What Makes A TED Talk Popular?

Ruixi Wang, Xin Zhao, Yuhui Ren, Zhizheng Li

Instructor: Prof. Rong Liu

1 Introduction

TED is a platform for sharing ideas, and TED Talks are influential videos from expert speakers on education, business, science, tech, creativity and so on. [1] In recent few years, TED Talks appeal to global citizens due to various merits, such as they are short but focused, shot like movies, highlight diverse speakers, discuss specific inspiring and informative topics, go beyond classrooms and offices. [2] And all these advantages are exactly the motivation that make our group willing to explore what else makes a TED talk popular with our knowledge.

After the initial analysis on the relations between specific dimensions such as on duration, tags, and popularity of topics in TED, it's time to further explore the common determining factors of topics that attracts viewers. From the initial analysis we've already known that the top 3 popular tags are science, culture, health. Is there any common factor that the most popular tags share? For instance, are people attracted by their instinct such as emotion preference? To burrow the hidden factors especially emotion we use sentiment analysis. After analyzing each transcript of topic by scoring eight main emotions, we adopt each topic into a logistic regression by using eight emotion as predictor variables and whether the topic is popular or not as the dependent variable. Relying on the regression model, we can see if some of the eight emotions are positively dominant on tempting viewers.

2 Objectives and Expected Contributions

The main objective of our work is to investigate what make a TED Talk popular within the following aspects:

- Is there any relation between speech duration and speech popularity? How does the duration affect the popularity?
- Which aspects are people interested in? Which tags for a TED Talk can make it have more views? Does number of tags for a TED talks affect the views?
- Which speakers are popular?
- Are there any specific words can make a topic gain more click? How about the sentiment for these words?
- How does the speech description affect view?
- What's emotions do general topics share? Is there any extraordinary topic that has different emotion pattern in hand?
- Whether there are some peculiar emotion patterns which tend to attract more viewers?

We expect our work can make following contributions:

- This work can provide insights about what makes a TED Talk popular. Through sentiment analysis, we could provide tips for speakers about how to create the topics and descriptions that can inspire people's interest most.
- We can give advice about which aspects' speeches should be made more.

3 Methodology

3.1 data collection

When we first time crawling the data from the website, we made some mistakes, for example: we used to locate some of information from the homepage such like topics, views, duration, and speakers and located the descriptions and relevant tags from the video paging. However, this will result in some information may located in wrong order. Because when we run to crawl the whole website, we may meet the anti-crawl project and ban our website. Or it may kill itself because of some errors. When we met the situation above, the information order will be missed up. For instance, the topic A may have the write speakers, duration, views because all of information are read from the homepage and the description and relevant for topic E because those information are from video paging.

To repair this problem, we crawled all the url links in the homepage and get all the details from the video paging. This will at least put all the information in the right order.

In addition, sometimes we may meet the anti-crawl program from the website. We set the program will read paging for 3 seconds and efficiently avoid most of anti-crawl programs. Moreover, the project will kill itself for no reason sometimes. Although if we rerun the program, it will be repaired but it is hard for us to locate where the project stopped last time. It will last a lot of times. So, we add a little routine, we use label 'topics' to test if the program had given the whole information. If it tests the same information, it will not be re-written in the CSV. It really saves a lot of time and effort.

In the final project, we have two datasets. Dataset 1 is the early data that we scraped from official TED Talks website (https://www.ted.com/talks), which contains information about each speech, including speech topics, speech descriptions, etc. Dataset 2 contains transcripts of TED Talks that we downloaded from Kaggle (https://www.kaggle.com/rounakbanik/ted-talks?select=transcripts.csv). Unfortunately, the dataset 2 is insufficient for popularity analysis. So, we also scraped transcript from official TED Talks website and match it with existing topics for further usage.

3.2 Data Preprocessing

The scraped duration for each video were in seconds, which was not easy to read while easy for data analysis. So firstly, we transformed the number for the duration to a more straightforward way, for example from 1262 seconds to 0:21:02 (0 hours 21 minutes 02 seconds) and call them as lengths.

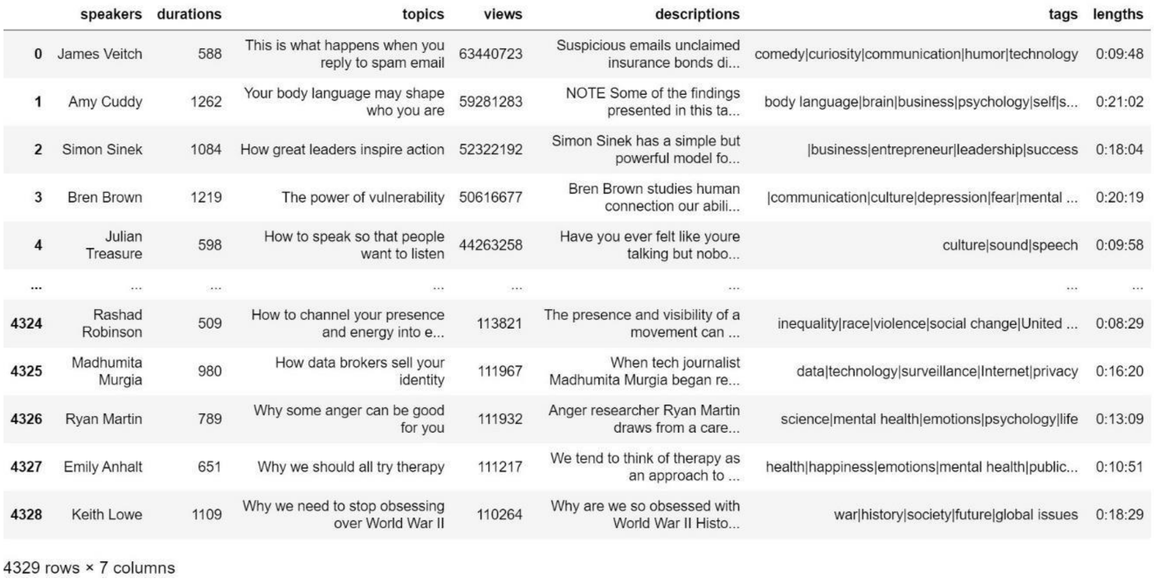
Secondly, we found there are some meaningless symbols in the text of speaker column, descriptions column and topics column, such as '?€?', we removed these meaningless symbol for future better interpretation and text mining.

Moreover, we also removed some needless suffixes, such as '(TED)' in lecturers' name, 'TEDx', 'TED-Ed', 'TEDMD' etc in tags to avoid meaningless high-frequency speakers' name or tags.

In some cases, if there are multiple speakers attend one same talk, the raw data used ‘+’ to connect speakers’ names. We replaced ‘+’ with the word ‘and’ to remain the same meaning.

3.3 Exploratory Data Analysis

To start with, we generated a data overview to display popular topics and the relational information about the topics, shown as figure 1.



	speakers	durations	topics	views	descriptions	tags	lengths
0	James Veitch	588	This is what happens when you reply to spam email	63440723	Suspicious emails unclaimed insurance bonds di...	comedy|curiosity|communication|humor|technology	0:09:48
1	Amy Cuddy	1262	Your body language may shape who you are	59281283	NOTE Some of the findings presented in this ta...	body language|brain|business|psychology|self|s...	0:21:02
2	Simon Sinek	1084	How great leaders inspire action	52322192	Simon Sinek has a simple but powerful model fo...	|business|entrepreneur|leadership|success	0:18:04
3	Bren Brown	1219	The power of vulnerability	50616677	Bren Brown studies human connection our abili...	|communication|culture|depression|fear|mental ...	0:20:19
4	Julian Treasure	598	How to speak so that people want to listen	44263258	Have you ever felt like youre talking but nobo...	culture|sound|speech	0:09:58
...	...	...	...	...	...	...	...
4324	Rashad Robinson	509	How to channel your presence and energy into e...	113821	The presence and visibility of a movement can ...	inequality|race|violence|social change|United ...	0:08:29
4325	Madhumita Murgia	980	How data brokers sell your identity	111967	When tech journalist Madhumita Murgia began re...	data|technology|surveillance|Internet|privacy	0:16:20
4326	Ryan Martin	789	Why some anger can be good for you	111932	Anger researcher Ryan Martin draws from a care...	science|mental health|emotions|psychology|life	0:13:09
4327	Emily Anhalt	651	Why we should all try therapy	111217	We tend to think of therapy as an approach to ...	health|happiness|emotions|mental health|public...	0:10:51
4328	Keith Lowe	1109	Why we need to stop obsessing over World War II	110264	Why are we so obsessed with World War II Histo...	war|history|society|future|global issues	0:18:29

4329 rows × 7 columns

Figure 1 Data Overview

Next, we tokenized the tags column and got the count of each unique tag, then sorted with descending sequence. Figure 2 below shows the top 15 tags. The topics of science, culture and health are far much popular than other topics. The 8th issue tags are global issues for most cases.

