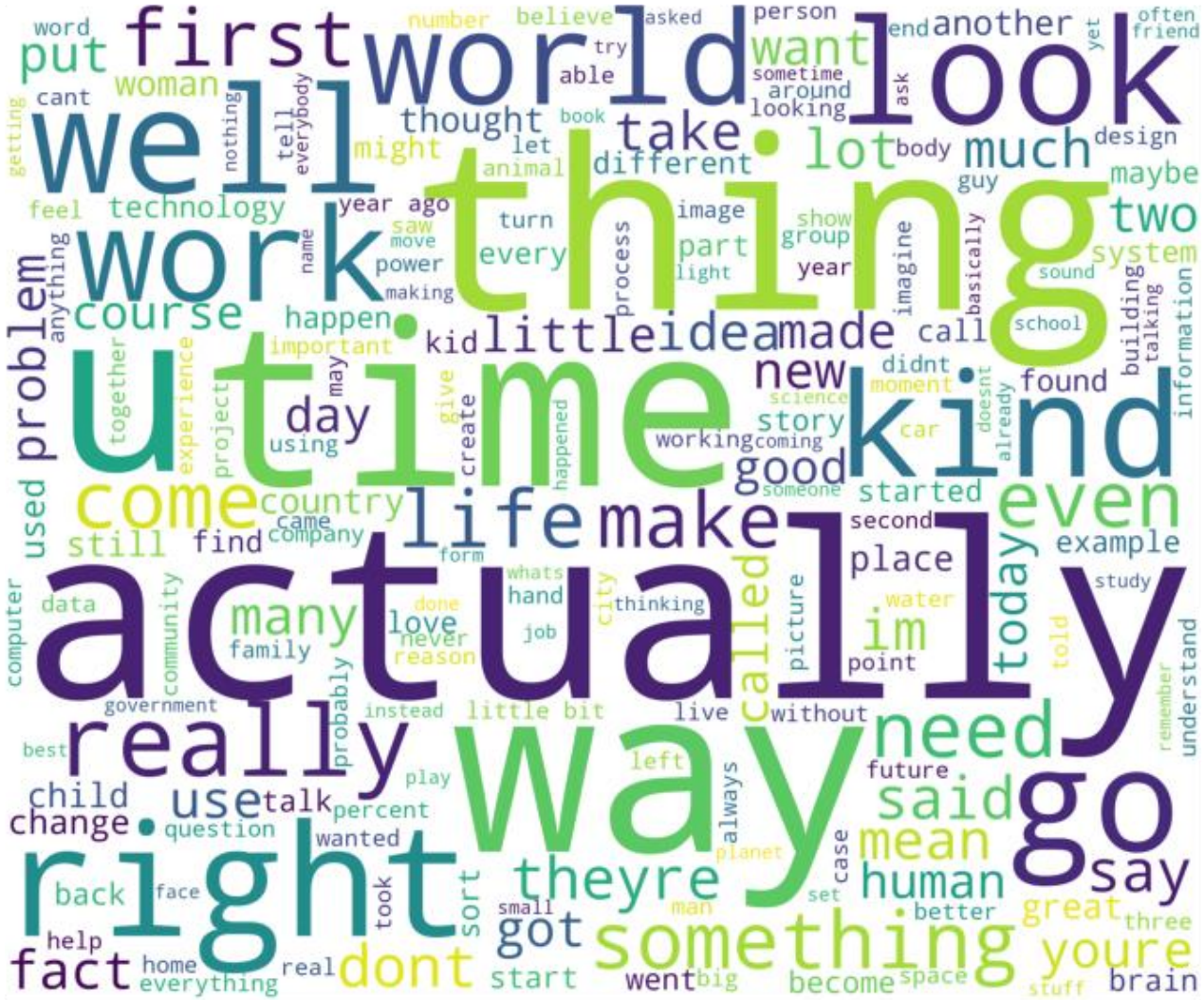
Visualization:



