Paper 168-2019

Analyzing Factors Influencing Viewer Count of TED Talks through Text Analytics

Yasha Pastaria, Oklahoma State University; Miriam McGaugh, PhD, Oklahoma State University

ABSTRACT

The objective of this research paper is to explore the TED Talk data and generate some insights that others may find helpful when designing their own talk. This process would include understanding the trend of popularity of TED Talks over the years in terms of views, comments and ratings along with exploring possible drivers like occupation of the speaker, duration of the TED Talk, number of speakers etc. The analysis will be useful to consumers in understanding where TED Talks are heading over the years with respect to their popularity and identify its drivers. It will ultimately help them design the best TED Talk and avoid the mistakes of the worst ones.

INTRODUCTION

TED Talks aim to provide a platform for thinkers, visionaries and teachers, so that people around the globe can gain a better understanding of the biggest issues faced by the world with a desire to help create a better future. TED makes money through conference attendance fees, sponsorships, foundation support, licensing fees and book sales. Everyone who buys a pass to attend a TED conference is helping share free TED Talk videos with the world. The profits made by the TED conferences are directed toward initiatives such as global public health, poverty alleviation and biodiversity. It is imperative to assist this endeavor by identifying factors that affect view count and help generate more revenue for these worthy causes.

METHODOLOGY

The methodology used for this research has been divided into 3 stages shown in the diagram below:

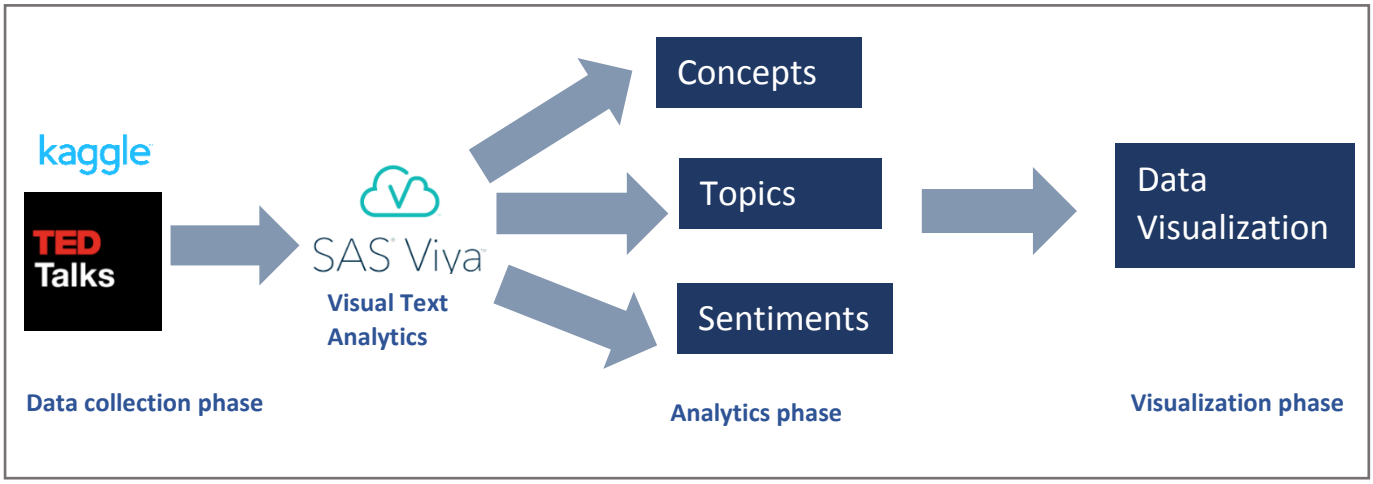


Figure 1: Methodology

DATA COLLECTION

The data source for the analysis was a Kaggle Dataset called TED Talk Data. The main dataset contains metadata about every TED Talk hosted on the TED.com website until September 21, 2017. The data has been collected by scraping the official TED Website and was made available under the Creative Commons License.

The main TED dataset contained a total of 2,550 rows and 14 variables where each row represented data of a particular TED Talk. Another dataset, Transcripts, had 2,400 rows and two variables. This dataset contained transcript information for each of the TED Talks. The datasets were directly imported into SAS Viya Visual Text Analytics environment and were later used for analysis.

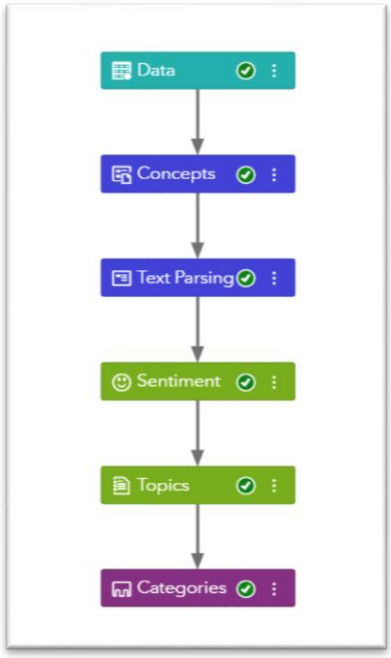
Variable Name	Description
Comments	The number of comments made on the talk
Description	A blurb of what the talk is about
Duration	The duration of the talk in seconds
Event	The TED/TEDx event where the talk took place
Film Date	The Unix timestamp of the filming
Languages	The number of languages in which the talk is available
Speaker	First named speaker of the talk
Name	The official name of the TED talk. Includes the title and the Speaker
Num Speakers	Number of Speakers in the talk
Publish Date	Date on which the TED talk was published
Ratings	A dictionary of various ratings given to the talk
RelaTED Talks	List of recommended talks to watch next
Speaker Occupation	Occupation of the main speaker
Tags	Themes associated with the talks
Title	Title of the talk
URL	URL of the talk
Views	Number of views on the talk

Table 1: Metadata

ANALYTICS PHASE

BUILDING PIPELINES USING MODEL STUDIO

SAS Viya Model Studio was used to explore the data and extract useful information through pipelines. The Model Studio pipeline in figure 2 performs six steps during the analysis. Each of the steps is described further.



1. Data- This is the first node for the pipeline. It takes data from the TED_main table in the PUBLIC CASLIB library. The node automatically creates a uniqueID variable that is a unique identifier of the data. The description column contained text data and was assigned "text" role for the project.

