Ted Talk Dataset Analysis

Dheeraj Nair (014705587)
Vaishak Melarcode Kallampad (015017496)
Viswamithra Vallabhaneni (015570516)
Yasaswi Mandava(015910583)
Prof. Mahima A. Suresh, SJSU CMPE 255

Abstract **: There has been an exponential growth in the digital content being uploaded on the internet over the years. This makes room for interesting analysis of both the content and the behavior of people who consume this content. In this project we try to analyze the Ted Talk dataset and come up with interesting insights from it. We also aim at using the content information to make meaningful recommendations. We have tried to integrate the results of our analysis and the recommendations functionality together in the form of a fun application for the users to explore.**

I. Introduction

The Internet has without doubt connected people across the world but more importantly it has driven the growth of many platforms where people can share ideas. Youtube is one of the main platforms we can consider as an example. The affordable prices of mobile phones and internet make such platforms easily accessible to a huge population. YouTube is the second most visited website in the world after Google hosting 122 million active users daily[9]. People do spend a lot of time on this website and for multiple purposes from entertainment, news to educational videos. In this project we plan to work on a small subset of the vast content that YouTube has, focusing only on Ted Talk videos. Ted talks are short and powerful talks organized with the motive of spreading ideas[1]. Ted is a non-profit organization which began in 1984 as a conference hosting people from different fields ranging from technology to entertainment to design. Today it deals with almost any topic you name including science, business and global issues. The diverse representation of the speakers and the issues chosen to talk about attracts an even more diverse audience. Ted talks being more of an educational or motivational nature can give us a lot of information from its contents. YouTube in general tries to engage its audience using the varied content in its database as well as the users' behavioral patterns[11]. Ted talks being informative and inspirational engages a user solely on the basis of its contents. So we have tried to engage the users of our applications by showing them content which they chose and then recommending similar content.

II. Dataset

The dataset for the project was obtained from kaggle[2]. It has three parts:

- Speaker data
- Talk data
- Transcript data

Name	Rows	Columns
speaker_data	4442	5
talk_data	4322	8
transcript_data	4442	2

Fig 1- Dataset Overview

Speaker data - The speaker data consists of the following information related to the speaker:

- Talk (the talk which he has given)
- Speaker Name
- Speaker Title (titles such as Dr., General, Imam etc)
- Speaker Occupation
- Speaker bio (brief description)

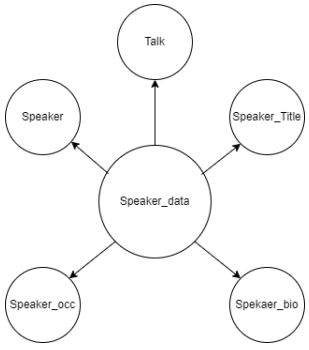


Fig 2- Speaker data

Talk data - This file contains specific information with regards to the talk.

- Talk Name
- Talk Description
- Ted Event
- Duration of talk
- Tags related to talk
- Number of Views
- Published on (Datetime)
- Recorded at (Datetime)

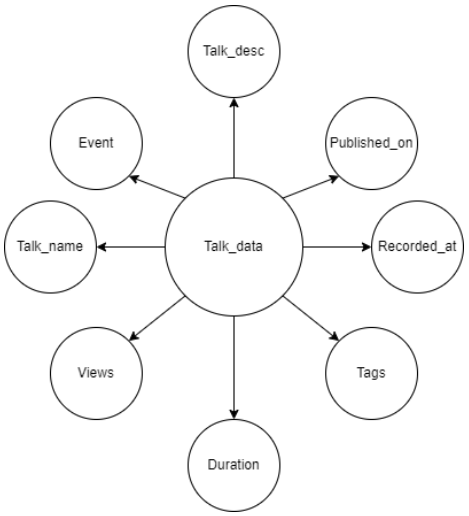


Fig 3 - Talk data

Transcript Data - This file contains information regarding the transcript.

- Talk Title(Name)
- Transcript of the talk

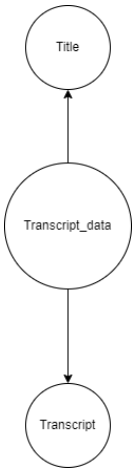


Fig 4 - Transcript data

We only have information regarding the talk and the speaker. Our dataset does not contain any information regarding the audience or viewers.

