**Syracuse University**

**IST-736 Final Project**

**Text Mining on TED Talks**

Authored by: Jean Deleon, Kelly Schillacci, Thulasi Ram Ruppa KrishnanSulav Rupakheti

IST 736

Professor Ami Gates

Contents

[INTRODUCTION 3](#_Toc19048854)

[ANALYSIS AND MODELS 5](#_Toc19048855)

[CLUSTERING 7](#_Toc19048856)

[TOPIC MODELING 11](#_Toc19048857)

[SUPERVISED CLASSIFICATIONS 19](#_Toc19048858)

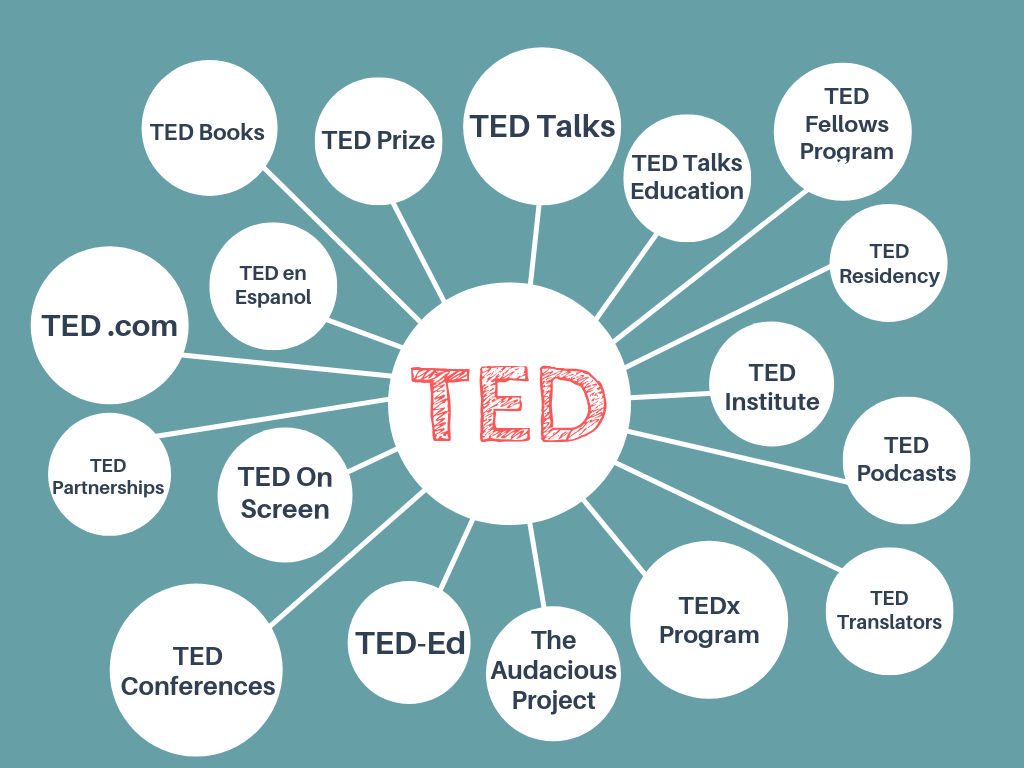
[COMPARISON OF CLASSIFICATION MODELS 20](#_Toc19048859)

[CONCLUSION 21](#_Toc19048860)

## 

## **INTRODUCTION**

TED Conferences LLC (Technology, Entertainment, Design) is a media organization that posts talks online for free distribution under the slogan "ideas worth spreading." TED was conceived by Richard Saul Wurman in February 1984 as a conference; it has been held annually since 1990. TED's early emphasis was on technology and design, consistent with its Silicon Valley origins. It has since broadened its perspective to include talks on many scientific, cultural, political, and academic topics. It is owned and curated by Chris Anderson, a British-American businessman, through the Sapling Foundation.



How did one-off conference about technology, entertainment and design become a viral video phenomenon and a worldwide community of passionate people? TED Talks is the flagship video series of great talks and performances, filmed at TED conferences, independent TEDx events and on other stages worldwide. TED.com is the flagship website for both distributing our 2,000+ TED Talks and sharing ideas. At TED Conferences, speakers appear on the main stage to give 18-minute talks and shorter presentations, including music, performance, and comedy. The TEDx program lets individuals, organizations and communities worldwide hold local, independent TED-like events. To date, more than 13,000 TEDx events have been held in 150 countries.

