
# TED TALKS Topic Modeling

## Natural Language Processing MVP Project

Hayat Aldhahri & Juri AlSayigh 11/9/21 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. The objective is to utilize NPL modeling to extract the topics from the transcript of all TED talks ever recorded.

# MVP Goal

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.
We started with data preprocessing by joining the two dataset and then cleaning the data .

## ProfileReport was performed for EDA.

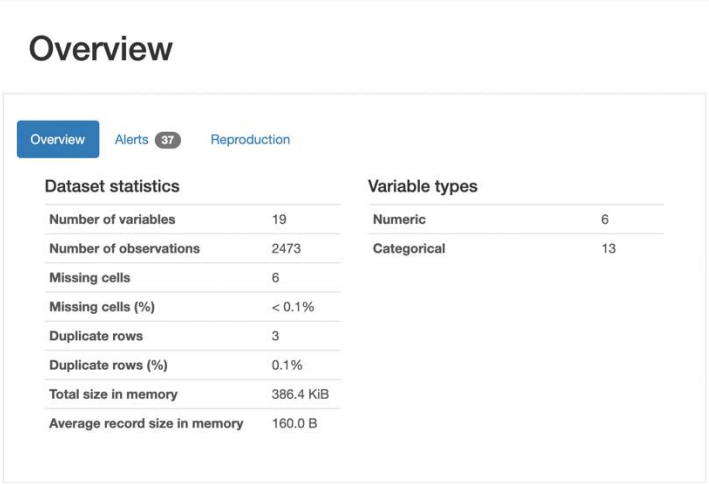
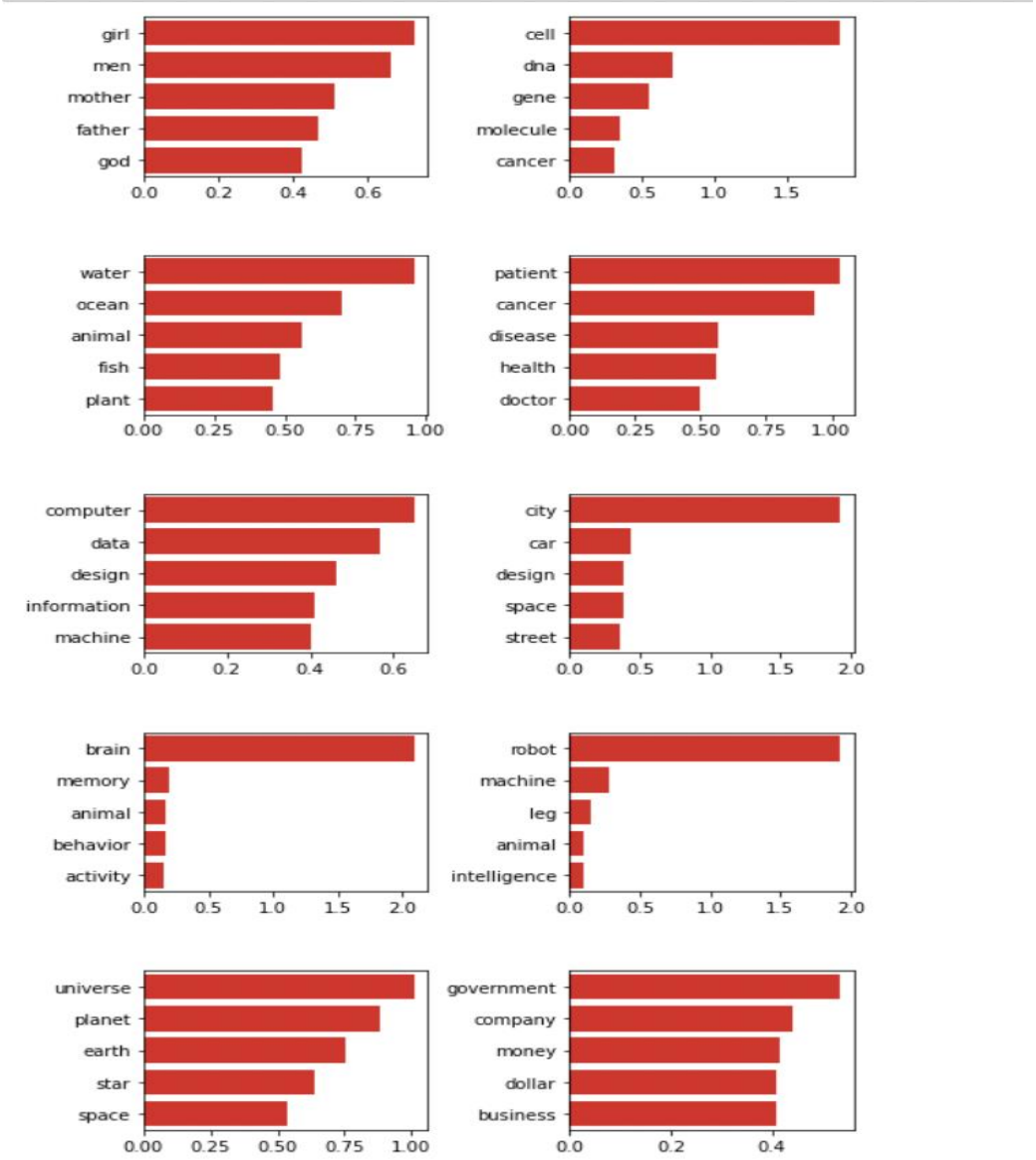