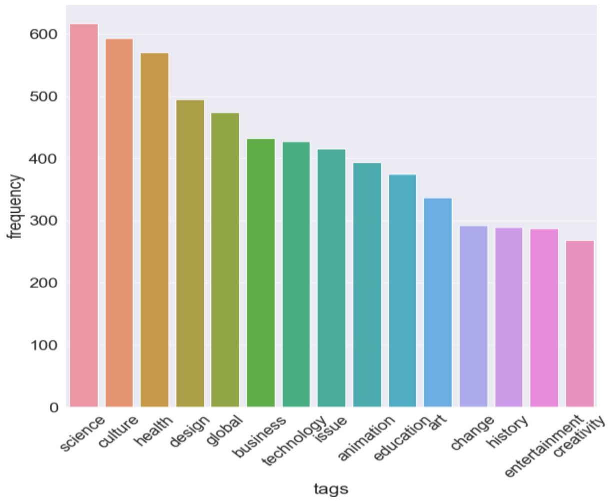
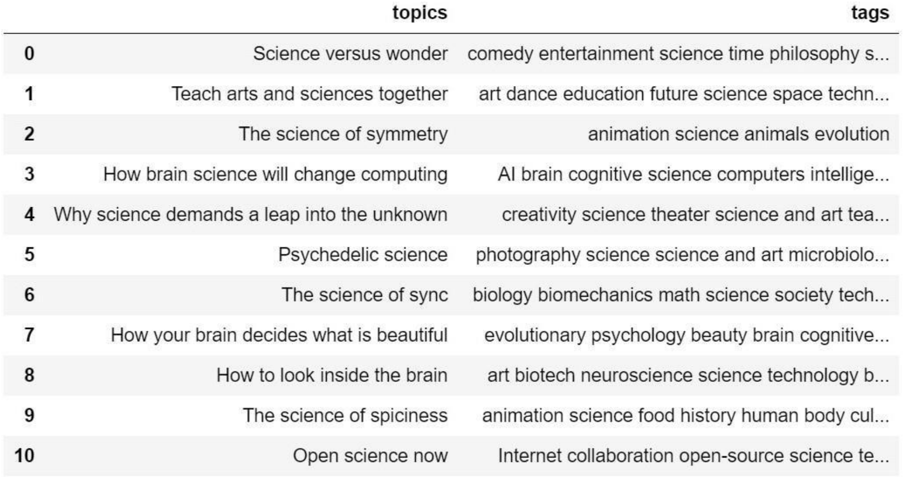


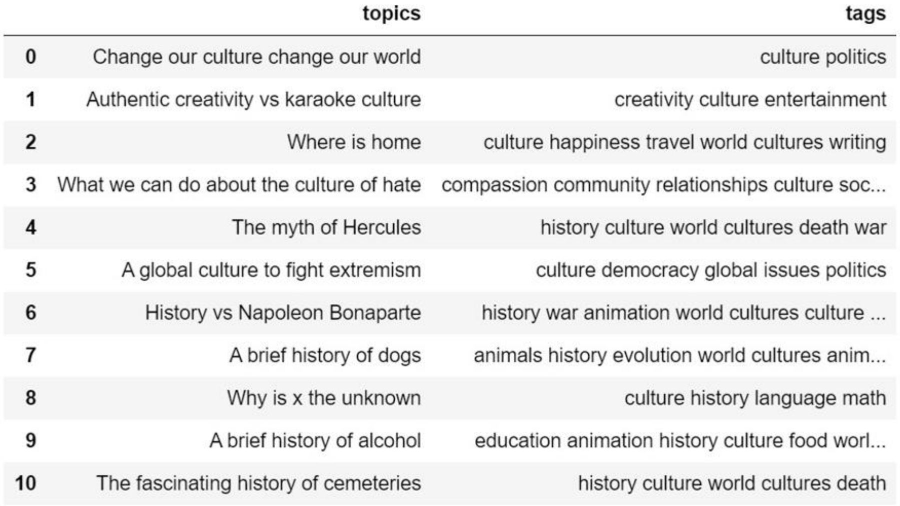
Figure 2 Top 15 Tags

In addition to find top 15 popular tags for TED talks, we also implement tf-idf algorithm to find out top 10 topics that highly related to tags, shown in figure 3, figure 4 and figure 5. We can use them as auto recommender to recommend related topics to tags or use them for deep text mining later.



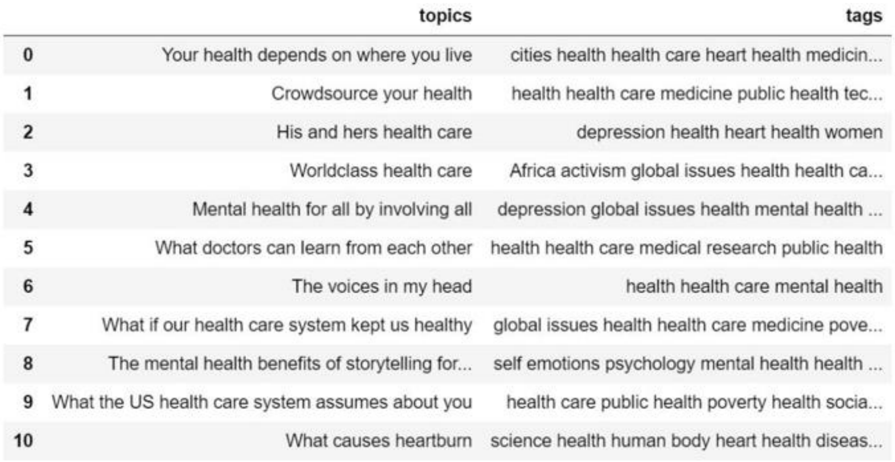
	topics	tags
0	Science versus wonder	comedy entertainment science time philosophy s...
1	Teach arts and sciences together	art dance education future science space techn...
2	The science of symmetry	animation science animals evolution
3	How brain science will change computing	AI brain cognitive science computers intellige...
4	Why science demands a leap into the unknown	creativity science theater science and art tea...
5	Psychedelic science	photography science science and art microbiolo...
6	The science of sync	biology biomechanics math science society tech...
7	How your brain decides what is beautiful	evolutionary psychology beauty brain cognitive...
8	How to look inside the brain	art biotech neuroscience science technology b...
9	The science of spiciness	animation science food history human body cul...
10	Open science now	Internet collaboration open-source science te...

Figure 3 Science Tag Related Topics



	topics	tags
0	Change our culture change our world	culture politics
1	Authentic creativity vs karaoke culture	creativity culture entertainment
2	Where is home	culture happiness travel world cultures writing
3	What we can do about the culture of hate	compassion community relationships culture soc...
4	The myth of Hercules	history culture world cultures death war
5	A global culture to fight extremism	culture democracy global issues politics
6	History vs Napoleon Bonaparte	history war animation world cultures culture ...
7	A brief history of dogs	animals history evolution world cultures anim...
8	Why is x the unknown	culture history language math
9	A brief history of alcohol	education animation history culture food worl...
10	The fascinating history of cemeteries	history culture world cultures death

Figure 4 Culture Tag Related Topics



	topics	tags
0	Your health depends on where you live	cities health health care heart health medicin...
1	Crowdsource your health	health health care medicine public health tec...
2	His and hers health care	depression health heart health women
3	Worldclass health care	Africa activism global issues health health ca...
4	Mental health for all by involving all	depression global issues health mental health ...
5	What doctors can learn from each other	health health care medical research public health
6	The voices in my head	health health care mental health
7	What if our health care system kept us healthy	global issues health health care medicine pove...
8	The mental health benefits of storytelling for...	self emotions psychology mental health health ...
9	What the US health care system assumes about you	health care public health poverty health socia...
10	What causes heartburn	science health human body heart health diseas...

Figure 5 Health Tag Related Topics

Furthermore, we use tokenized topics column to plot word cloud to find most frequent words in topics, shown in figure 6. These words are make, new, life, world, future, good, help, solve, brain, learn, power, art, human, and so on.

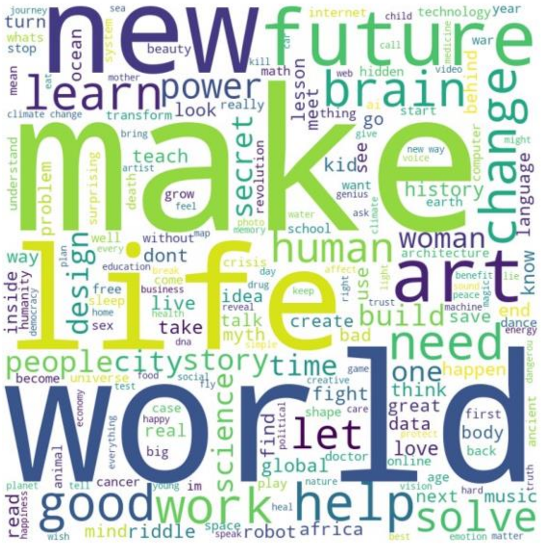


Figure 6 Topic Word Cloud

We sort our data by descending views and extracted the speakers who made the top 5 viewed TED Talks. Figure 7 shows the top 5 speakers. For the top 5 speakers, four of them are English- speakers and only one Bren Brown is French-speaker. English here is the most popular or most frequent language used in Ted Talks.

