What Makes a Good TED Talk?

Executive Summary

TED is a nonprofit devoted to spreading ideas, usually in the form of short, powerful motivational talks. Ted speakers are very good at giving such talks. Ever individual who has ever attended or watched a TED talk truly admires them. This study explains the reasons why TED talks are so addictive to watch and how a person can give a good motivational talk. In this project, we studied 2500 talks from all the TED events to understand what and why TED talks are so highly viewed and rated when it comes to motivational talks. The difference between a good TED talk and an exceptional TED talk is also one of the key findings of this study. Key methods include sentiment analysis, text mining, topic modelling to answer questions pertaining to why they are so special.

Introduction

Every individual in their lives look for motivation to achieve their goals and ambitions. In today's fast paced world where time is a commodity as precious as gold, we do not have the time and resources to look for motivation. Here is where TED talks have helped individuals shape their lives and gain that much needed insightful motivation.

How many times have we heard someone give a talk, and walked away asking ourselves, "What was he or she even talking about?". This is the cardinal sin of speeches. Even if the audience hasn't had to pay money to attend a talk, they're giving something potentially more valuable: their time. Hence it is imperative for the speaker to not only have a proper structure to but also have a good storytelling approach. Ted speakers are very good at giving such talks.

This project focuses on deducing the elemental reason why TED talks are considered the benchmark for influential speeches.

Sentiment Analysis

Sentiment analysis is the process of distilling actionable insights from text. The essential components of a sentiment analysis project include problem description, text identification, text organization, feature extraction, analysis and conclusion. Sentiment Analysis in R can be performed by many ways such as Semantic Parsing and Bag of Words. In Semantic parsing we care about the word type and order. A feature rich and tree structure tagging is done. Whereas in

Bag of words each word is treated as single token without considering type or order. This project utilizes several such ways to analyze TED talks.

Data Cleaning and Preparation

Our data includes 2500+ Ted talks obtained from Kaggle, and there are 2 separate charts in the dataset. First one, Ted Main, includes attributes of the number of comments, lecturer, themes, ratings, durations, etc. The second table, Transcripts, includes the full transcript of all 2500+ ted talks.

comments	description	duration	event	film_date	languages	main_speake	name	num_speake	published_da	ratings	related_talks	speaker_occ	tags	title	url	views
4553	Sir Ken Robir	1164	TED2006	1140825600	60	Ken Robinso	Ken Robinso	1	1151367060	[{'id': 7, 'nam	[{'id': 865, 'h	Author/educ	['children', 'c	Do schools ki	https://ww	47227110
265	With the sam	977	TED2006	1140825600	43	Al Gore	Al Gore: Ave	1	1151367060	[{'id': 7, 'nam	[{'id': 243, 'h	Climate advc	['alternative	Averting the	https://ww	3200520
124	New York Tir	1286	TED2006	1140739200	26	David Pogue	David Pogue:	1	1151367060	[{'id': 7, 'nam	[{'id': 1725, 'l	Technology c	['computers'	Simplicity sel	https://ww	1636292
200	In an emotio	1116	TED2006	1140912000	35	Majora Carte	Majora Carte	1	1151367060	[{'id': 3, 'nam	[{'id': 1041, 'l	Activist for e	['MacArthur	Greening the	https://ww	1697550
593	You've never	1190	TED2006	1140566400	48	Hans Rosling	Hans Rosling	1	1151440680	[{'id': 9, 'nam	[{'id': 2056, 'l	Global health	['Africa', 'Asi	The best stat	https://ww	12005869
672	Tony Robbin:	1305	TED2006	1138838400	36	Tony Robbin:	Tony Robbin:	1	1151440680	[{'id': 7, 'nam	[{'id': 229, 'h	Life coach; e:	['business', 'c	Why we do v	https://ww	20685401
919	When two yo	992	TED2006	1140739200	31	Julia Sweene	Julia Sweene	1	1152490260	[{'id': 3, 'nam	[{'id': 22, 'he	Actor, comec	['Christianity	Letting go of	https://ww	3769987
46	Architect Jos	1198	TED2006	1140652800	19	Joshua Princ	Joshua Princ	1	1152490260	[{'id': 9, 'nam	[{'id': 750, 'h	Architect	['architecture	Behind the d	https://ww	967741
852	Philosopher	1485	TED2006	1138838400	32	Dan Dennett	Dan Dennett	1	1153181460	[{'id': 3, 'nam	[{'id': 71, 'he	Philosopher,	['God', 'TED E	Let's teach re	https://ww	2567958
900	Pastor Rick V	1262	TED2006	1140825600	31	Rick Warren	Rick Warren:	1	1153181460	[{'id': 21, 'na	[{'id': 94, 'he	Pastor, autho	['Christianity	A life of purp	https://ww	3095993
79	Accepting his	1414	TED2006	1140912000	27	Cameron Sin	Cameron Sin	1	1153786260	[{'id': 3, 'nam	[{'id': 1749, 'l	Co-founder, .	['activism', 'a	My wish: A c	https://ww	1211416
55	Jehane Nouj:	1538	TED2006	1140912000	20	Jehane Nouj:	Jehane Nouj:	1	1153786260	[{'id': 1, 'nam	[{'id': 2228, 'l	Filmmaker	['TED Prize', '	My wish: A g	https://ww	387877
71	Accepting th	1550	TED2006	1140652800	24	Larry Brillian	Larry Brillian	1	1153786260	[{'id': 8, 'nam	[{'id': 1153, 'l	Epidemiologi	['TED Prize', '	My wish: Hel	https://ww	693341
242	Jeff Han shov	527	TED2006	1139184000	27	Jeff Han	Jeff Han: The	1	1154391060	[{'id': 9, 'nam	[{'id': 685, 'h	Human-comp	['demo', 'des	The radical p	https://ww	4531020
99	Nicholas Neg	1057	TED2006	1140652800	25	Nicholas Neg	Nicholas Neg	1	1154391060	[{'id': 3, 'nam	[{'id': 2043, 'l	Tech visionar	['children', 'd	One Laptop p	https://ww	358304
325	Violinist Sire	1481	TED2006	1140652800	31	Sirena Huang	Sirena Huang	1	1154995860	[{'id': 1, 'nam	[{'id': 2273, 'l	Violinist	['entertainm	An 11-year-o	https://ww	2702470
305	Pianist and c	1445	TED2004	1077753600	32	Jennifer Lin	Jennifer Lin:	1	1154995860	[{'id': 1, 'nam	[{'id': 2273, 'l	Pianist, comp	['creativity', '	Improvising c	https://ww	1628912