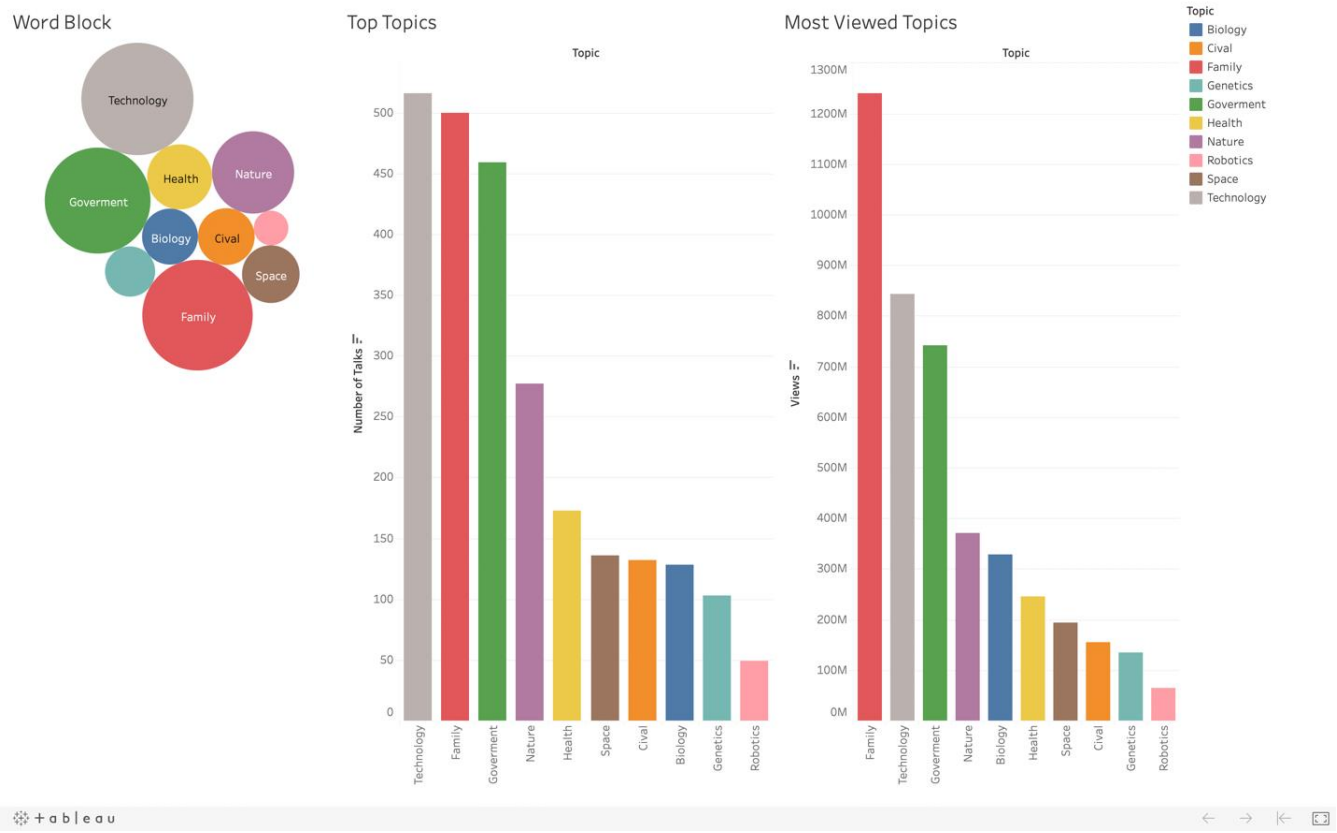
The above figure illustrates the distribution for each topic. Topic "Robotics" appears to have the strongest outliers among the others. As for the most frequent topics, Topics "Family, Technology and Government" appear to be most frequent and well distributed. The below is a word cloud created for the Topics, which shows the most frequent topics appeared in Ted Talks.



As for the transcript, a word cloud can also be created to illustrate the most frequent words mentioned in Ted Talk since 1994.



A dashboard has been created in Tableau to show most viewed topic and most appeared topics.



It appears that most appeared topic in Ted Talk is “Technology”. However, topics labeled “Family” are most viewed and most popular in the platform and had gathered over than 1.2 billion views since 1994.

Tools

- Pandas and NumPy packages to data manipulation.
- Seaborn and Matplotlib libraries for data visualization.
- Nltk and spaCy for text preprocessing.
- sklearn's CountVectorizer and TFIDF.
- Sklearn preprocessing.
- re: Regular expression.
- Gensim for LDA Topic modeling.
- MNF Topic modeling .
- Word cloud for topic visualization .
- Teablu..
- Jupyter notebook that hosts the code.
- Prezi for presentation.

أكاديمية سدايا
SDAIA Academy

Communication:

Please refer to the presentation for more insights.
(https://prezi.com/view/4xzpr5FTeBXe4qou9Sfd/)
Code can be accessed in the following link to Github.
(https://github.com/Jurisayigh/NLP-Project/blob/main/NLP_Code.ipynb)
(https://github.com/hayataldhahri/NLP)
Also Tableau dashboard can be accessed through the following public link
https://public.tableau.com/app/profile/hayat4538/viz/NLP_TedTalk/Dashboard1