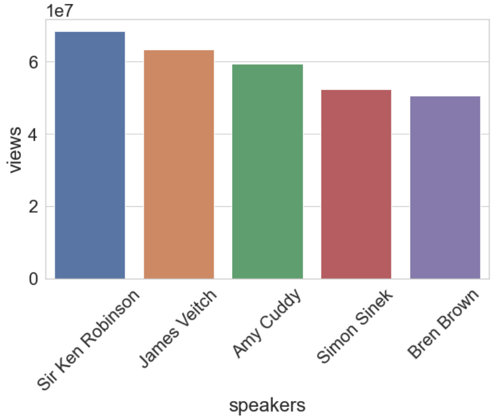


Figure 7 Top 5 Speakers

We try to explore the data by implementing some naive sentiment analysis in both topics and description. We found that:

- count of positive words in topics are 1225
- count of negative words in topics are 1206
- count of positive words in descriptions are 1112
- count of negative words in descriptions are 927

It seems that positive and negative topics are evenly distributed. Then, we want to explore that whether audience preferences positive or negative topics by filtering data into top 500 popular topics. Then, we found that:

- count of positive words in top 500 topics are 195
- count of negative words in top 500 topics are 155

It seems that audience still evenly distributed but slightly trend to preference positive topics. Moreover, we generated a figure to show the correlation between number of views and the number of tags in figure 8. The most viewed ted talks generally have 3 to 9 tags.

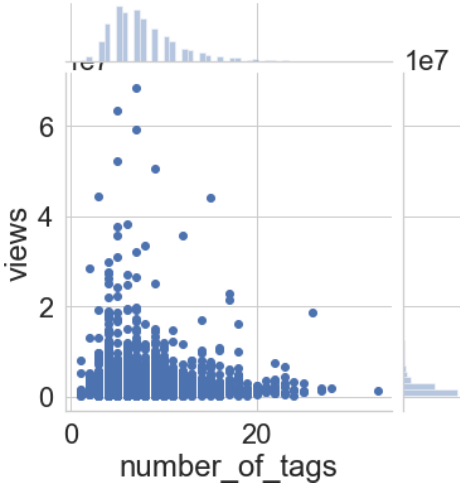


Figure 8 Correlation Between Number of Views and The Number of tags

We generated a figure to show the correlation between number of views and the speech duration (unit: seconds) in figure 9. Most viewed talks have 60 to 2000 seconds of duration; That is, 1 minutes to around 33 minutes are the favorable length of talks.

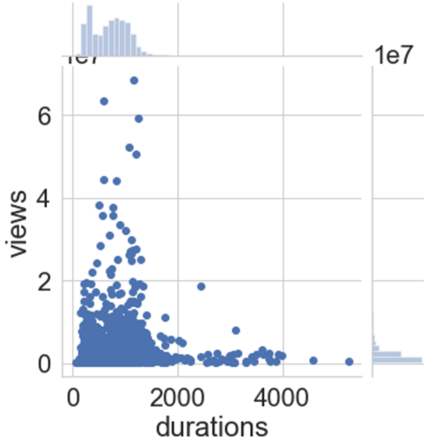


Figure 9 Correlation Between Number of Views and the Speech Durations

3.4 Topic Clustering

By using cosine distance for initializing clustering model, we artificially set 10 numbers of clusters. After getting words with top 20 tf-idf weight in the centroid, we name 10 clusters by the top 20 words in each cluster. The result showed as following:

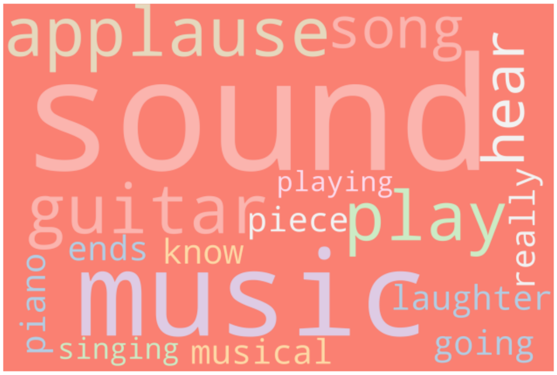


Figure 10 Cluster1: Music

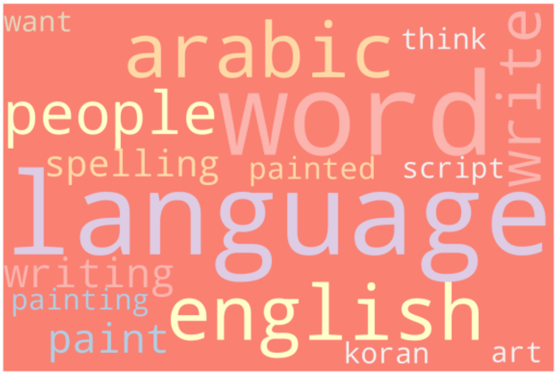


Figure 11 Cluster2: Language

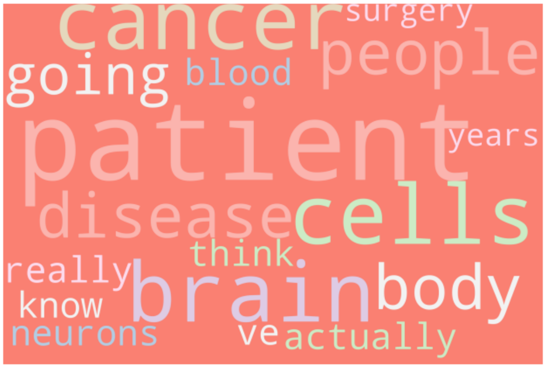


Figure 12 Cluster3: Health

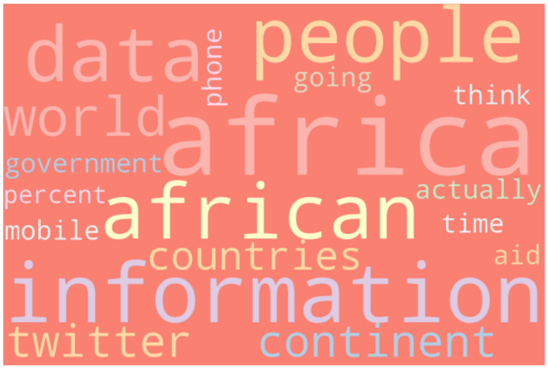


Figure 13 Cluster4: Global

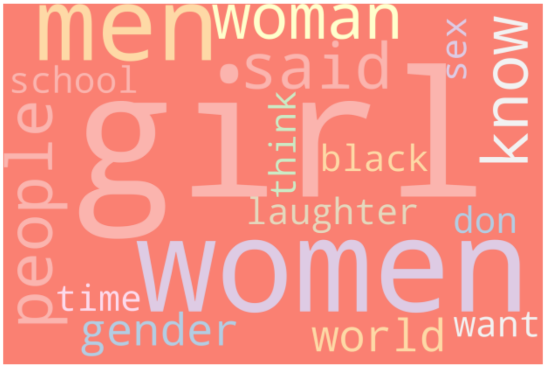


Figure 14 Cluster5: Gender and Ethical issues

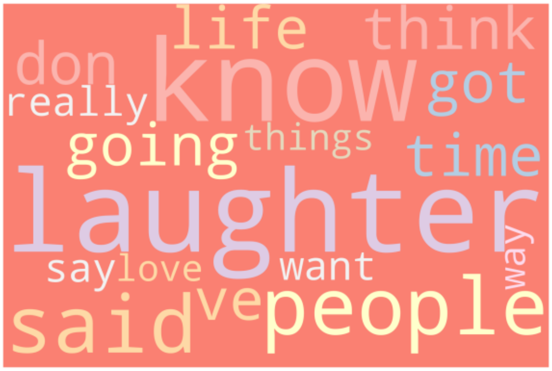


Figure 15 Cluster6: Entertainment

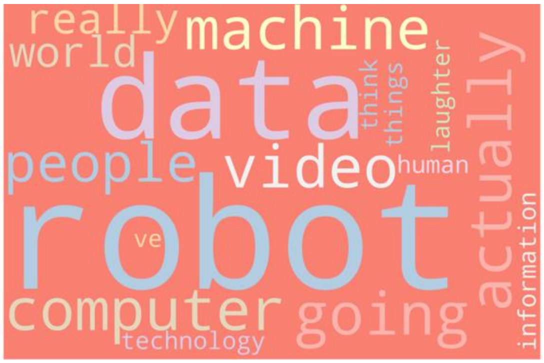


Figure 16 Cluster7: Gender and Ethical issues

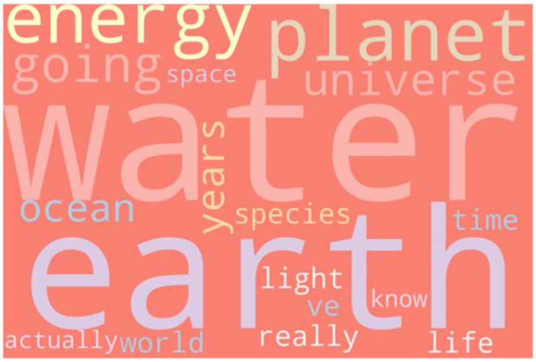


Figure 17 Cluster8: Entertainment

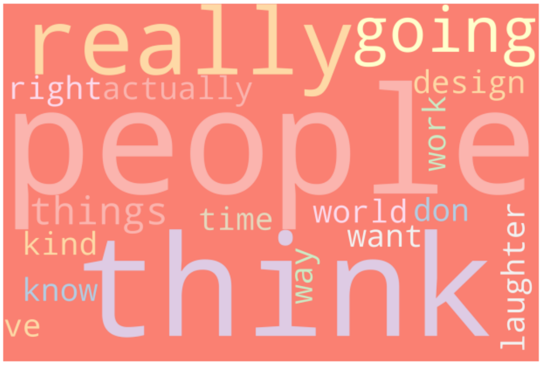


Figure 18 Cluster9: Creativity

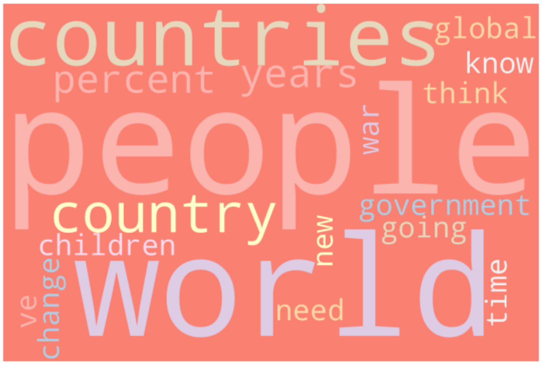


Figure 19 Cluster10: Change

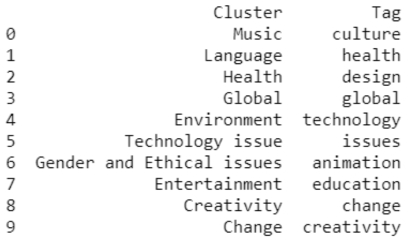

```
                      Cluster         Tag
0                       Music     culture
1                    Language      health
2                      Health      design
3                      Global      global
4                 Environment  technology
5            Technology issue      issues
6  Gender and Ethical issues    animation
7               Entertainment   education
8                  Creativity      change
9                      Change  creativity
```

Figure 20 Comparison between clusters and most popular tags

The clustering is reasonable and successful. The reason it that the ten clusters almost perfectly include the most popular tags of topics. That is, the most frequent words(Top 20) in each cluster are able to represent the majority of the cluster.

4 Experiments

4.1 Sentiment Mining

4.1.1 Analysis with Vader

Firstly, we combined topics column and description column in dataset 1 as one topic+desc column, because we wanted to combine all text information together in each speech, so that we only need to do the analysis once.

By using the SentimentIntensityAnalyzer in nltk.sentiment.vader, it is easy to get sentiment metrices which derived from rating the inputted text. The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). [3]

In our experiment, we extracted only compound score instead of the positive, neutral and negative score, to see the general emotion of each speech.

From figure 21, we can see that the upper half part (compound score > 0) of the figure is denser than the lower half part (compound score < 0). That means positive emotion is more than negative emotion.

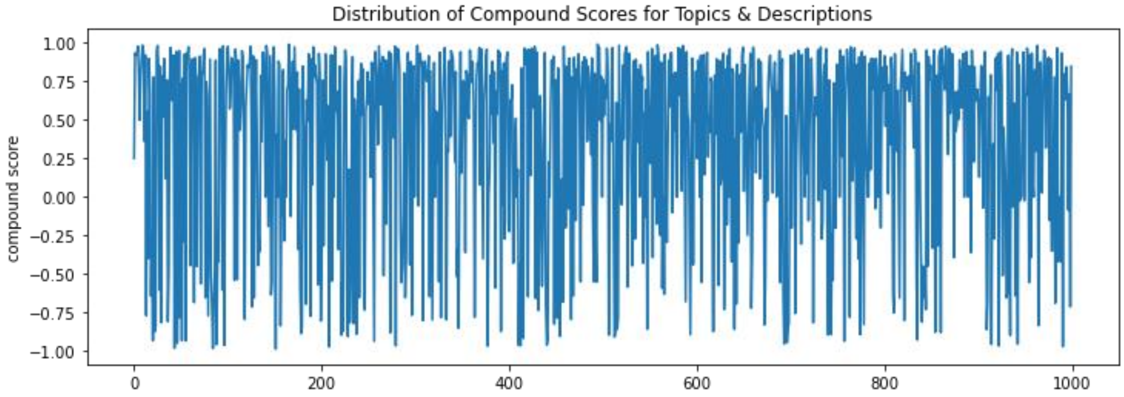


Figure 21 Distribution of Compound Scores for Topics & Descriptions

Except the text in topics and descriptions in each speech, we also detected the emotion in whole transcripts (from dataset 2). Figure 22 shows the compound scores which we got from the transcripts. Similar as the emotion information in figure 21, we found the upper half part (compound score > 0) of figure 22 in denser that the lower half part (compound score < 0), which means positive emotion dominates the TED Talks again.

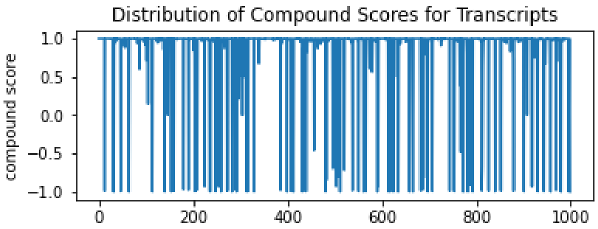


Figure 22 Distribution of Compound Scores for Transcripts

4.1.2 Analysis with Textacy

To validate our finding more accurately, we tried another library in python, which is textacy, to detect the sentiment in TED Talks.

Textacy is a Python library for performing higher-level natural language processing (NLP) tasks, built on the high-performance spaCy library. [4] It can detect DepecheMood (AFRAID, AMUSED, ANGRY, ANNOYED, DONT_CARE, HAPPY, INSPIRED, SAD) and directly give scores of the moods as well, DepecheMood is a high-quality and high-coverage emotion lexicon for English and Italian text, mapping individual terms to their emotional valences. [5]

Firstly, we divided each speech into 10 sections, and detected the emotion from each single section, to see if some specific emotion changing patterns exist or not, for example a sad story in the beginning of a speech but then happy ending at the end of the speech. Table 1 shows an example of mood scores of 10 sections from a randomly selected speech. And we also generated a plot for these mood scores, as shown in figure 23.

Table 1 Emotion Scores for a Randomly Selected Speech

	AFRAID	AMUSED	ANGRY	ANNOYED	DONT_CARE	HAPPY	INSPIRED	SAD
0	0.090804	0.121346	0.120167	0.139926	0.135001	0.098103	0.203327	0.091326
1	0.094469	0.121029	0.119500	0.141718	0.139080	0.102911	0.183794	0.097500
2	0.095660	0.126724	0.119156	0.141931	0.132503	0.101892	0.184496	0.097638
3	0.095897	0.126303	0.115016	0.139041	0.133054	0.104111	0.188234	0.098343
4	0.099263	0.125058	0.112713	0.137216	0.130785	0.104046	0.188595	0.102325
5	0.094048	0.126136	0.109671	0.134816	0.132176	0.105739	0.194814	0.102600
6	0.089429	0.129288	0.105942	0.130541	0.130896	0.113984	0.195153	0.104768
7	0.092766	0.128346	0.107435	0.129965	0.130827	0.112192	0.195167	0.103302

8	0.092585	0.126686	0.108613	0.130985	0.130204	0.110905	0.196035	0.103986
9	0.091848	0.125368	0.109792	0.131926	0.129793	0.111189	0.195803	0.104282

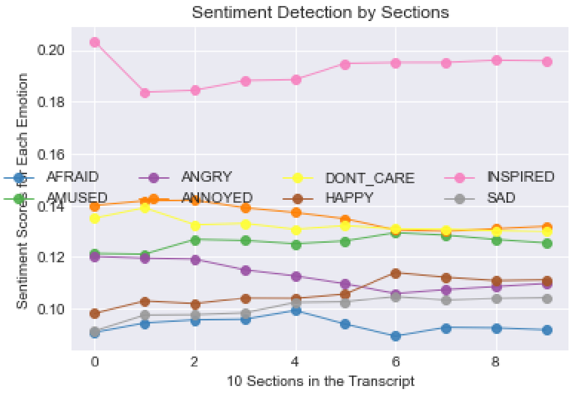


Figure 23 Line Plot for Mood Scores for a Randomly Selected Speech

To get a universally applicable pattern or rule, we used this same method to try more speeches, to see if the mood score sequence in figure 23 is a typical pattern or not.