Table 1. Ted Main

transcript	url
Good morning. How are you?(Laughter)It's been great, hasn't it? I've been blown away by the whole thing. In fact, I'm leaving.(Laughter)There have been three themes	https://ww
Thank you so much, Chris. And it's truly a great honor to have the opportunity to come to this stage twice; I'm extremely grateful. I have been blown away by this confer	https://ww
(Music: "The Sound of Silence," Simon & Garfunkel)Hello voice mail, my old friend.(Laughter)I've called for tech support again. I ignored my boss's warning. I called on a	https://ww
If you're here today â€" and I'm very happy that you are â€" you've all heard about how sustainable development will save us from ourselves. However, when we're not	https://ww
About 10 years ago, I took on the task to teach global development to Swedish undergraduate students. That was after having spent about 20 years together with Africa	https://ww
Thank you. I have to tell you I'm both challenged and excited. My excitement is: I get a chance to give something back. My challenge is: the shortest seminar I usually do	https://ww
On September 10, the morning of my seventh birthday, I came downstairs to the kitchen, where my mother was washing the dishes and my father was reading the pape	https://ww
I'm going to present three projects in rapid fire. I don't have much time to do it. And I want to reinforce three ideas with that rapid-fire presentation.The first is what I lik	https://ww
It's wonderful to be back. I love this wonderful gathering. And you must be wondering, "What on earth? Have they put up the wrong slide?" No, no. Look at this magnifi	https://ww

Table 2. Transcripts

Detailed information of the attributes are listed as followed.

comments	The number of first level comments made on the talk
description	A blurb of what the talk is about
duration	The duration of the talk in seconds
event	The TED/TEDx event where the talk took place
film_date	The Unix timestamp of the filming
languages	The number of languages in which the talk is available

main_speaker	The first named speaker of the talk
name	The official name of the TED Talk. Includes the title and the speaker.
num_speaker	The number of speakers in the talk
published_date	The Unix timestamp for the publication of the talk on TED.com
ratings	A thingified 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

Table 3. Data columns information

Following steps were taken to clean the dataset, after having a brief knowledge of the dataset. First step was to check and delete the missing values, alter the date, duration and views into a viewable form. Then the dataset was arranged in the order of views and the top 500 and bottom 500 were selected.
These selected transcripts were cleaned first using qdap functions. They included removal of the numbers, symbols, whitespaces and punctuations. Lastly a Document Term Matrix (DTM) was created for top and bottom ted talks. By breaking the transcript data in string form into tokens, a weighted DTM was created by calculating the term frequency and inverse document frequency.

TED events

Speaker and themes

Before we investigate how TED speakers give a speech, it is important to understand their background. From figure 1 we can see the most appearance of speakers by occupation. From figure 2 we can see the most popular topics. Interestingly, it can be observed that while technology is the most popular theme of the TED talks, writers are more dominant when it came to professions. Hence after looking into this interesting analysis in more detail, we found out that the professions of the speakers mentioned were not necessarily their education. For example, an engineer could also be writer.