2. Concepts- The next node is the Concepts nodes, which extracts useful information from text data and analyzes information in context. A concept can be a property such as country, gender, name, book title and so on. SAS Visual Analytics offers two kinds of concepts, Predefined Concepts, which are concepts whose rules are already written, and Custom Concepts where rules can be written for recognizing concepts that you think are important.

As can be seen from Figure 3, SAS Viya has nine predefined concepts such as dates, people, places, measurements, mentions of currency. A custom concept, TEXT_CONCEPT, was created for the unique purpose of treating concepts related to technology such as new technologies, technologist, technological and so on as similar concepts.

Figure 2: SAS Visual Text Analytics Pipeline in Model Studio

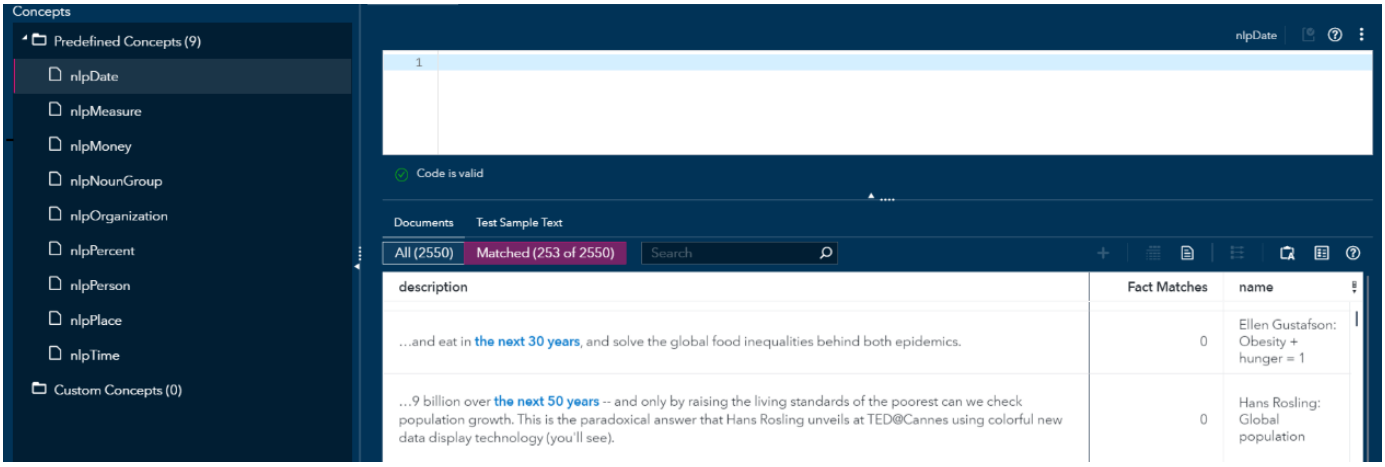


Figure 3: Concept Interactive Viewer

From Figure 4, it can be seen that this was done by examining similarity scores of textual elements with each other. For example, "new technology" has a similarity score of 0.861 with "technology" and should be treated as the same concept.

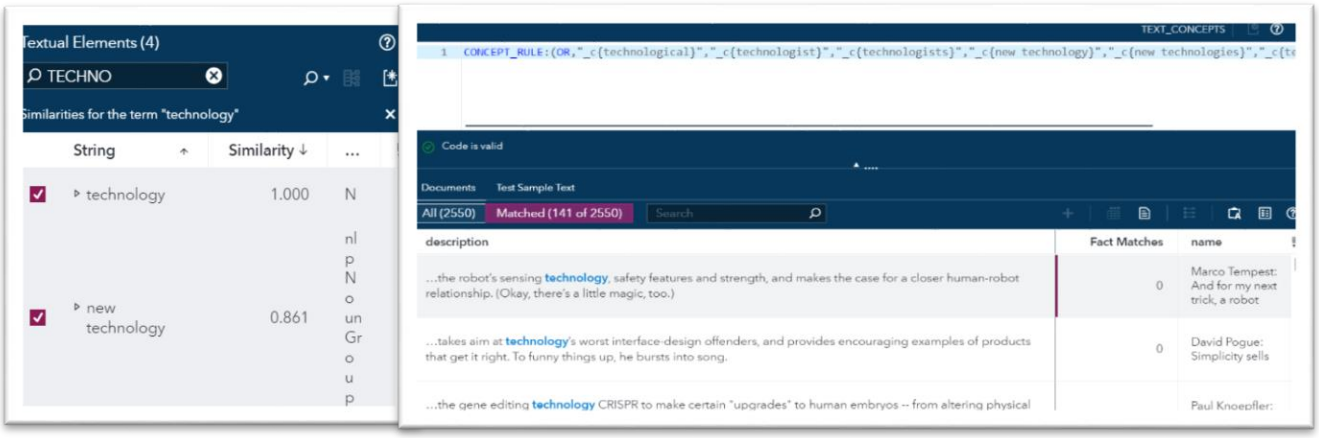


Figure 4: Defining a Custom Concept

3. Text Parsing- The Text Parsing node performs tokenization of data, detects part-of-speech, checks for spelling mistakes and puts synonyms under their parent terms. From figure 5, the interactive viewer gives flexibility to keep and drop terms with the frequency and number of documents in which they appear. Terms such as too, not, go, way, as and so on have been dropped.