Next, figure 24 shows the mood scores for 12 more randomly selected speeches. INSPIRED (pink) is usually the mood with highest score in nearly all speeches. DONT_CARE (yellow), ANNOYED (orange) and AMUSED (green) are moods with relative high scores. ANGRY (purple) and HAPPY (brown) are mood with relative low scores. Finally, AFRAID (blue) and SAD (grey) are moods with lowest scores.

According to our experiments, the above pattern is not just applied to the twelve speeches tested in figure 4, but also applied to the majority of the other speeches which are not shown in figure 4 due to space limit.

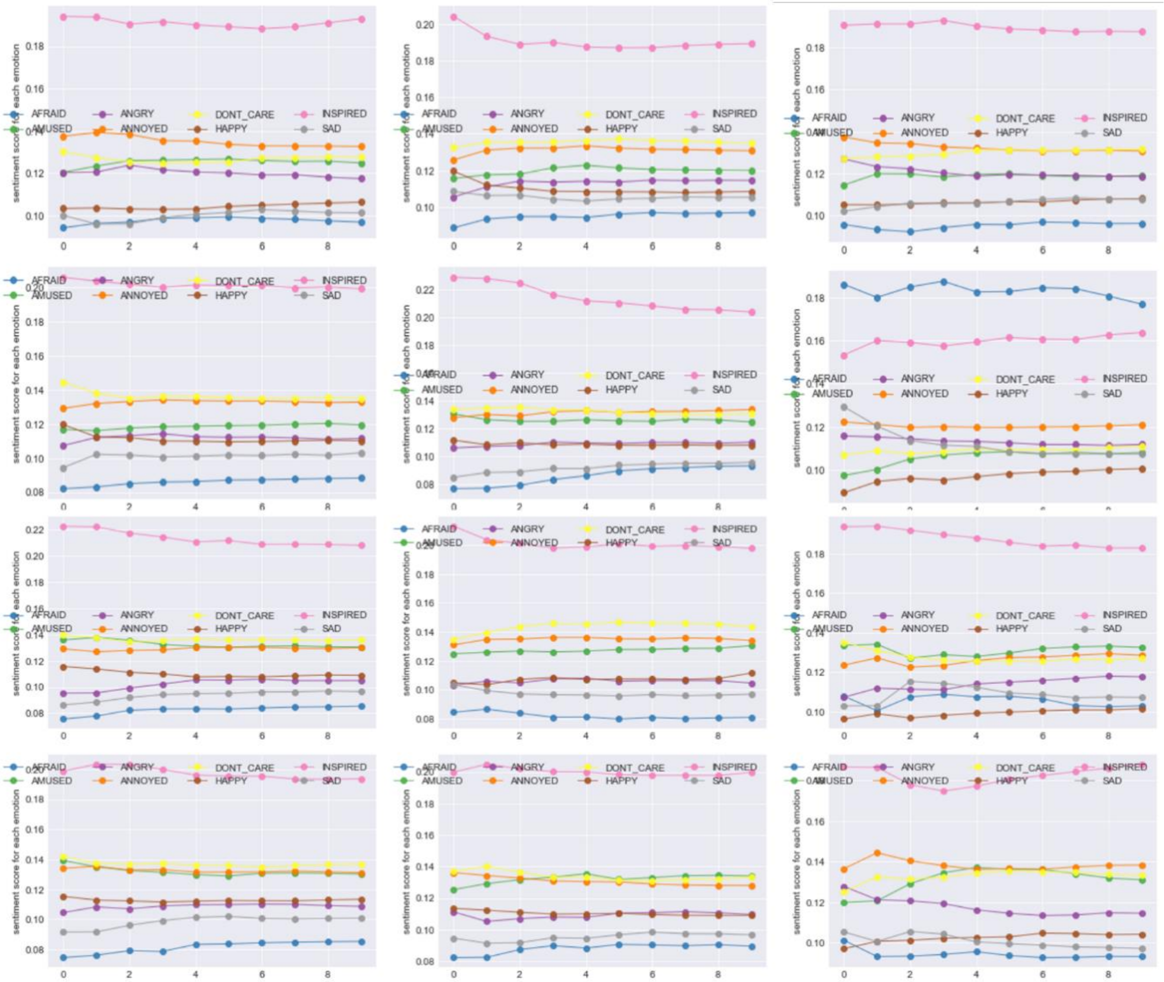


Figure 24 Line Plot for Mood Scores for more Randomly Selected Speeches

But the problem with emotion detection by sections is that the sections are evenly separated from each speech. Sometimes, a change of emotion might just be separated by the section dividing. So, we also tried to detect the emotion sentence by sentence in each speech, using the same algorithm as above.

We tokenized each speech into single sentences, and then got the mood scores for each sentence and plotted the mood scores as well. Moreover, we also used exponential smoothing method to make the lines in the plots look smoother.

Figure 25 shows the emotion score changes sentence by sentence in nine randomly selected speeches. From these plots we can still see INSPIRED (pink), which is positive emotion, is almost always the highest, AFRAID (blue) and SAD (grey), which are negative emotion, are almost always the lowest.

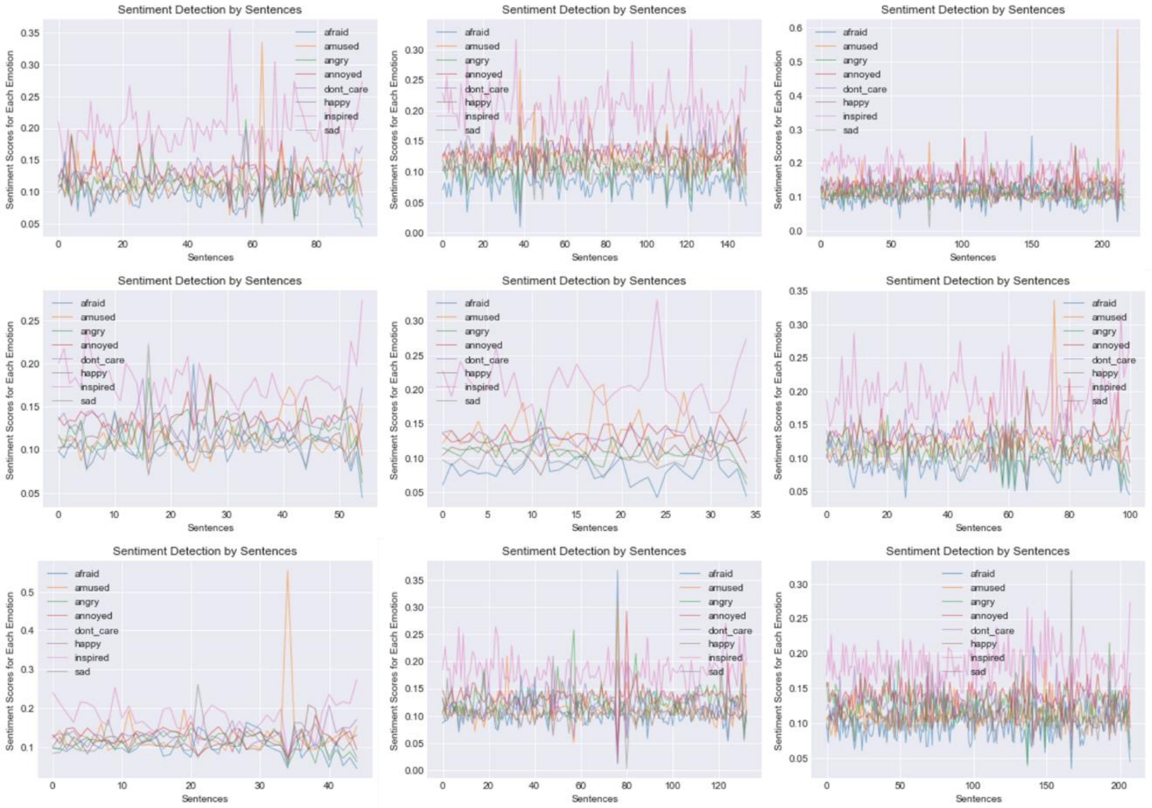


Figure 25 Line Plot for Mood Scores for more Randomly Selected Speeches

So, the sentiment detection with textacy again indicates that TED Talks have much more positive emotion than negative emotion.

However, this pattern does not show some change of emotion, only shows an overall extent of each emotion in every speech. In order to detect some emotion changes in speeches, we tried to explore more sentiment score distributions.

With our exploration, we found that many of the speeches have spikes of AMUSED (orange) and AFRAID (blue), which shows in figure 26. But these spikes of AMUSED and AFRAID are not related to specific topics, just to generic speeches in our dataset.