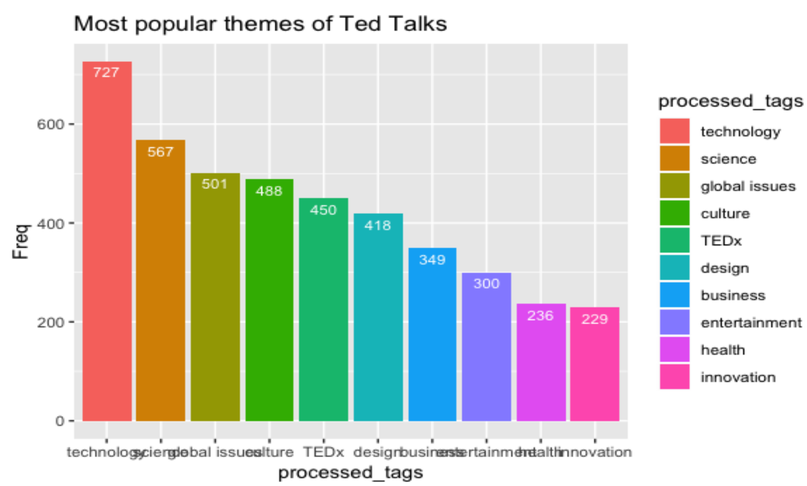

Figure 1. Most Popular Themes of Ted Talks

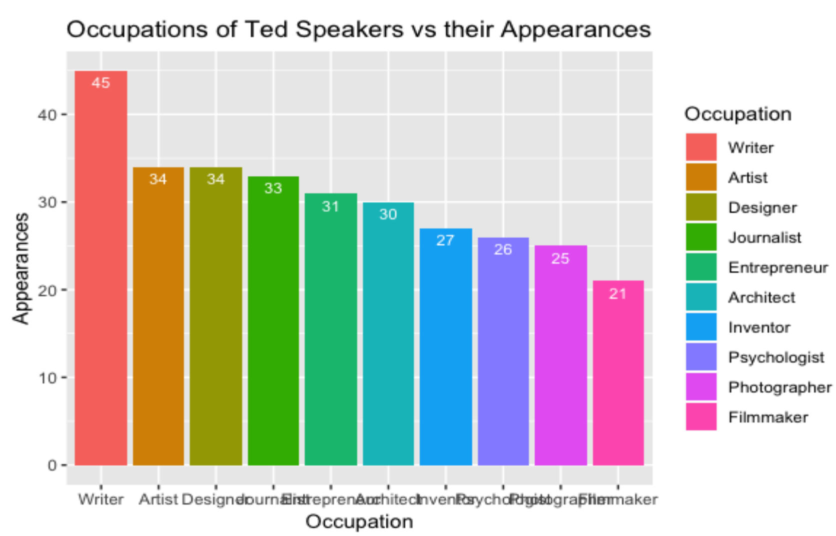

Figure 2. Occupations of Ted Speakers

Power Words

Words that are used most frequently have been coined as power words. To ensure that we had to get the right important words, we used tf-idf weighting. What tf-idf gives is how important is a word to a document in a collection, and that's why tf-idf incorporates local and global parameters, because it takes in consideration not only the isolated term but also the term within the document collection. From figure 3 we can see words like brain, kids, children, school being used the most in TED talks. From the power words we can infer that TED talks put emphasis on overall development of individuals during the early stages of life. They also talk about holistic development of an individual.