*Figure 1: ProfileReport Overview of the Dataset*

Our objective for the NLP to extract and categorize the ideas and thoughts in the transcripts into specific topics. we started using LDA for the model and below are the findings

```
{0: ['economist', 'sustainable', 'survey', 'financial', 'investment'],
 1: ['brain', 'music', 'robot', 'sound', 'play'],
 2: ['la', 'sorry', 'oh', 'welcome', 'chose'],
 3: ['economist', 'sustainable', 'survey', 'financial', 'investment'],
 4: ['economist', 'sustainable', 'survey', 'financial', 'investment'],
 5: ['city', 'water', 'data', 'community', 'company'],
 6: ['economist', 'sustainable', 'survey', 'financial', 'investment'],
 7: ['economist', 'sustainable', 'survey', 'financial', 'investment'],
 8: ['economist', 'sustainable', 'survey', 'financial', 'investment'],
 9: ['web', 'internet', 'link', 'page', 'government']}
```

*Figure 2: LDA Results*



We can notice from the above results there are words that are not related to each other and repeated words.

To optimize the topic molding we used NMF in our final evaluation, and blew are the results.

```
{0: ['girl', 'men', 'mother', 'father', 'god'],
 1: ['cell', 'dna', 'gene', 'molecule', 'cancer'],
 2: ['water', 'ocean', 'animal', 'fish', 'plant'],
 3: ['patient', 'cancer', 'disease', 'health', 'doctor'],
 4: ['computer', 'data', 'design', 'information', 'machine'],
 5: ['city', 'car', 'design', 'space', 'street'],
 6: ['brain', 'memory', 'animal', 'behavior', 'activity'],
 7: ['robot', 'machine', 'leg', 'animal', 'intelligence'],
 8: ['universe', 'planet', 'earth', 'star', 'space'],
 9: ['government', 'company', 'money', 'dollar', 'business']}
```

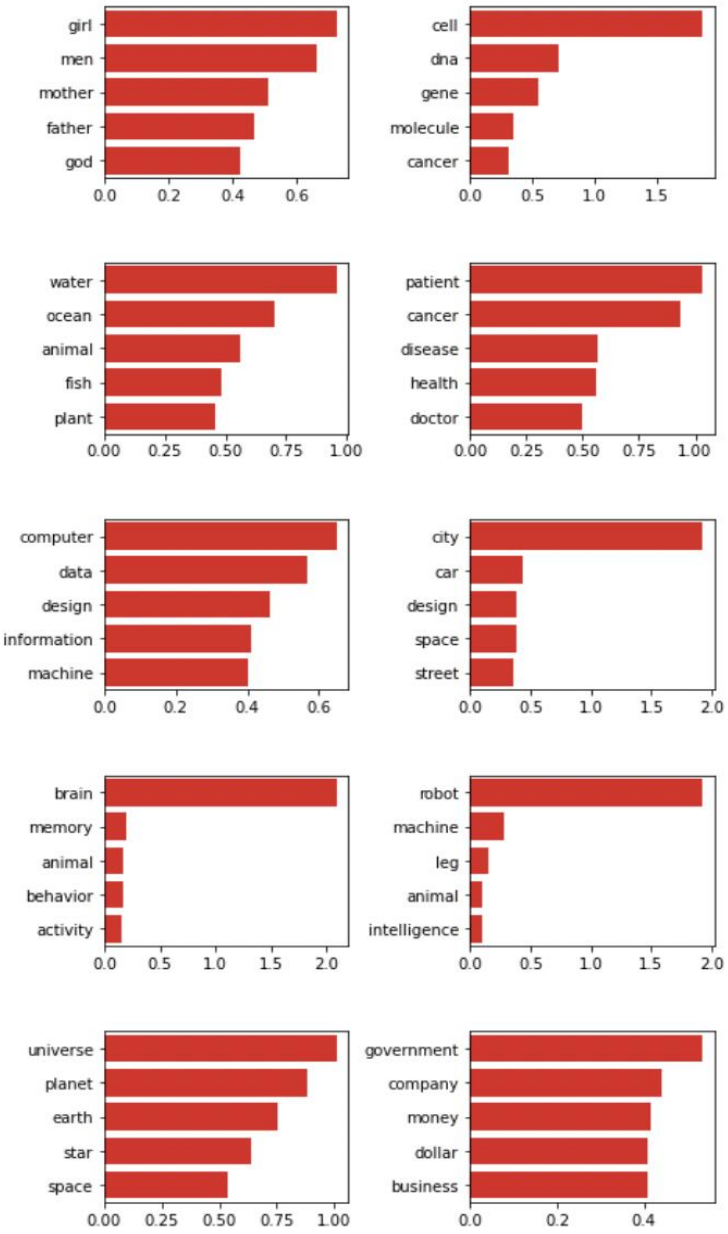



*Figure 3: NMF output*

It can be clearly seen that with NMF, the topic modeling results are much better in terms of less 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”.