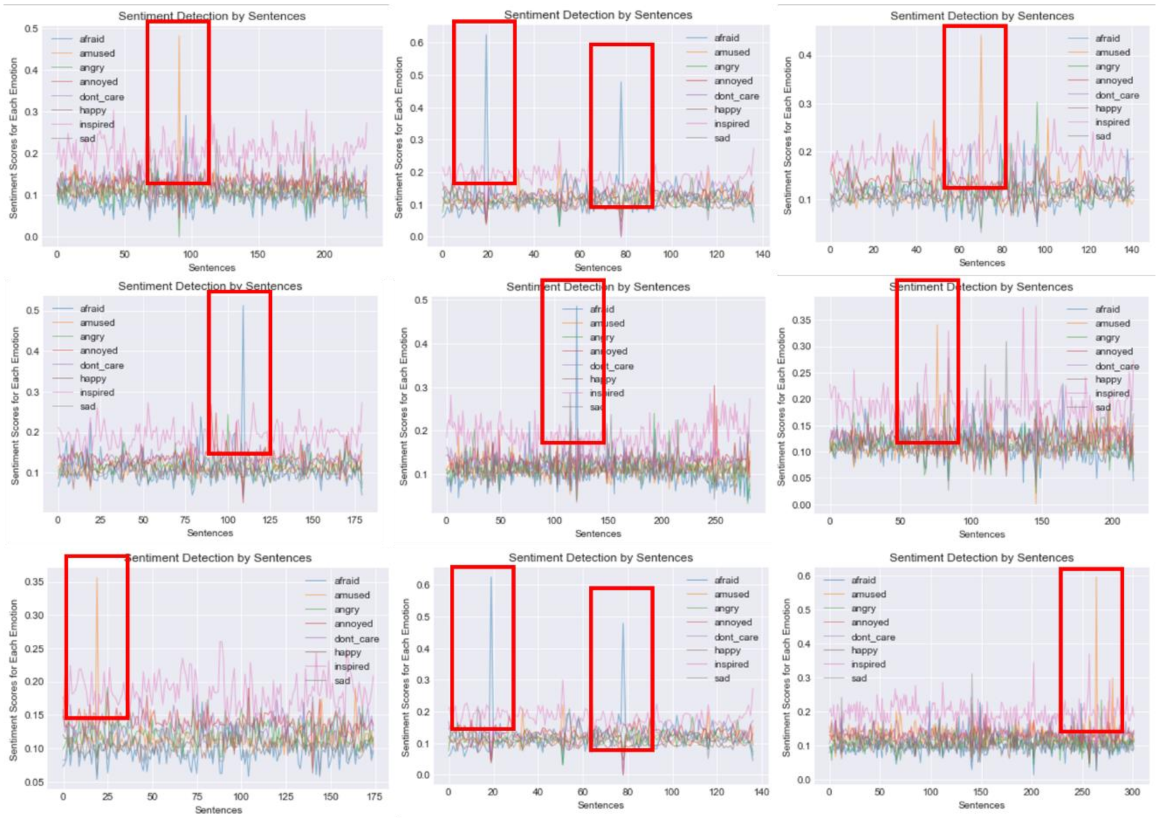
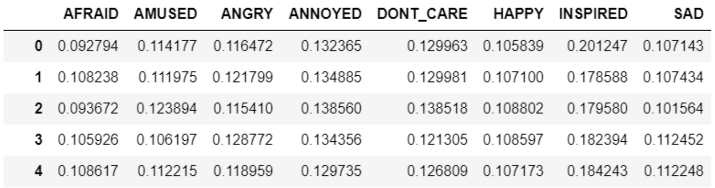


Figure 26 Emotion Score Distributions with Specific Spikes

4.2 Sentiment Clustering Analysis

In order to seek more patterns among sentiment, topics, popularity, and so on, we generate a sentiment score dataset for sentiment clustering analysis.

The dataset in figure 27 consists of 8 sentiment scores. Each row represents each talk, and each column represents each emotion. Unlike topic clustering, sentiment clustering cannot use tfidf vectorizer to look for the vector for each talk. Accordingly, the sequential pattern of sentiment score for each talk is the better way to calculate the pair wise distance. Basically, the idea is that we consider the sequence of sentiment scores as a pattern we rely on to find the pair wise distance for each talk. Thus, the package fastdtw is introduced to implement the idea of sequential pattern pair wise distance. The package takes two series of data as input and use Euclidean distance to find distance between patterns.



	AFRAID	AMUSED	ANGRY	ANNOYED	DONT_CARE	HAPPY	INSPIRED	SAD
0	0.092794	0.114177	0.116472	0.132365	0.129963	0.105839	0.201247	0.107143
1	0.108238	0.111975	0.121799	0.134885	0.129981	0.107100	0.178588	0.107434
2	0.093672	0.123894	0.115410	0.138560	0.138518	0.108802	0.179580	0.101564
3	0.105926	0.106197	0.128772	0.134356	0.121305	0.108597	0.182394	0.112452
4	0.108617	0.112215	0.118959	0.129735	0.126809	0.107173	0.184243	0.112248

Figure 27 sentiment score dataset

Once we found the distance for each talk, we can reshape the distance to better fit in the clustering model. This project uses Gaussian Mixture Models to fit in the calculated dtm. This model uses the lowest BIC score to find out that the best GMM model is 9 components with spherical covariance type. Eventually, this model uses manifold embedding method to visualize the sentiment clustering model.

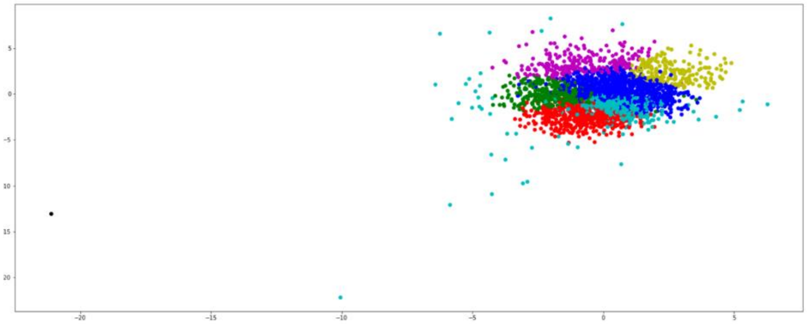


Figure 28 sentiment clustering

After combining the sentiment clustering result with other data, we gathered some interesting results.

Figure 29 shows that the most popular sentiment patterns fall in with cluster 3, the second is cluster 2, and third is cluster 8.

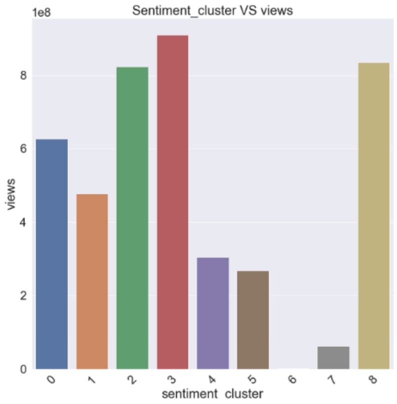


Figure 29 sum of views for each sentiment cluster

The figure 28, 29 and 30 together demonstrate that cluster 6 is an anomaly talk or outlier. This cluster only contains one talk, it is because that the "don't care" score is so high that makes the distance of this talk far from other talks.

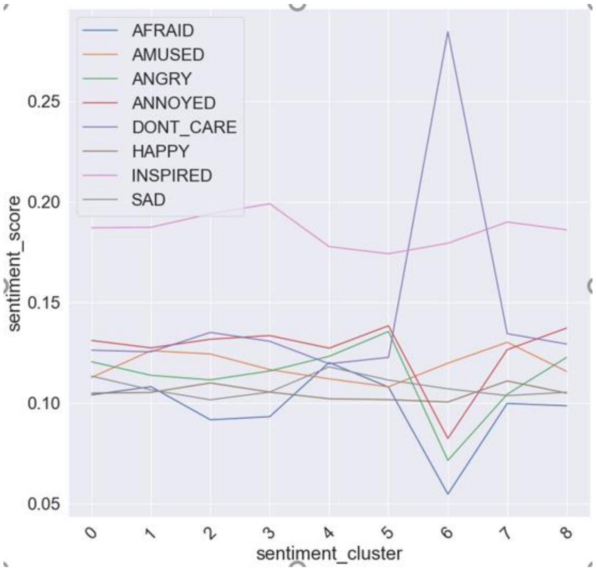


Figure 30 sum of views for each sentiment cluster

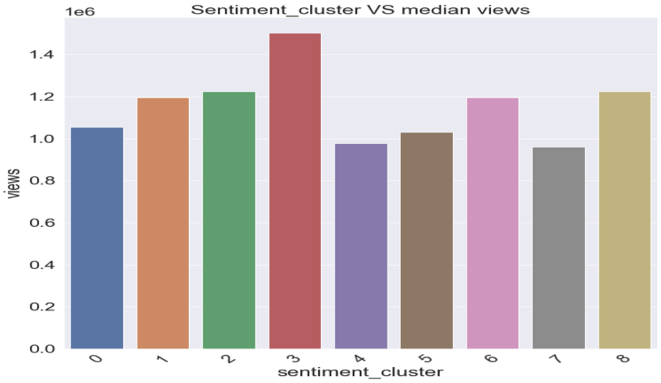


Figure 31 sentiment cluster with its average sentiment score

Though cluster 3 still has the highest views in median views for each cluster graph, other clusters are relatively close with the median views. It implies that cluster 0, 2, and 8 contain the extremely popular talks that highly influenced the sum of views.