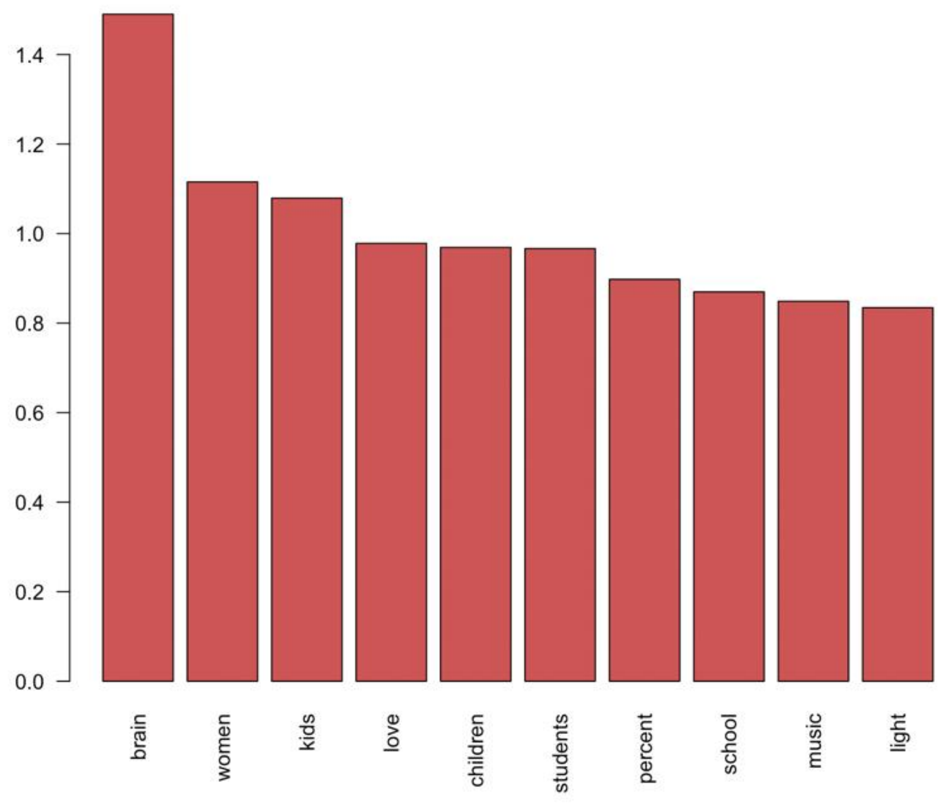

Figure 3. Power words

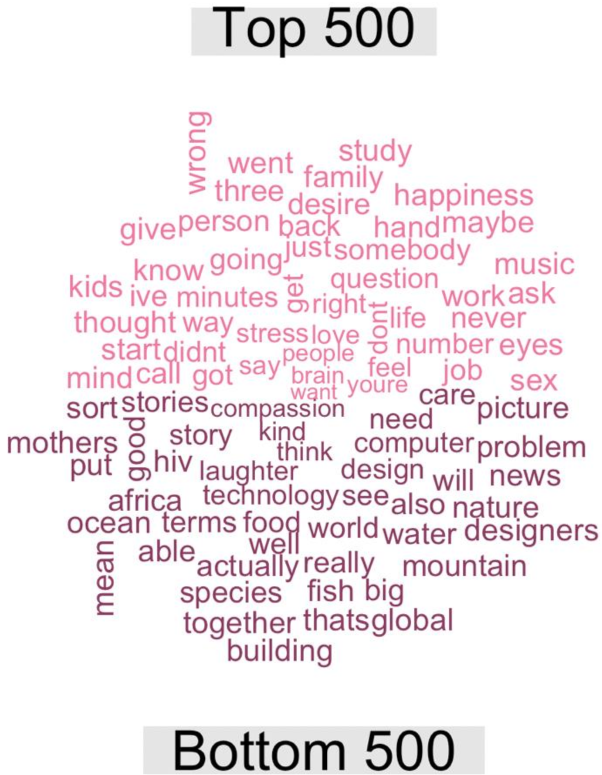

Figure 4. Word Cloud (Top 500 and Bottom 500)

Sentiment Lexicons

Sentiment lexicons can be defined as a database of lexical units for a language along with their sentiment orientations. They can be expressed as a set of tuples of the form (lexical unit, sentiment). Here, the lexical units may be words, word senses, phrases, etc. They are a list of words that are scored in some way (binary). In this project we used 3 essential lexicons – NRC, Bing and AFINN lexicons.

NRC Lexicon

The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). These eight emotions have been taken from Plutchik's wheel of emotions (figure 4). According to Plutchik, the primary emotions were learned through evolution. In his mind all other emotions are derivative of the basic eight. For example, remorse is a combination of disgust and sadness. Within the wheel itself each emotion has an opposite across from it and similar ones are adjacent.