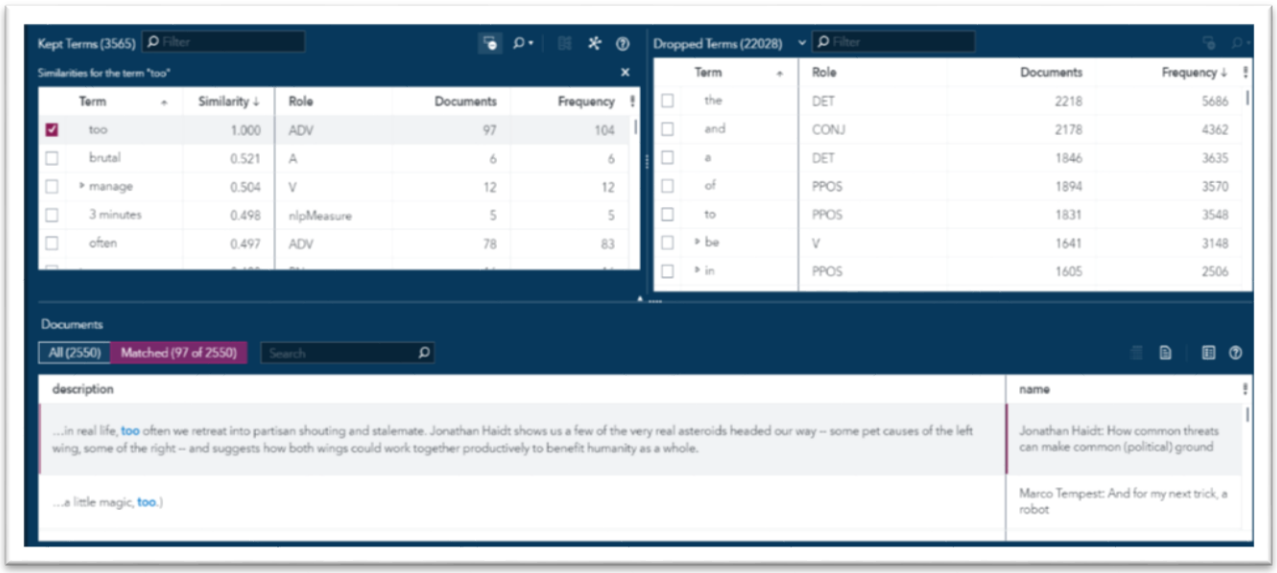


Figure 5: Dropping Terms through Text Parsing Node

As we can see form the below term matrix that "talk" and "world" are the most frequent words. We can also observe that the frequency of "talk" has increased because all of the synonyms of "talk" have been merged together.

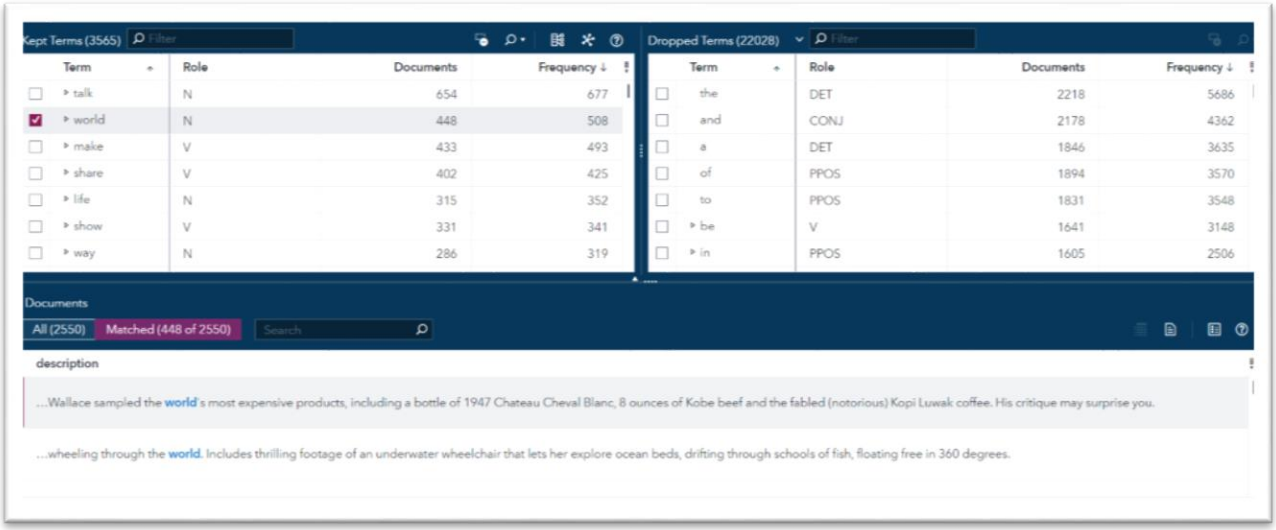
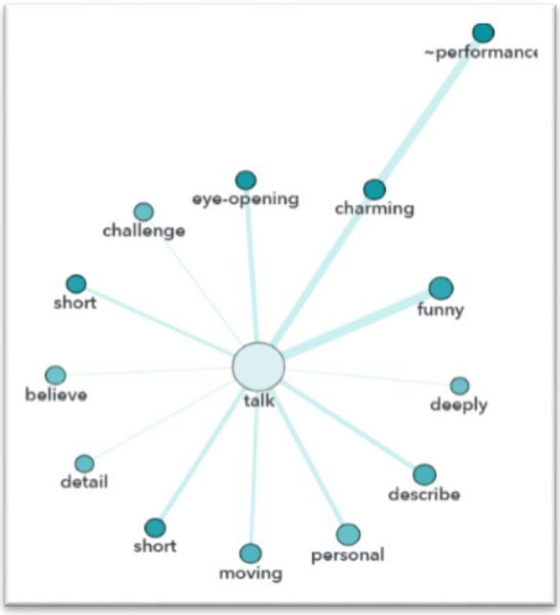


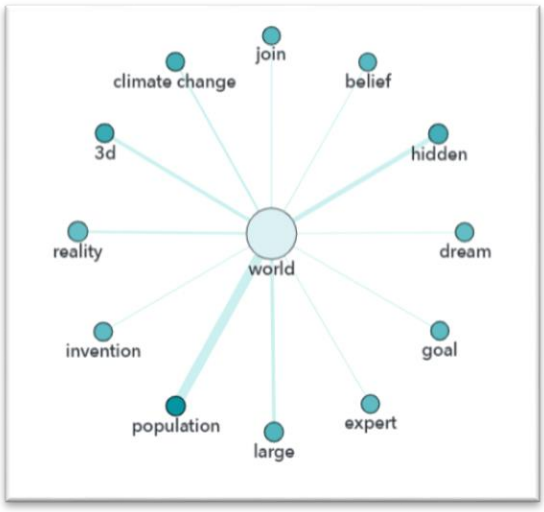
Figure 6: Text Parsing Node Interactive Viewer

3.1 Term Map – Term maps can be viewed through the interactive filter inside the text parsing node. Term map is an association analysis that links terms based on their similarity score (i.e the number of times two terms occur in a document together). The following term map has been obtained using the most frequent terms in the TED Talks .