III. Data Preprocessing

As we can see from the Dataset overview in Fig 1 that the number of rows in all 3 files are not the same. There were some null values in all the files. The steps taken for data preprocessing were:

- Imported all 3 files using pandas and stored as a DataFrame
- Removed rows with Null values in the talk name from all the 3 dataframes
- Dropped the column of speaker_title from speaker data because it is not relevant and has majority values as Null.
- Merged all 3 dataframes into 1 dataframe based on the common column talk_name and removed the duplicated columns.
- Saved the cleaned dataframe as a new csv file.

The cleaned dataset has 4016 talks in total with the following columns:

```
In [4]: cleaned_df.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 4016 entries, 0 to 4015
        Data columns (total 12 columns):
         #   Column         Non-Null Count  Dtype
        ---  ------         --------------  -----
         0   talk           4016 non-null   object
         1   speaker        4016 non-null   object
         2   speaker_occ    4016 non-null   object
         3   speaker_bio    4016 non-null   object
         4   talk_desc      4016 non-null   object
         5   event          4016 non-null   object
         6   views          4016 non-null   int64
         7   duration       4016 non-null   int64
         8   tags           4016 non-null   object
         9   recorded_at    4016 non-null   object
         10  published on   4016 non-null   int64
         11  transcript     4016 non-null   object
        dtypes: int64(3), object(9)
        memory usage: 376.6+ KB
```

Fig 5 - Cleaned DataFrame

IV. Analysis and Findings

We have tried to analyze the cleaned data and gather some insights from it which might be interesting to the users. We have used matplotlib in order to display graphical representation of our findings.

Increase in the number of Ted Talks over the years

We used the recorded date of each ted talk to create a graphical representation of the number of talks in each year. The below figure was obtained as a result.

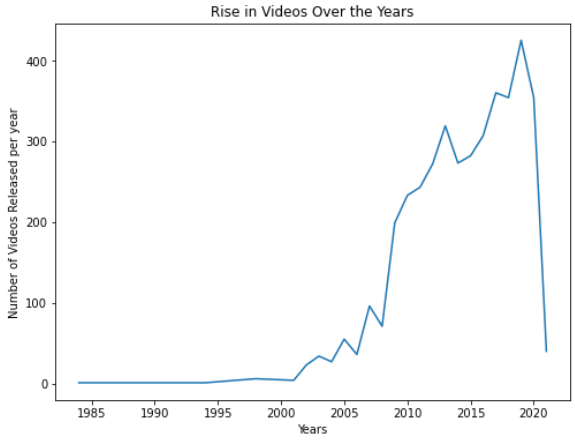


Fig 6 - Rise in Ted Talks

As seen in the graph we have data starting from 1984 all the way upto 2020. Initially the number of talks and conferences were less. After 2000 we can see a slight increase in the number of talks which drastically shot up around 2007-2008, the time when the internet started becoming more accessible. This shows the important role played by the advancements in technology on the number of talks recorded per year.

Most Popular Themes

We used the tags provided for each talk to understand which are the topics that were most common and were chosen by the speaker to talk about.

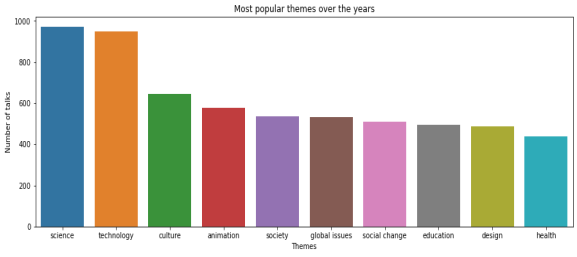


Fig 7 - Most popular themes

We find the most number of talks are based on science and technology, followed by culture, animation, society, global issues, social change, education , design and health. We restrict to showing only the top 10 themes to the user. This also portrays the audiences' choice of topics.