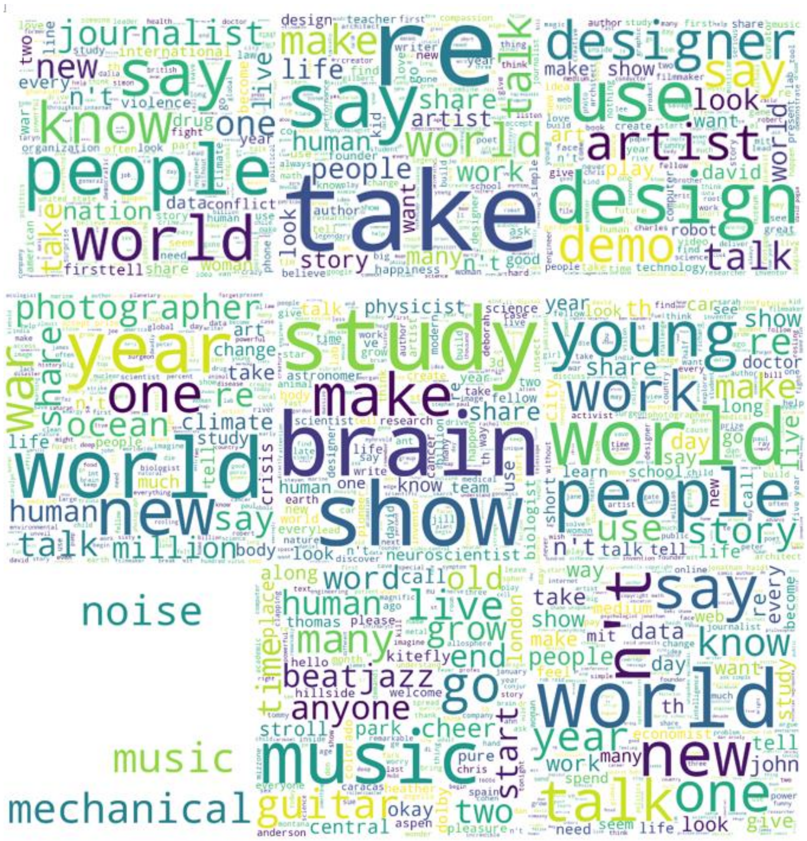


Figure 32 Word cloud of each sentiment cluster

The above word clouds are word cloud of each sentiment clusters by order. We can see that the topic is related to people, youth, story, and world share similar sentiment patterns. Study, brain, and research share similar sentiment patterns. Design and artist share similar sentiment patterns. Topics related to human, people, life, work, and story share similar sentiment patterns and have the highest views. What draws my interests is that the word cloud in cluster 4 shows that the topics of photographer, ocean, and war share similar sentiment patterns. By doing this, we can find many interesting results about the relationship between some topics and sentiment patterns.

4.3 Logistic Regression

Problem Statement Understanding the problem statement is the first and foremost step. This would help you give an intuition of what you will face ahead of time. Let us see the problem statement -

The objective of this task is to detect if the ted talk is popular. For the sake of simplicity, we say how emotion can affect the popularity of Ted Talk. So, the task is to find the relationship between the emotion and the popularity. And I did a contrast. Because I thought the number of comments is also a display for the people emotion. The more emotion shows that people have more strong emotion for the Ted Talk. The first LR is with comments and the second one is without comments

Formally, given a training sample of tweets and labels, the higher of views (times that the video had been watched) shows that the video is more popular. I used the mean value of views to decide whether the ted talk is popular. If the views are higher than the mean value represent that the ted talk is popular, and the value will return to 1, otherwise, return to 0 your objective is to predict the labels on the test data.

Thanks to my teammates, they have already done the Text PreProcessing and Cleaning. I just use a few columns from the data and build the model.

4.3.1 Model 1: Data with comments

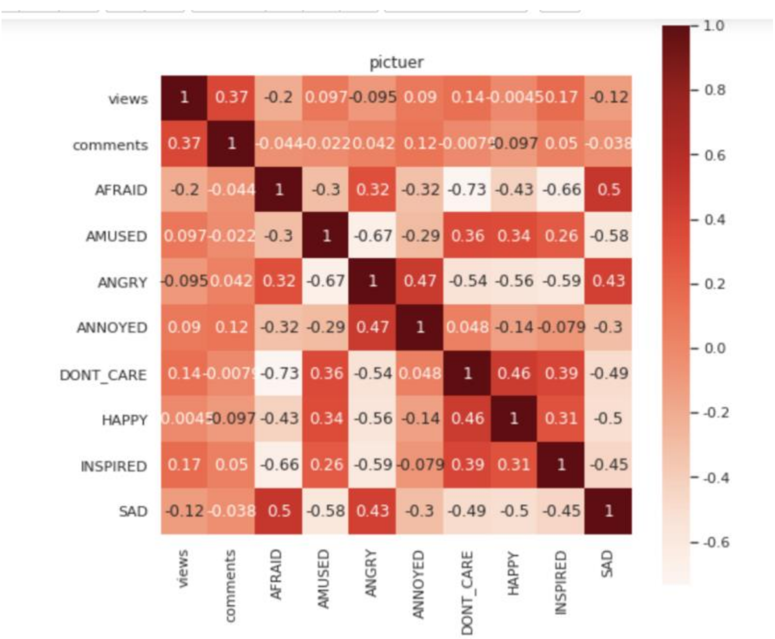
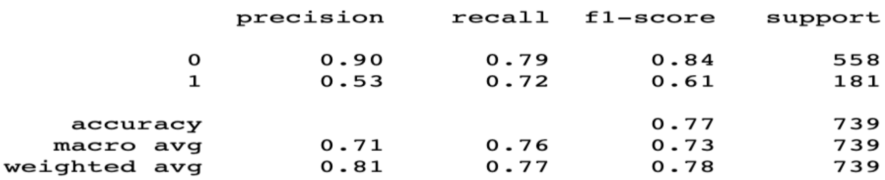


Figure 33



```
              precision    recall  f1-score   support

           0       0.90      0.79      0.84       558
           1       0.53      0.72      0.61       181

    accuracy                           0.77       739
   macro avg       0.71      0.76      0.73       739
weighted avg       0.81      0.77      0.78       739
```

Figure 34

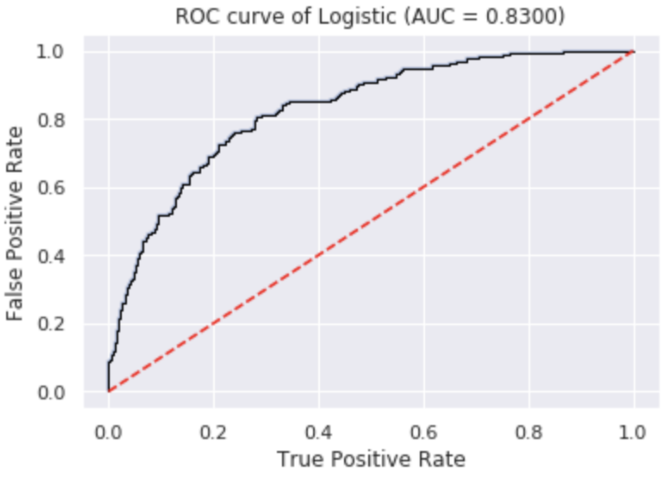
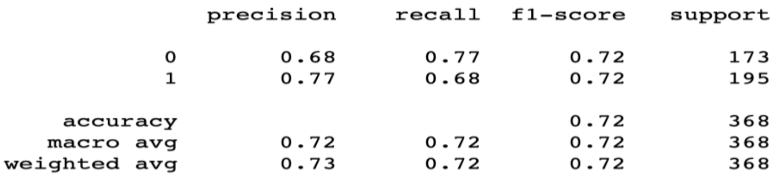


Figure 35



```
              precision    recall  f1-score   support

           0       0.68      0.77      0.72       173
           1       0.77      0.68      0.72       195

    accuracy                           0.72       368
   macro avg       0.72      0.72      0.72       368
weighted avg       0.73      0.72      0.72       368
```

Figure 36

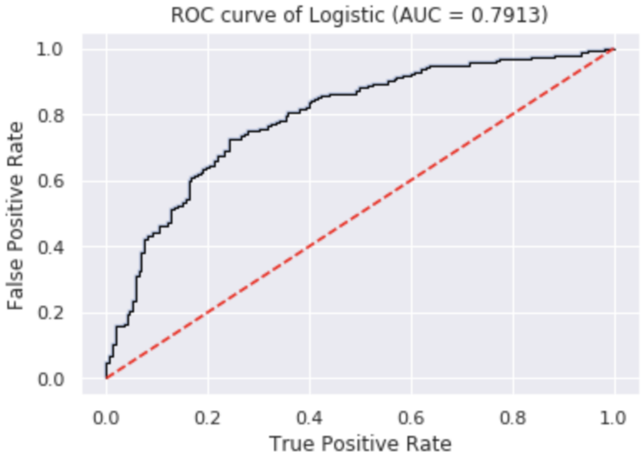


Figure 37

The two conclusions of AUC and precision is a little waved. The result is quite similar. However, the second result is after down sampling. For the first result，The AUC is a little higher but the precision for value 1 is not accurate. So, I add a down sampling algorithm to make the value of 1 and 0 balanced. After the algorithm, the accuracy for value 1 improved a lot. And the AUC is just a little bit lower. Then I use the same methods in model 2: data without comments.