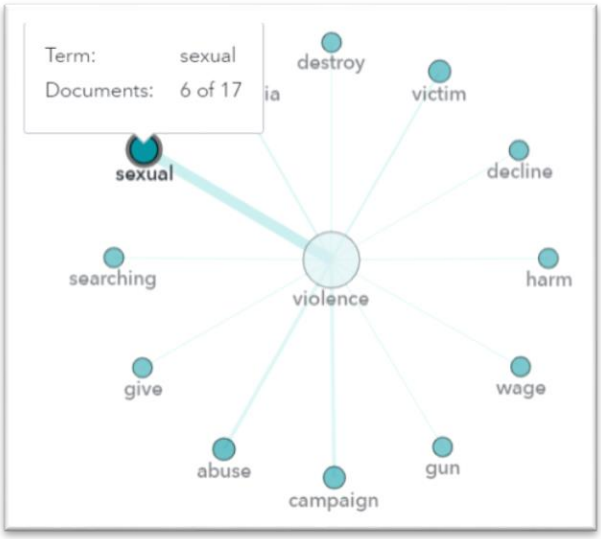
TALK – the term map is linked with various adjectives that define the characteristics of a talk and what people expect from a good talk. The word "talk" has the highest association with the word "funny" followed by "charming" and "performance". The word "talk" appeared with "funny" 55 times out of 83 documents, with "charming" 28 of 40 times and with "eye opening" 22 out of 29 times. This gives us a little idea about what people perceive as a good TED talk. People also love talks that are short and with a deep message behind it.

Figure 7: Term Map for the Word TALK



WORLD- Today, people are starting to become active participants in the fight against global issues and as a result, progress is being made. However, there are still individuals unaware of pressing issues around them. TED Talks try to bring these people up-to-date with global issues around the world. As can be seen from the term map, "world" is the most talked word in the TED Talks. The association between "world" and other terms highlight issues such as "population" and "climate change". Terms like "goal", "join", and "belief" in association with "world" highlight the collective effort of people to join hands and make the world a better place.

Figure 8: Term Map for the Word "WORLD"



VIOLENCE- Global issues are important but there is another issue we often overlook, which is Violence. The term map for "violence" has the strongest association with the word "sexual". According to United Nations Foundation, violence against women and girls is one of the most fundamental and pervasive violations of human rights in the world. It's association with other terms show it comes in various different forms. It is rooted in "abuse", "destruction" and harmful social norms. The words victim, campaign, searching show that TED talks have been able to garner people's attention to combat violence.

Figure 9: Term Map for the Word VIOLENCE

4. Sentiment -The Sentiment node generates sentiment scores. These scores help in identifying the overall attitude or tone (positive, negative and neutral) of the document.

5. Topics - The Topics node generates naturally occurring themes in the data. It groups important terms in the document together via unsupervised machine learning methods. It uses Latent Dirichlet Allocation (LDA) technique, which is based on Singular Value Decomposition. The interactive viewer provides features such as merging topics, splitting topics and assigning topics to categories. Figure 10 shows two similar topics that have been merged, after which other topics will be recalculated.

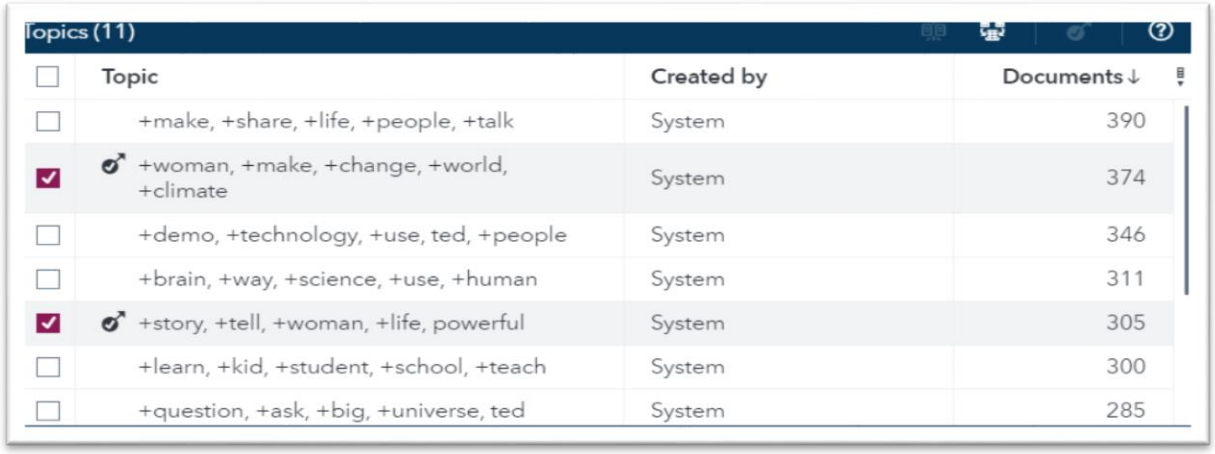


Figure 10: Merging Similar Topics in the Topics Node

In figure 11, the table on the left shows important topics and the number of documents in which they appear. The table on the right contains terms from these topics. The topics are matched in the documents along with the overall sentiment of the document. The topics are from different categories such as world and economy, ocean and animals, brain and behaviours and so on. Because the sentiment node preceeded the topics node, an overall sentiment of the document is also displayed.