Most Popular Speakers

We have used the speakers' information to find out the most popular speakers of all time. For this we calculated the average view count of each speaker and then sorted it in descending order to get the top 10 speakers. To find the average views we found the total views for a speaker and then divide it by the number of talk shows done by the speaker. The figure below shows the most popular speakers based on average view counts.

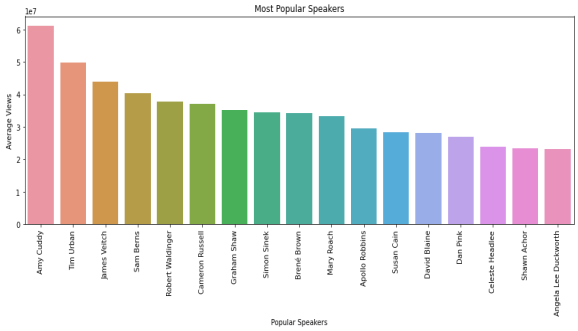


Fig 8 - Most popular Speakers

Most Popular Ted Events

Ted talks which were initially launched in the United States have spread to almost every continent today. There are ted events taking place in different parts of the world ,that too in multiple languages. We have tried to use the 'event' column which given the event information to which a ted talk belongs and the 'views' column of that ted talk to come up with the most popular ted events. The events which occur frequently and have a high viewership end up as the most popular ted events.

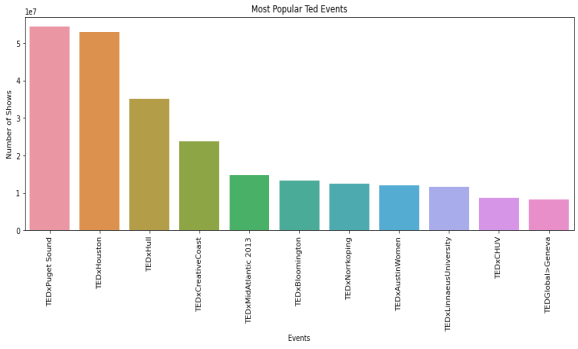


Fig 9 - Most Popular Ted Events

Must Watch Videos

In our application we display the top 25 ted talk videos of all based on the views of the videos. The more views imply higher popularity of the video. This list that we get is shown in figure 10 below.

```
These are our top 25 videos:
```

	talk	speaker	talk_desc	event	views
4012	Do schools kill creativity?	Sir Ken Robinson	Sir Ken Robinson makes an entertaining and pro...	TED2006	70176973
1959	This is what happens when you reply to spam email	James Veitch	Suspicious emails: unclaimed insurance bonds, ...	TEDGlobal>Geneva	63683410
2833	Your body language may shape who you are	Amy Cuddy	(NOTE: Some of the findings presented in this ...	TEDGlobal 2012	61030600
3437	How great leaders inspire action	Simon Sinek	Simon Sinek has a simple but powerful model fo...	TEDxPuget Sound	54351663
3299	The power of vulnerability	Brené Brown	Brené Brown studies human connection -- our ab...	TEDxHouston	52871680
1909	Inside the mind of a master procrastinator	Tim Urban	Tim Urban knows that procrastination doesn't m...	TED2016	49876247
2343	How to speak so that people want to listen	Julian Treasure	Have you ever felt like you're talking, but no...	TEDGlobal 2013	45961304
2147	The next outbreak? We're not ready	Bill Gates	In 2014, the world avoided a global outbreak o...	TED2015	40591593
1317	My philosophy for a happy life	Sam Berns	Born with a rare genetic disorder called proge...	TEDxMidAtlantic 2013	40315294
1964	What makes a good life? Lessons from the longe...	Robert Waldinger	What keeps us happy and healthy as we go throu...	TEDxBeaconStreet	37902552
2753	Looks aren't everything. Believe me, I'm a model.	Cameron Russell	Cameron Russell admits she won "a genetic lott...	TEDxMidAtlantic	37103025
1315	Why people believe they can't draw	Graham Shaw	Most people think they can't draw, but communi...	TEDxHull	35124222
3624	10 things you didn't know about orgasm	Mary Roach	"Bonk" author Mary Roach delves into obscure s...	TED2009	33320592
2555	The art of misdirection	Apollo Robbins	Hailed as the greatest pickpocket in the world...	TEDGlobal 2013	29455967
3011	The power of introverts	Susan Cain	In a culture where being social and outgoing a...	TED2012	28333512
3500	How I held my breath for 17 minutes	David Blaine	In this highly personal talk from TEDMED, magi...	TEDMED 2009	28045289

Fig 10 - Must Watch Videos

When comparing with Ted's official website we see that the videos which we recommend as must watch are in fact the most popular 25 videos as displayed on the official site as well. The screenshot below from www.Ted.com validates the results that we got.