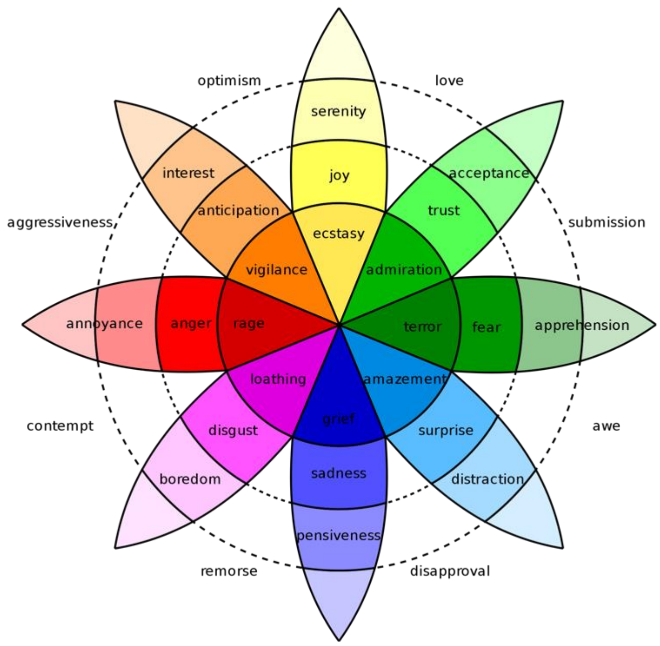

Figure 5. Plutchik's wheel of emotions

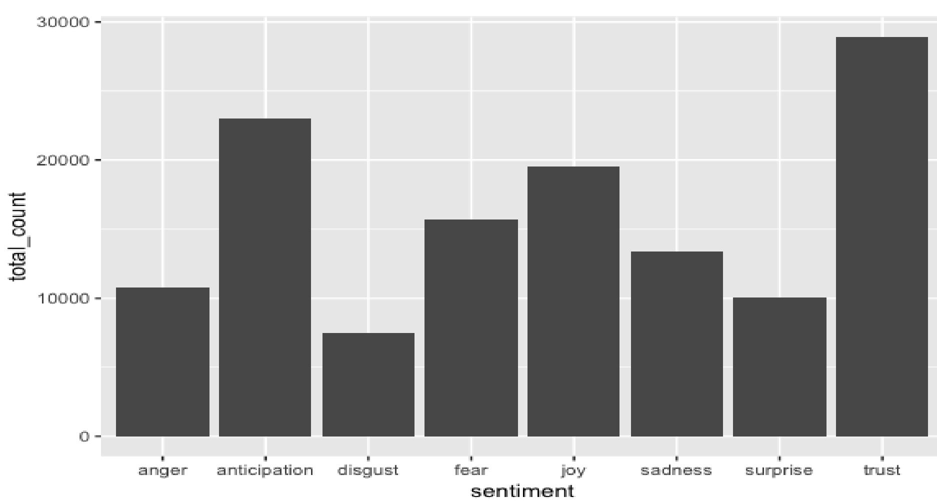

Figure 6. NRC Emotion Lexicon Top 500

The chart above depicts the total count of words corresponding to the particular emotion. We can see that trust and anticipation score the highest. Thus, we can infer that TED speakers build trust and engage with the audience with a sense of optimism while keeping the viewers engaged.

Moreover, a sentiment like fear also exists in a significant amount. Many TED speakers talk about various issues in the society we live in. Issues like global warming, epidemic spread, climate change etc. Some viewers might find topics like these as worrisome and might bring about some fear.

Bing Lexicon

The Bing lexicon categorizes words in a binary fashion into positive and negative categories. From the two plots shown below, we studied sentiment of TED speakers over a timeline. X axis represents the timeline and y axis stands for polarity of words. Polarity in sentiment analysis refers to identifying sentiment orientation (positive, neutral, and negative) in written or spoken language. On one hand, we can observe that the starting polarity is around +3 and ending polarity is about 8 in Figure 7. This means that as the time goes by, the TED speaker's talk will end on a higher positive note. On the other hand, Figure 8 shows that the starting polarity is higher than ending polarity. It also explains the reason why the bottom 500 talks do not get high views. Hence the study of Bing lexicons tells us that it is imperative for a speaker to end their talks on a positive and optimistic note.