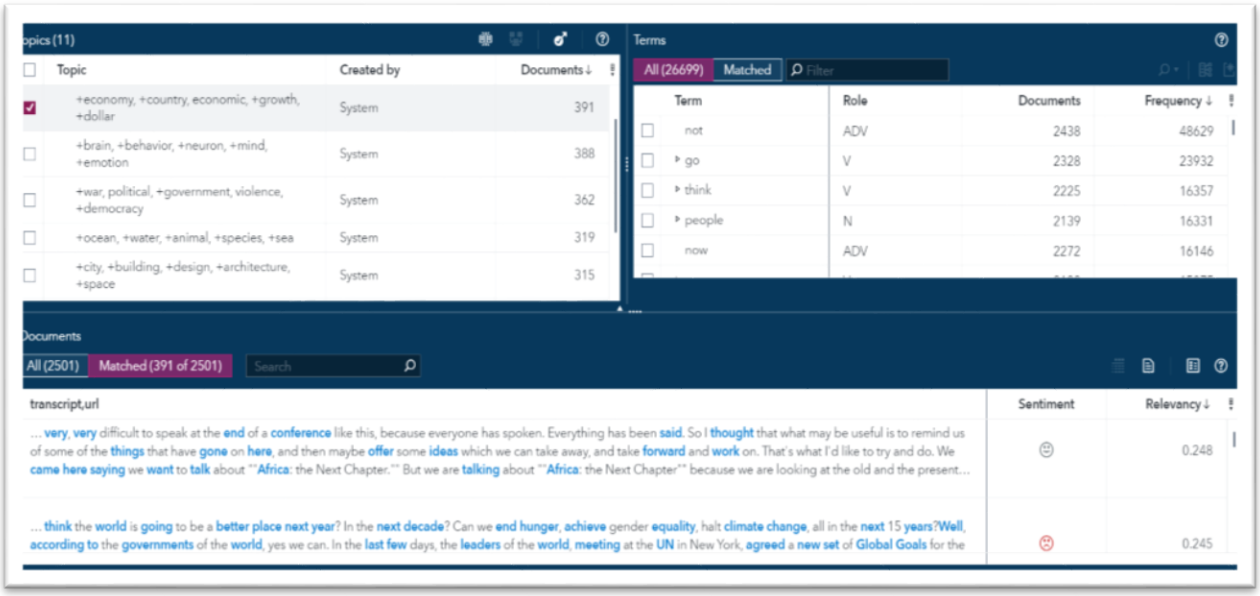


Figure 11: Frequent Topics and Their Sentiment

The first thing we can see from the word cloud in figure 12 is that the most frequent word is "talk", which is obvious because we are analyzing TED Talks. The second word is "world". This is reflective of TED Talk's mission of welcoming people from every discipline and culture who seek a deeper understanding of the world. The makers of TED Talks believe

passionately in the power of ideas to change attitudes, lives and, ultimately, the world. When we look at the topics in figure 13, we see that most of them contain these terms in one, or another format, but also contain positive words like "build". The bar for each topic is colored in the ratio of how many negative (blue), neutral (orange) and positive (green) sentiment polarity terms occur. In this case most of the topics have mainly positive terms, possibly due to the message that every TED Talk spreads about creating change in the world. The topic story, tell, women, life contains more negative terms compared to other topics and this can be due to women sharing their life experiences in these TED talks.

Figure 12: Word Cloud of Topics in TED Talks

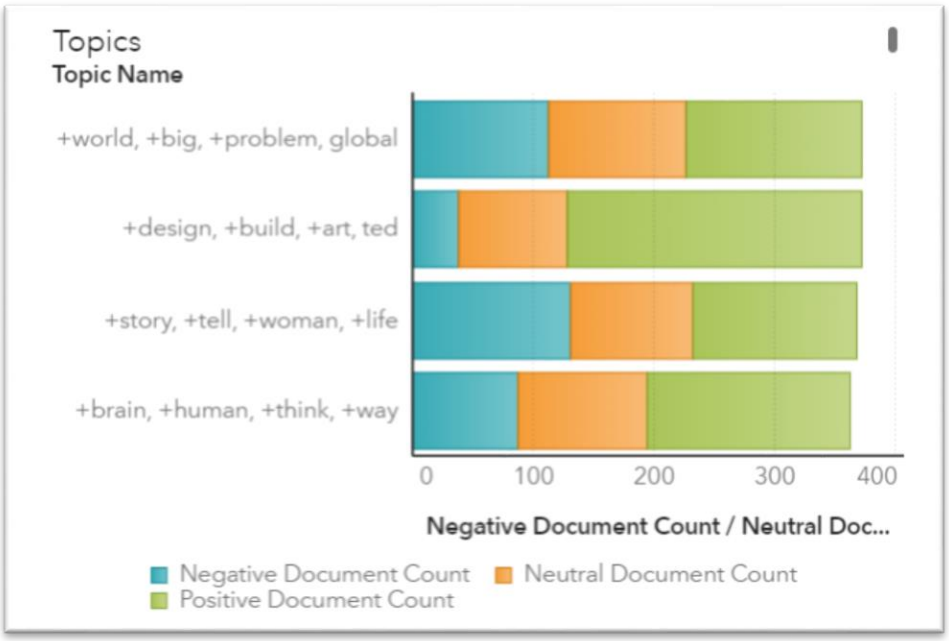


Figure 13: Most Frequent Topics in TED Talks

Every Ted Talk in the dataset had tags associated with it in the Tags column. The figure 14 shows the most important and frequent tags that appeared in the TED talks. These were technology, global issues and science. This is understandable considering the boom of technologies every decade such as big data, artificial intelligence, machine learning, blockchain and so on. It is also interesting to see the occupations of speakers that people like listening to. The word cloud in figure 15 shows the most popular speaker occupations, and these are mostly Writers, Designers and Entrepreneurs.

Figure 14: Most Frequent Tags in TED Talks

Figure 15: Popular Speaker Occupations in TED Talks

6. Categories- Categories help in identifying a group of documents that share some similar characteristic. Three popular topics identified from the topics node were promoted as categories as can be seen in the figure 16. Categorization is done through linguistic rules rather than statistical weighting of terms. Each category is defined using keywords and important contexts. The model then evaluates each document against these categories based on the rules defined. The output data from the categories node was further used to create visualizations as can be seen in the figure 17. The figure shows the frequency of occurrence of each category in all the documents and the keywords that are used to identify categories from each document.