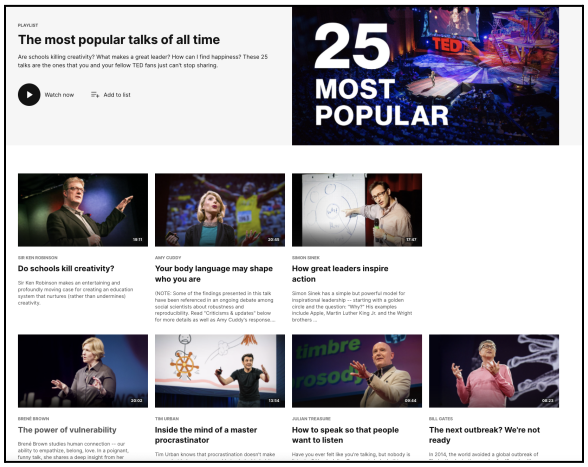


Fig 11 - Most Popular videos on Ted website

Popular Videos by Theme

We display the most popular themes based on the information in figure 7, in the form of a list to the users. A user can select a theme of his choice and then we display the top 50 ted talks related to that topic. This allows the users to choose from a limited set of really popular videos reducing their confusion a bit. Figure 12 shows how the list of popular videos are shown for a given topic selected by the user.

```
Here are few popular talk themes:

          1 science

          2 technology

          3 culture

          4 animation

          5 society

          6 global issues

          7 social change

          8 education

          9 design

          10 health

Which one would you like to checkout. Enter number:
```

```
Which one would you like to checkout. Enter number:
3

Here are the top 50 talk on  culture
```

	talk	speaker	event	views
4012	Do schools kill creativity?	Sir Ken Robinson	TED2006	70176973
3299	The power of vulnerability	Brené Brown	TEDxHouston	52871680
2343	How to speak so that people want to listen	Julian Treasure	TEDGlobal 2013	45961304
2753	Looks aren't everything. Believe me, I'm a model.	Cameron Russell	TEDxMidAtlantic	37103025
3624	10 things you didn't know about orgasm	Mary Roach	TED2009	33320592
3011	The power of introverts	Susan Cain	TED2012	28333512
2860	Strange answers to the psychopath test	Jon Ronson	TED2012	26975392
3557	The danger of a single story	Chimamanda Ngozi Adichie	TEDGlobal 2009	26846075
3687	Your elusive creative genius	Elizabeth Gilbert	TED2009	20146770
2707	My escape from North Korea	Hyeonseo Lee	TED2013	20122033
3992	The surprising science of happiness	Dan Gilbert	TED2004	19503481
2774	A Saudi, an Indian and an Iranian walk into a ...	Maz Jobrani	TEDxSummit	18071113

Fig 12 - Top Talks on Culture

The users can enter the associated number of the talk they find interesting to get more details.

Talk Details

When a user enters the number associated with a talk we display the following details to the user:

- Talk Name
- Speaker Name
- Speaker Occupation
- Talk Description
- Event
- Views
- Tags
- Wordcloud
- Transcript
- Recommendations based on this talk

All of the above information except wordcloud and recommendations are directly fetched from the cleaned dataset. Wordcloud is generated for each talk

using the transcript column[10]. The wordcloud can give the users an overview of the transcript and can help them decide whether to watch a particular talk or not.