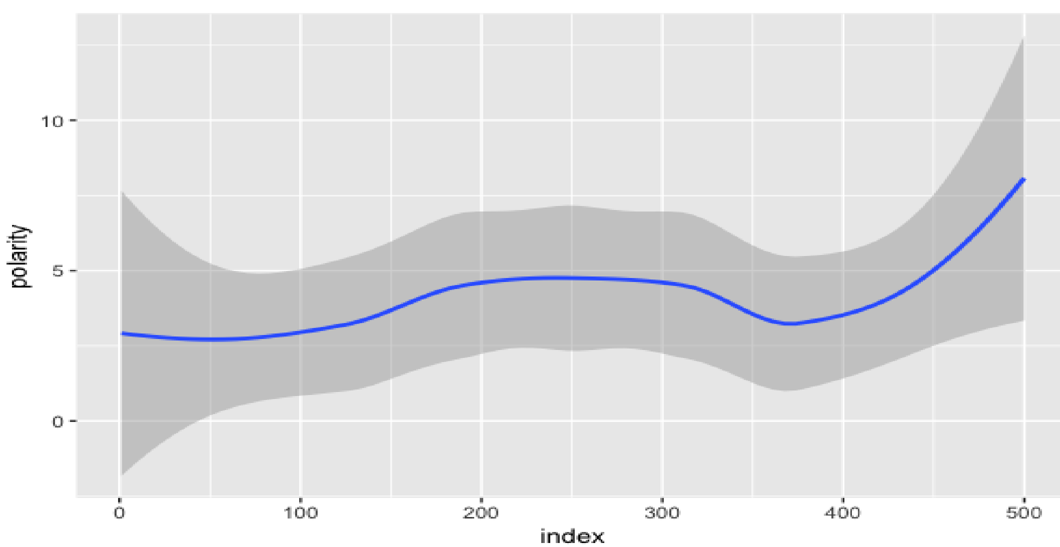

Figure 7. Bing Lexicon Top

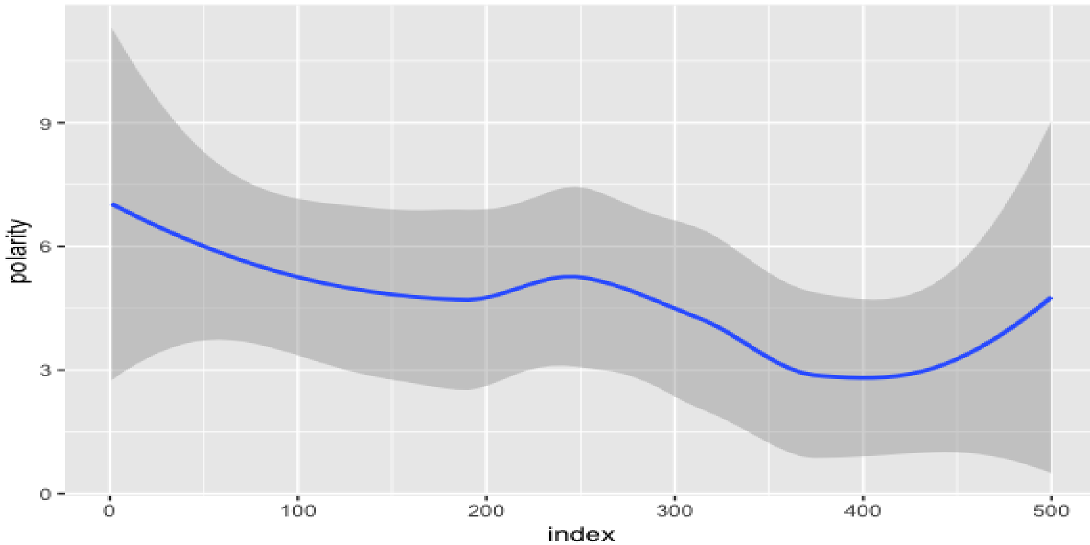

Figure 8. Bing Lexicon Bot

AFINN Lexicon

AFINN lexicon is a new word list for sentiment analysis. It assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. There are thousands of words (including a few phrases), each labeled with a sentiment strength and targeted towards sentiment analysis on the text. Unlike Bing lexicon, AFINN lexicon gives us a non-binary score that helps us understand how positive or negative the sentiments can be. Bing lexicons only give us the overall polarity binary score. By giving words different positive or negative scores, we plotted a graph to measure the density of polarity of words. Figure 9 depicts the density plot where we can observe that the highest density of words appears around +2.5 or -2.5. Some of the negative scores in the plot exhibit how TED speakers use a storytelling approach, sharing with the viewers their journey of overcoming failures and achieving success.

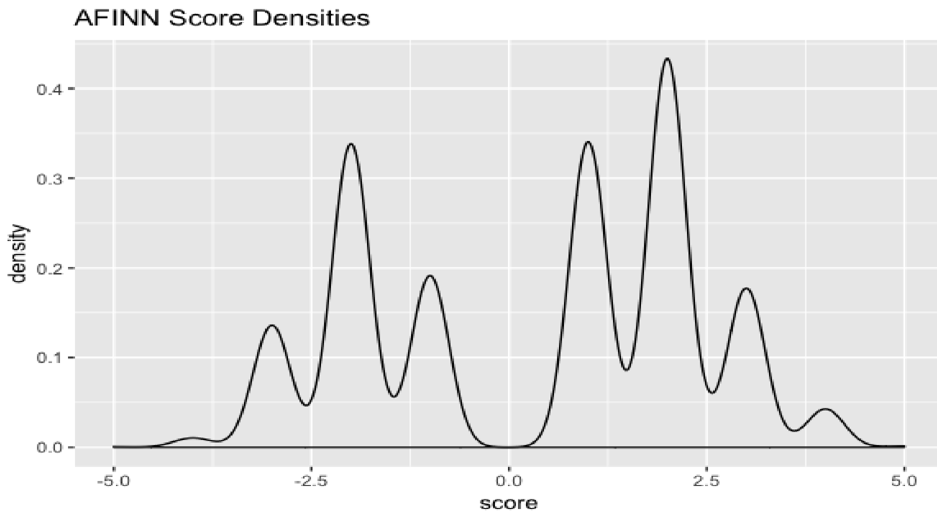

Figure 9. AFINN Score Densities

Topic Modelling

Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. To understand more about how TED speakers, give their speech, we compared TED talks to other speeches on the similar topics. Topic modelling between three different talks on women empowerment (TED, UN & Political speech) was carried out to gain insight on the differences of power words based on the intent of deliverance. All the high frequent words from the three talks were grouped and three clusters were made.

LDA

Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions. Figure 10 depicts the top terms for 3 topics. From the results we can clearly see that topic 1 refers to TED talks as the TED talks talk about development of individual. Topic 2 depicts the UN speech as terms like rights, social, change is often used at UN talks that bring about an empowerment and transformation ideas. The final cluster is about a political speech.