Figure 16: Promoted Categories and Rules

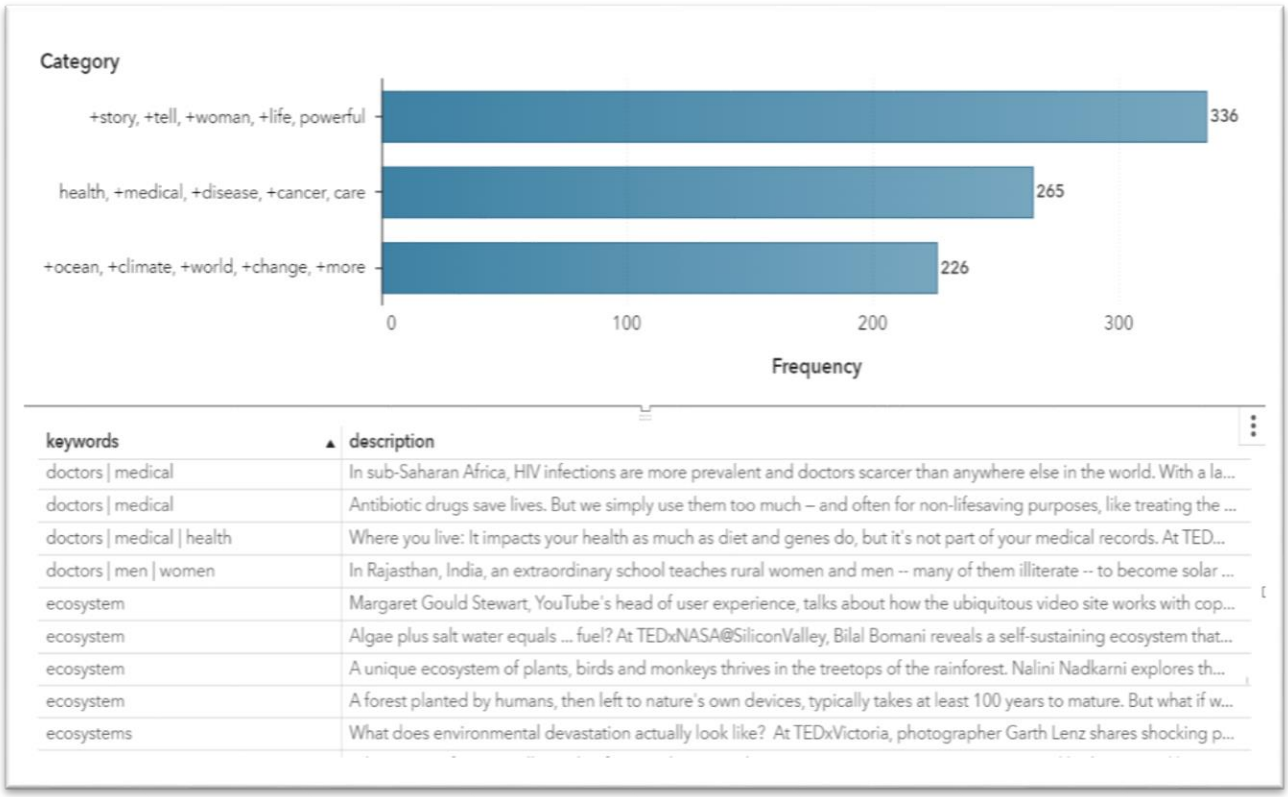


Figure 17: Popular Categories and Keywords for Identifying Them in a Document

VISUALIZATION PHASE

We examined the frequent topics that appear in TED Talks along with their tags and keywords. We also promoted categories. The below bubble plot (figure 18) shows the comments and views in these three categories sized by their frequency percent. The category with topic (+story, +woman, +life, +powerful) seems to be the most popular among TEDsters with 39.58% values.

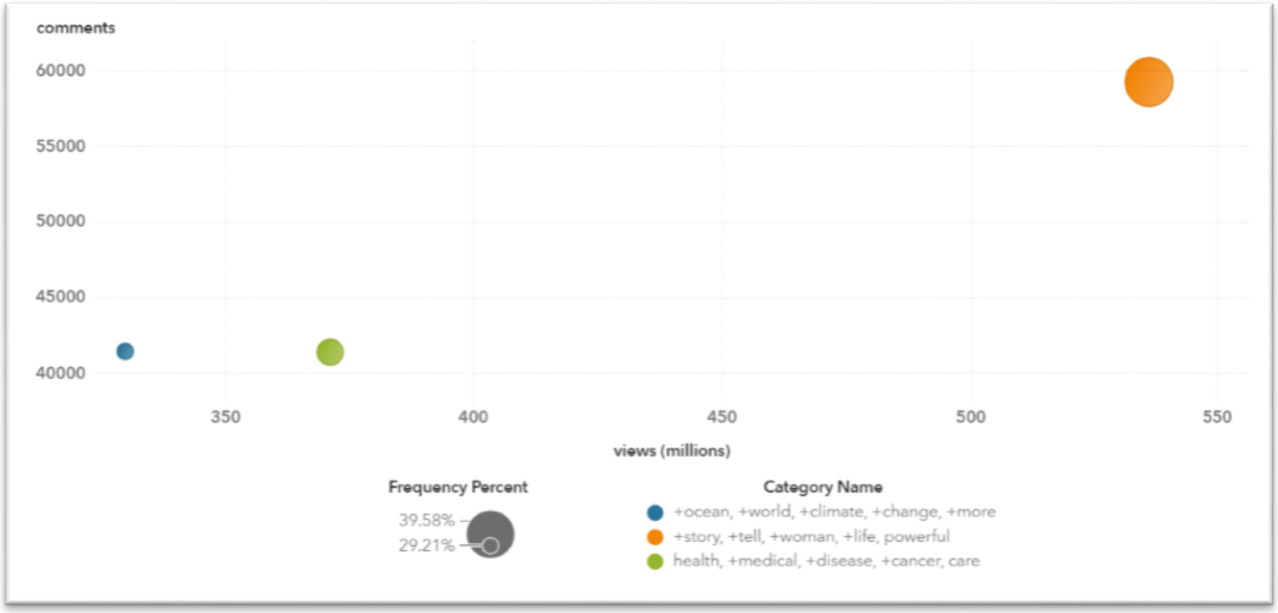


Figure 18: Bubble plot of comments by views sized by Frequency Percent

While these topics appear famous and important, it is important to analyze their trend over the years. The below trend line chart makes a comparison between the three topic categories between the year 2006 to 2017. The topic Health seems to have an increasing trend over the years with sudden falls in years like 2013. We saw technology term rising in the year 2012-2013 and this can be one of the reasons for the less popularity of this topic in those years.

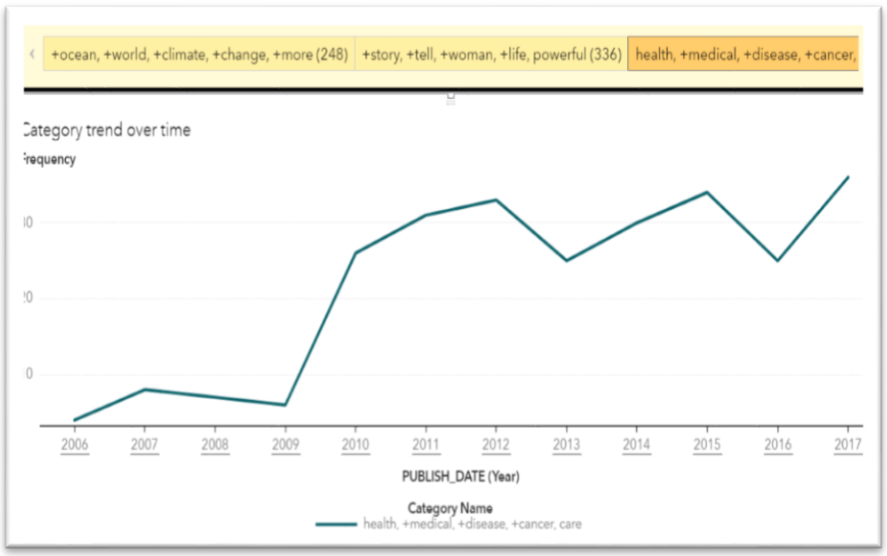


Figure 19: Trend Chart for Health Topic

The Heat map below shows the most popular months when TED talks are viewed. These months are usually between March to May with March being the most popular and August being the least popular.