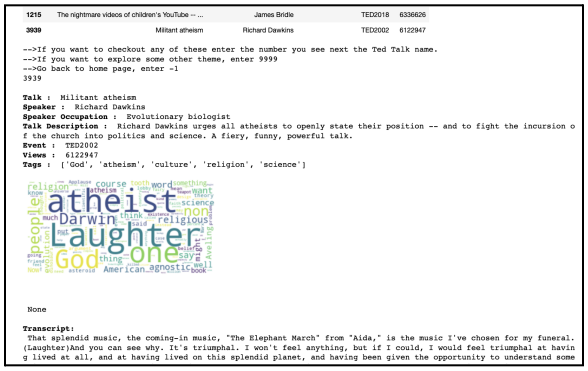


Fig 13 - Talk Details

V. Recommendation System

As mentioned above, we show recommendations to the user based on his selection of a talk. We again use the transcripts column for this purpose. The words in the transcript column are converted into vectors using TF-IDF vectorizer. Term Frequency-Inverse Document Frequency (Tf-Idf) is used to measure the importance of a particular word in the transcript. If a word occurs a lot in one ted talk but is rare in the other ted talks, then that word holds a high importance in that particular ted talk.

Term Frequency-Inverse Document Frequency (Tf-Idf)

```
Text=cleaned_df['transcript'].tolist()
tfidf=text.TfidfVectorizer(input=Text,stop_words="english")
matrix=tfidf.fit_transform(Text)
print(matrix.shape)
```

```
(4016, 68448)
```

Fig 14 - TF-IDF Vectorizer

Once transcripts are converted into word vectors we know the words which hold most significance in a talk. We then use cosine similarity on the vector representation of each transcript and compare it to all the other transcripts to find the similar transcripts.

Cosine Similarity

Unigrams

```
In [6]: from sklearn.metrics.pairwise import cosine_similarity
        sim_unigram=cosine_similarity(matrix)
        sim_unigram.shape
Out[6]: (4016, 4016)
```

Fig 15 - Cosine Similarity

We have tried vectorizing using unigrams , bigrams and trigrams and have got different sets of recommendations for each using the same process. The recommendations obtained using unigrams were most accurate as can be seen in figure 16.

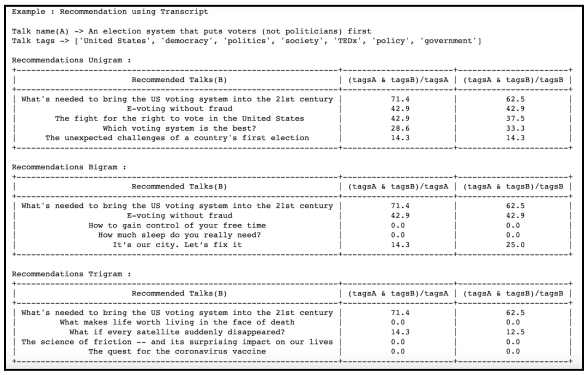


Fig 16 - Recommendation using transcripts

This can be due to the fact that using tf-idf we are not able to get the exact context of all the words so using bigrams and trigrams does not have a positive impact.

To evaluate whether the recommendations which we make are relevant or not we have come up with a metric. Suppose we have a talk A and we are recommending a talk B for this talk. Both these talks have a set of Tags related to their contents. We come up with 2 metrics :

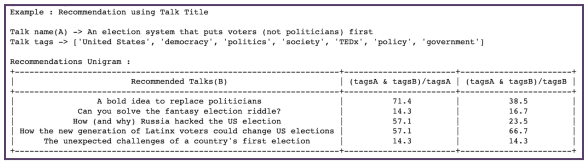
- (Number of common tags in A & B)/(number of tags in A) : a higher value indicates the common tags make up a higher portion out of the tags in A.
  (tagsA & tagsB)/tagsA

- (Number of common tags in A & B)/(number of tags in B) : a higher value indicates the common tags make up a higher portion out of the tags in B.
  (tagsA & tagsB)/tagsB

These metrics also back our apparent assumptions that unigrams gave better results than bigrams and trigrams.

We have also tried out using other columns like Talk Name and Talk Description to find similarity between 2 talks and then use that for making

recommendations. The recommendations obtained using these columns are also meaningful and are shown in the figure below.