4.3.2 Model 2: Data without comments

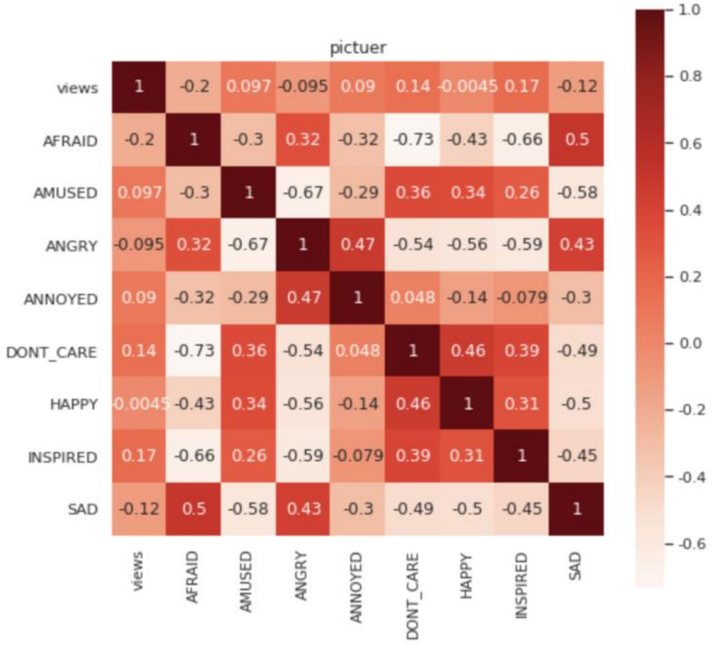
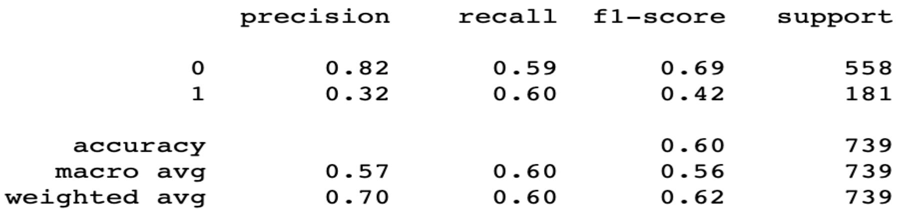


Figure 38



```
              precision    recall  f1-score   support

           0       0.82      0.59      0.69       558
           1       0.32      0.60      0.42       181

    accuracy                           0.60       739
   macro avg       0.57      0.60      0.56       739
weighted avg       0.70      0.60      0.62       739
```

Figure 39

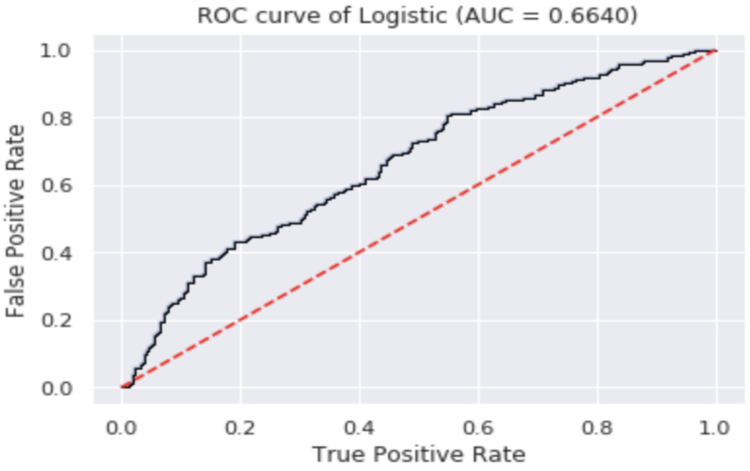
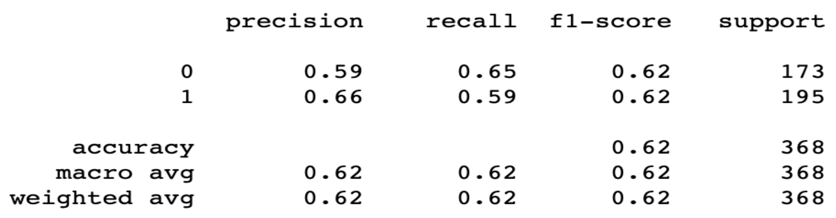


Figure 40

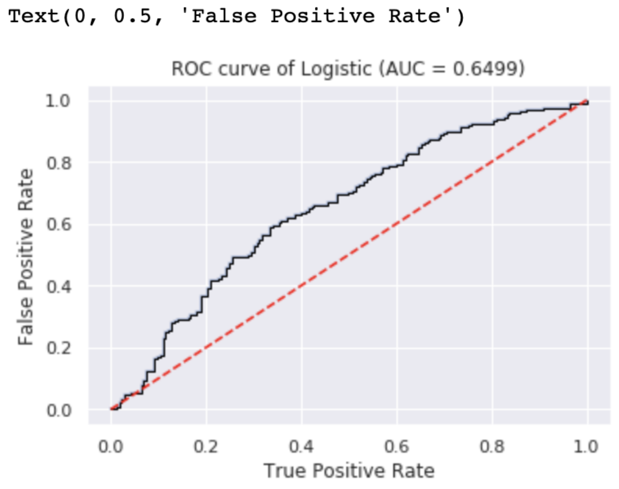


```
              precision    recall  f1-score   support

           0       0.59      0.65      0.62       173
           1       0.66      0.59      0.62       195

    accuracy                           0.62       368
   macro avg       0.62      0.62      0.62       368
weighted avg       0.62      0.62      0.62       368
```

Figure 41



```
Text(0, 0.5, 'False Positive Rate')
```

Figure 42

From both models above we can see that positive emotion like happy, inspired really help the popularity of Ted talk, and the negative emotion like sad and afraid are negative relate to the popularity. With the comments, the model will be more accurate. So I believe with the comments number will better to predict the popularity of the ted talks.

5 Conclusion

Tags does affect the views, the most popular tags are science, culture, health, design, global, etc. In some words, speak duration affects the popularity: 1 minute to around 33 minutes are the favorable length of talks. The top 5 popular speakers are Sire Ken Robinson, James Veitch, Amy Cuddy Simon Sinek and Bren Brown.

As for the conclusion about sentiment mining, we can conclude that usually positive sentiment is more than negative sentiment. More specifically, emotion of INSPIRED usually dominates most speeches, emotion of AFRAID and SAD are the least emotion I speeches. Moreover, among many of the speeches, there are some AMUSED and SAD emotions, which could be speakers' tricks to attract audience's attention.

Sentiment clustering is a powerful tool to gather many interesting insights about the relationship between topics, popularity, and sentiment patterns. We learned that Topics related to human, people, life, work, and story share similar sentiment patterns and belong to the most popular cluster. We also can get some interesting result such as topics related to photographer, and war virtually share similar sentiment patterns.

From the models above we can see that the positive emotion in the Ted Talk like happy, inspired will definitely improve the popularity of the Ted Talk. And the negative emotion like sad afraid is negative relate to the popularity. The number of comments will also improve the popularity.

6 Work Assignment

Task	Assignee	Signature
Data Collection	Zhizheng Li, Yuhui Ren	Yuhui Ren Zhizheng Li
Data Preprocessing	Xin Zhao, Ruixi Wang	Xin Zhao Ruixi Wang
Exploratory Data Analysis	Ruixi Wang, Xin Zhao	Xin Zhao Ruixi Wang
Topic Clustering	Ruixi Wang	Ruixi Wang
Mid-Term Report	Yuhui Ren, Zhizheng Li	Yuhui Ren Zhizheng Li
Sentiment Mining	Yuhui Ren	Yuhui Ren
Sentiment Cluster	Xin Zhao	Xin Zhao
Logistic Regression	Zhizheng Li	Zhizheng Li
Conclusion	All	Yuhui Ren Zhizheng Li Xin Zhao Ruixi Wang
Research Report Writing	All	Yuhui Ren Zhizheng Li Xin Zhao Ruixi Wang

Reference

[1] https://www.ted.com/about/our-organization

[2] https://thebiz.bentley.edu/9-reasons-ted-talks-are-so-popular/

[3] https://blog.quantinsti.com/vader-sentiment/

[4] https://pypi.org/project/textacy/

[5] https://textacy.readthedocs.io/en/stable/api_reference/resources.html