Figure 20: Heat Map for Popular Months

We all must have, once in our lives, sat through a TED Talk that left us inspired and energized or for some people boring and not as inspiring. What makes a TED Talk really interesting is the content and the message it delivers. While there are talks that deliver messages in a short time, there are also talks that can be really long to listen to. It will be interesting to see if there is any relationship between the duration of the talk and the number of views. The below scatter plot shows a correlation of 0.05 and a weak association between the two. This means that the content of the talk matters to audience not the duration.

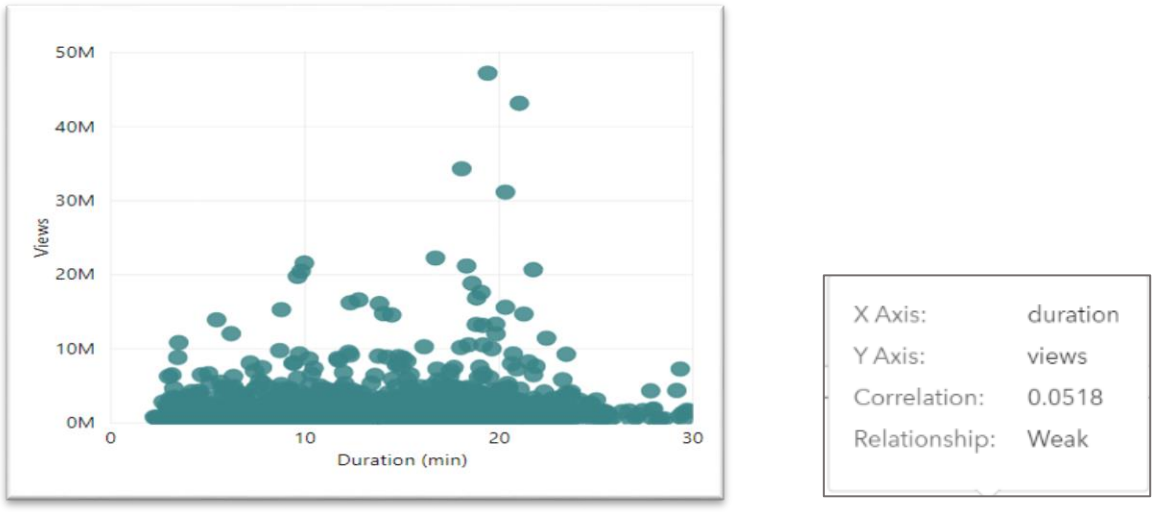


Figure 21: Scatter Plot between Duration and Views

The TED_main dataset had a ratings column with different ratings for each talk along with the count. The ratings were categorized into two groups using Python as Positive Ratings (values such as funny, inspiring, informative, beautiful, jawdropping) and Negative Ratings (values such as longwinded, ok, confusing). A scatter plot was generated in Power BI between the positive and negative ratings and automatic clusters were found. As can be seen from the below figure, the two clusters seem to be well defined with cluster 2 having more positive ratings and views compared to cluster 1. Cluster 2 comprises of all official TED Talks while cluster 1 comprises of all other TED events such as TED Global, TEDx, TED Women and so on.

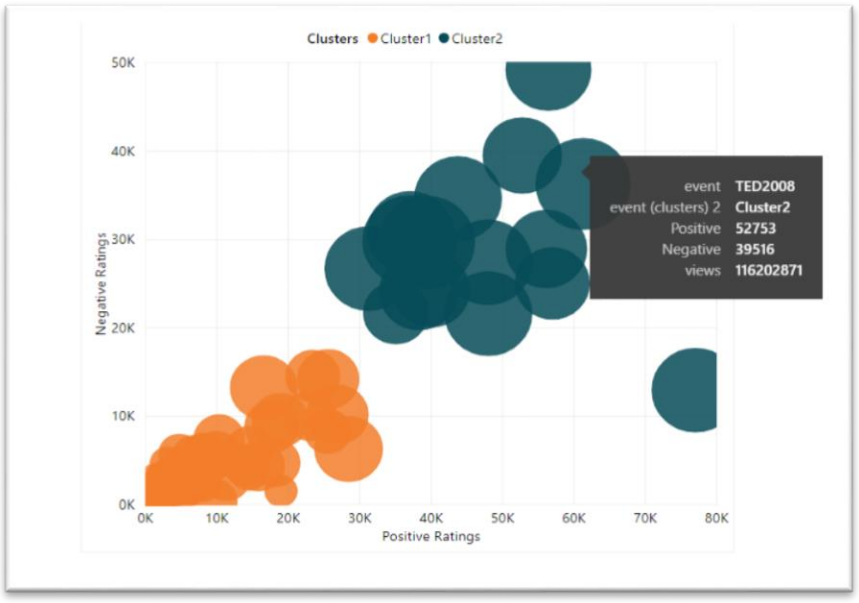


Figure 22: Scatter Plot between Positive and Negative Ratings for TED Talks

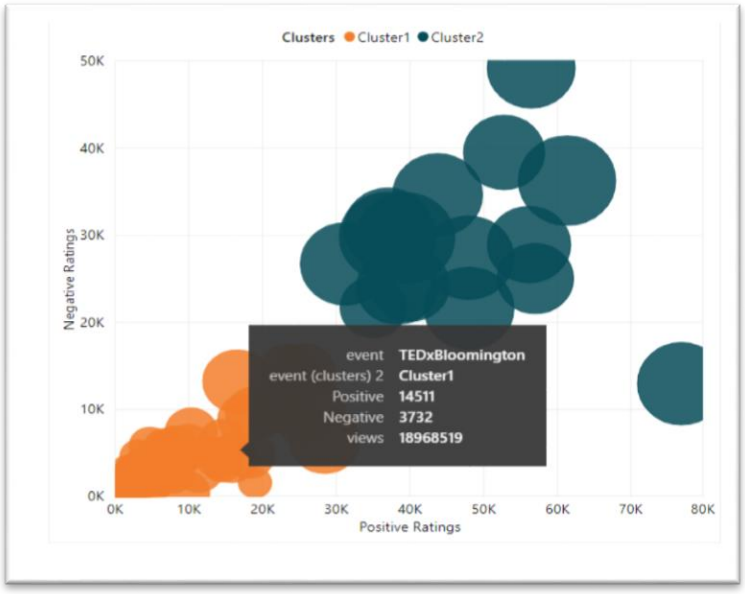


Figure 23: Scatter Plot between Positive and Negative Ratings for TED Events

After examining the topics of various TED talks, popular speaker occupations and months and themes across which they are popular; we would lastly analyze the top ten most viewed TED talks till the year 2017 and the top ten speakers with maximum views.