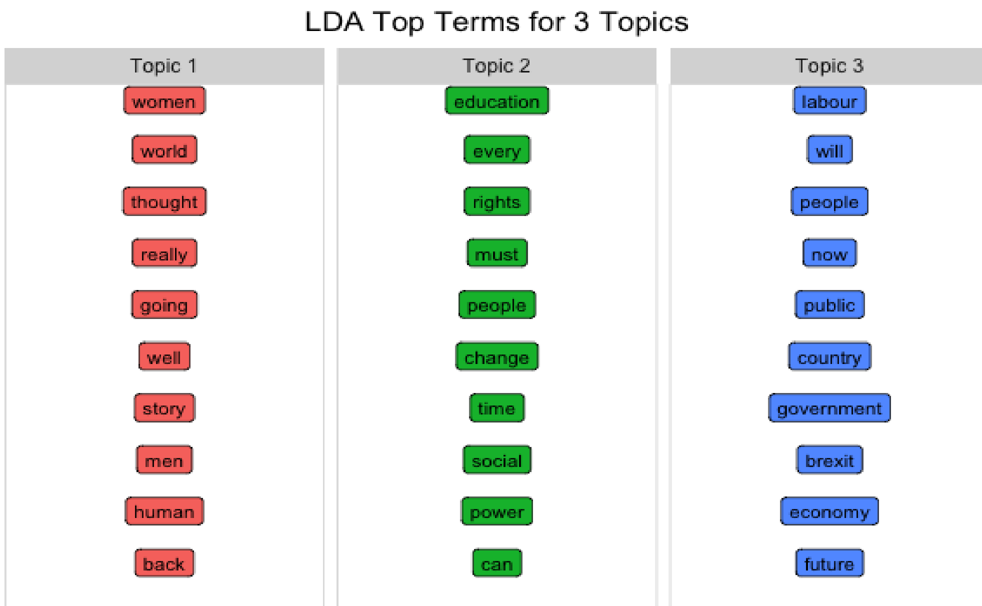

Figure 10. LDA Top Terms for 3 Topics

K-means

Data clustering is a kind of unsupervised classification, when the clusters are formed by evaluating similarities and dissimilarities of intrinsic characteristics between different cases. K-means clustering belongs to partitioning-based techniques grouping, which are based on the iterative relocation of data points between clusters. In the texting mining background, we need to classify the similar words appear in different talks. By executing K-means clustering, we tried to figure out that whether different fields' talks will have similarity or not.

Figure 11 shows us the results for K-means method. We can see that Topic 1 depicts a TED talk. The K- means does not seem to perform as well as LDA. It couldn't distinguish between the UN and TED talks well enough. LDA assigns a document to a mixture of topics. Therefore, each document is characterized by one or more topics (e.g. Document *D* belongs for 60% to Topic *A*, 30% to topic *B* and 10% to topic *E*). Hence, LDA can give more realistic results than k-means for topic modelling. Both methods can be tuned for better performance.

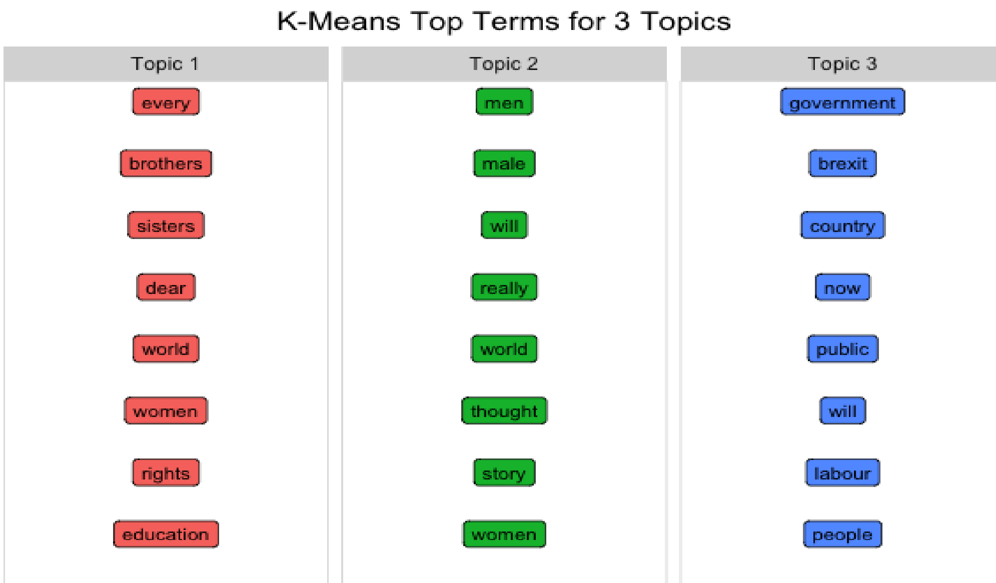

Figure 11. K-Means Top Terms for 3 Topics

Recommendation of similar speakers

Cosine Similarity

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter. In our project, we used key words of TED speakers' transcripts and speakers as the two vectors. Thus, if the two speakers have higher similarity on the key words, the similarity is higher for the two TED talks. If we regard all speakers as points in the coordinate, the angle between two points will be the measurement to judge the content similarity. There are two rules which are the extreme conditions. If the angle is 0 between two points, then the cosine value is 1. In this case, the two vectors will be oriented in identical directions. In other words, corresponding data sets will completely similar to one another. On the other hand, if the angle is 90 degrees, then the cosine value is 0. In the coordinate, the two vectors are perpendicular. However, it is not necessary to suggest that the two vectors are uncorrelated at all.