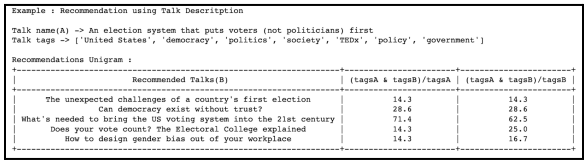


```
Example : Recommendation using Talk Title

Talk name(A) -> An election system that puts voters (not politicians) first
Talk tags -> ['United States', 'democracy', 'politics', 'society', 'TEDx', 'policy', 'government']

Recommendations Unigram :
+-----------------------------------------------------------------+---------------------+---------------------+
|                       Recommended Talks(B)                      | (tagsA & tagsB)/tagsA | (tagsA & tagsB)/tagsB |
+-----------------------------------------------------------------+---------------------+---------------------+
|                  A bold idea to replace politicians             |         71.4        |         38.5        |
|              Can you solve the fantasy election riddle?         |         14.3        |         16.7        |
|             How (and why) Russia hacked the US election         |         57.1        |         23.5        |
| How the new generation of Latinx voters could change US elections |         57.1        |         66.7        |
|       The unexpected challenges of a country's first election   |         14.3        |         14.3        |
+-----------------------------------------------------------------+---------------------+---------------------+
```

Fig 17 - Recommendations using Talk Name



```
Example : Recommendation using Talk Descritption

Talk name(A) -> An election system that puts voters (not politicians) first
Talk tags -> ['United States', 'democracy', 'politics', 'society', 'TEDx', 'policy', 'government']

Recommendations Unigram :
+-----------------------------------------------------------------+---------------------+---------------------+
|                       Recommended Talks(B)                      | (tagsA & tagsB)/tagsA | (tagsA & tagsB)/tagsB |
+-----------------------------------------------------------------+---------------------+---------------------+
|       The unexpected challenges of a country's first election   |         14.3        |         14.3        |
|                Can democracy exist without trust?               |         28.6        |         28.6        |
| What's needed to bring the US voting system into the 21st century |         71.4        |         62.5        |
|         Does your vote count? The Electoral College explained   |         14.3        |         25.0        |
|              How to design gender bias out of your workplace    |         14.3        |         16.7        |
+-----------------------------------------------------------------+---------------------+---------------------+
```

Fig 18 - Recommendations using Talk Description

For our main application we have kept the recommendations we obtain from transcripts since it encompasses a more detailed comparison of the talks. The recommendations are shown after other Talk details where the users can move on to the next video.

VI. Conclusion

This study offered a new concept for a TED talk video recommendation system and extended it into a command line application. We have used data mining techniques to show the recommendations using the Transcript data. The videos are recommended based on the most popular talks.

VII. Future Works

- Future work will include developing different playlists on topics such as Animals, Plants, or Religion.
- Creating the most popular playlist year-wise Example: Best of 2020, Best of 2019 etc.
- To get the context of the words right and make better recommendations, we could use different words vectorizers like word2vec or glove instead of tf-idf.

VIII. References

1. https://www.ted.com/about/our-organization
2. https://www.kaggle.com/
3. https://www.analyticsvidhya.com/blog/2021/04/a-guide-to-feature-engineering-in-nlp/
4. https://www.geeksforgeeks.org/data-visualization-with-python-seaborn/amp/
5. https://towardsdatascience.com/ted-talks-analysis-eda-for-beginners-df346bc431a6
6. Kharwal, Aman. “TED Talks Recommendation System with Machine Learning.”, June 21, 2021. https://thecleverprogrammer.com/2021/04/01/ted-talks-recommendation-system-with-machine-learning/
7. Maazouzi, Faiz, Hafed Zarzour, and Yaser Jararweh. "An effective recommender system based on clustering technique for ted talks." International Journal of Information Technology and Web Engineering (IJITWE) 15, no. 1 (2020): 35-51.
8. Oh, Jaehoon, Injung Lee, Yeon Seonwoo, Simin Sung, Ilbong Kwon, and Jae-Gil Lee. "TED talk recommender using speech transcripts." In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 598-600. IEEE, 2018
9. https://www.omnicoreagency.com/youtube-statistics/
10. https://www.analyticsvidhya.com/blog/2021/05/how-to-build-word-cloud-in-python/
11. https://www.thinkwithgoogle.com/data-collections/youtube-stats-video-consumption-trends/