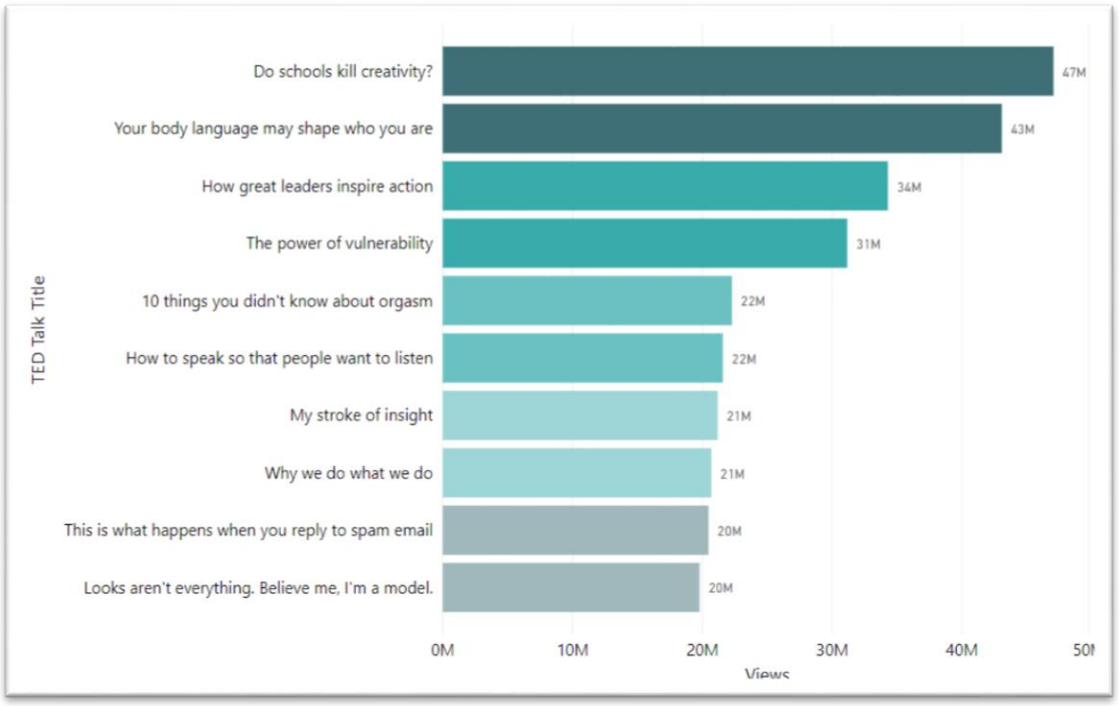


Figure 24: Top 10 Most Viewed TED Talks

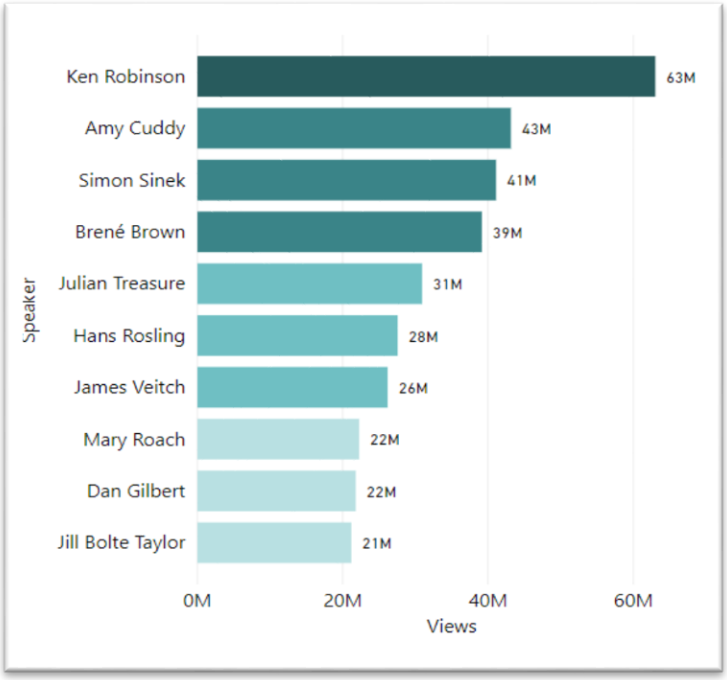


Figure 25: Top 10 Most Viewed Speakers

CONCLUSION

From the analysis we can see that:

- The most popular word in TED Talks is world and the most popular tags are technology and global issues.
- There are four topics most popular among TEDsters which are Health, Climate Change, Women Empowerment and World Issues.
- The most popular speaker occupations are Writer, Designer and Entrepreneur.
- Health related topics have seen an increasing trend over the years.
- The most popular months for viewing Ted Talks are March and May.
- The length of a talk has no relation with its popularity, people give value to content.
- Lastly, the most popular TED Talk of all time is "Do Schools Kill Creativity" and the most popular speaker of all time is Ken Robinson

REFERENCES

1. https://communities.sas.com/t5/SAS-Communities-Library/Discover-Main-Topics-on-MLKDayofService-Tweets-Using-SAS-Visual/ta-p/444259
2. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3856-2019.pdf
3. https://documentation.sas.com/?docsetId=ctxtug&docsetTarget=p0q83mpzlowlnun1fopgm5iv8mir.htm&docsetVersion=8.2&locale=en
4. https://video.sas.com/detail/video/5812813159001/creating-concept-rules-using-textual-elements-in-sas%C2%AE-visual-text-analytics

ACKNOWLEDGEMENT

I sincerely thank Dr. Goutam Chakraborty for his valuable guidance and motivation for accomplishing this paper. I also thank Dr. Miriam McGaugh for her constant support and suggestions throughout this study.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Yasha Pastaria
Master's in Business Analytics, Oklahoma State University, Stillwater
405-780-2133
Yasha.pastaria@okstate.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.