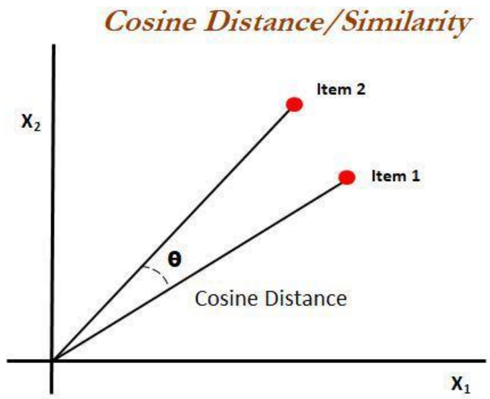

Figure 12. Cosine Similarity

Example

In our project, we extracted subgraph from our similarity network plot to depict a highly related network (figure 13). As we can see, Derek Sivers, Elon Musk, Elizabeth Gilbert and six more speakers are highly related in the network. The illustrated graph shoes that one of the eight speakers have similar key words. For example, if we see a video of Elon Musk on YouTube, there will be a high possibility of the appearance of Derek Sivers or Elizabeth Gilbert's videos on the recommended list. Hence this study makes it convenient for internet users to find more similar videos to the one's they have liked.

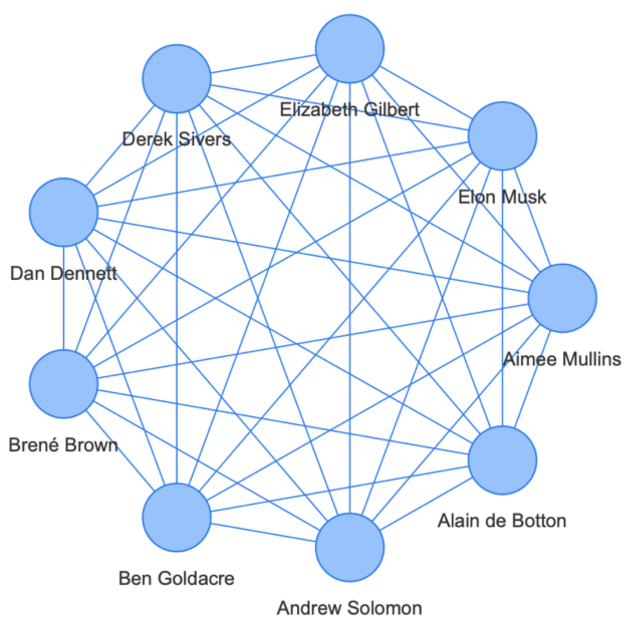

Figure 13. Similarity Network

Future Work

With further sentiment analysis, more processing about content in each specific category could be carried out. Accordingly, a speaker or viewer in a specific field can find more keywords on a topic, and the result can also be used to optimize search algorithms.

Time series sentiment analysis can bring breakthrough results. For example, in the finance industry, investment banks are utilizing real-time news to predict the market emotion simultaneously. The Bloomberg Radio can be an information source. With the transcripts of history data and the market index or stock price, sentiment analysis can help find the correspondence between the transcripts and trends of the market. This real-time analysis can be an indicator for trading decisions for investment companies.

Conclusion

While TED talks are special because of the stories they convey, performing sentiment analysis on them generated some interesting results on the similarities of these talks. We can analyze that although the TED talks are of progressive nature, they mostly revolve around wisdom, women, family and passion. Inspirations and suggestions for TED speakers might be derived from our study.

The greatest quality a TED speaker possesses is the ability to gain the trust of the audience. Also, creating an atmosphere of anticipation while storytelling is characteristic trait of great TED speakers.

References

[1] Singhal, A. (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43.
[2] John, P and Aaron, F. (2009) "(DISCRIMINANT FUNCTION ANALYSIS (DA))"
[3] Ayramo, S. & Karkkainen, T. (2006) "Introduction to partitioning-based cluster analysis methods with a robust example", Reports of the Department of Mathematical Information Technology; Series C: Software and Computational Engineering, C1, 1-36

[4] Efstathios, S. (2006), “Automatic Text Categorization in Terms of Genre and Author” Computational Linguistics, Volume 26 | Issue 4 | December 2000 p.471-495
[5] Marcelo, M. & Barbara, P. (2014), “Meta-level sentiment models for big social data analysis”, Volume 69, p.86-89.
[6] Pappas, Nikolaos, and Andrei Popescu-Belis. "Sentiment analysis of user comments for one-class collaborative filtering over ted talks." Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013.
[7] Mauro, Cettolo, Girardi Christian, and Federico Marcello. "Wit3: Web inventory of transcribed and translated talks." Conference of European Association for Machine Translation. 2012.
[8] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." Icwsm 11.538-541 (2011): 164.
[9] Maas, Andrew L., et al. "Learning word vectors for sentiment analysis." Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. Association for Computational Linguistics, 2011.