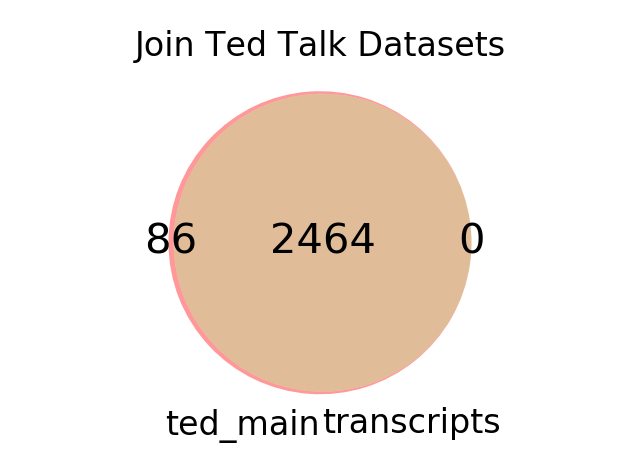
Educators can use any TED Talk and build a customized lesson around it. The TED Institute helps organizations unlock institutional knowledge and surface innovative thinking. The TED Partnerships team works with the world's leading corporations and foundations to connect the TED content and audience to their missions. The TED Talks Education one-hour program brings together a diverse group of teachers and education advocates delivering short, high-impact talks on the theme of teaching and learning. TED brings ideas worth spreading to the world with a growing collection of original content.



## **ANALYSIS AND MODELS**

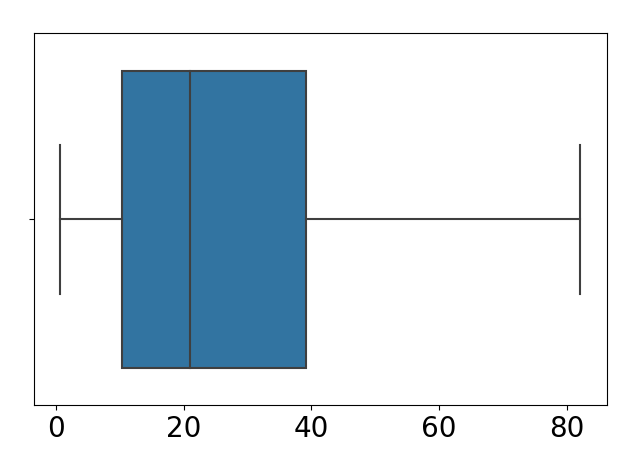
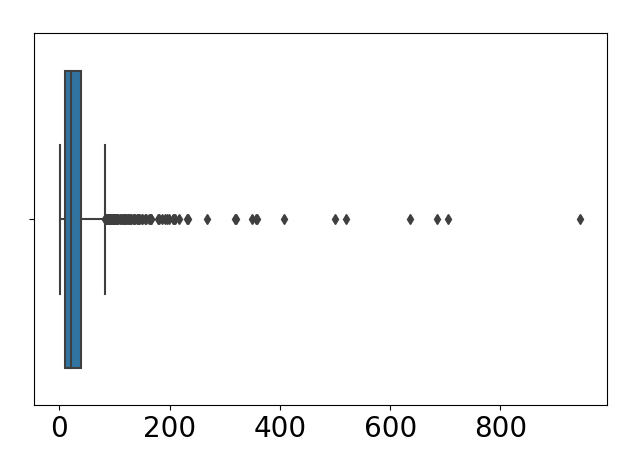
**About the Data**

There are two datasets retrieved from Kaggle. The datasets contain information about all audio-video recordings of the TED Talks uploaded to the official TED.com website until September 21st, 2017. The first dataset, TED main, contains information about all talks such as number of views, number of comments, descriptions, speakers and titles. The second dataset, TED transcripts, contains the audio transcribed to text data for all available talks on TED.com. TED Talks wants to find, organize, and share ‘ideas worth spreading’ for decades to come. For that reason, this analysis will attempt to model and predict the most popular talks from the past decade so that TED can focus on posting the most interesting talks.



**Figure 1.1 Join Ted Main and Ted Transcripts**

The datasets had to be preprocessed for bad data i.e. missing data, duplicates, formatting, and outliers. Because the metadata and the transcripts for each Ted Talk are separated into two files, the first step was to perform an inner join. Each talk was joined by a common key, the url column. As a result, the rows from both the TED main and TED transcripts found 2464 matches. The 86 rows found without a match in TED main were discarded. Additionally, the consolidated TED dataset was inspected for duplicate and incomplete records. Moving forward, each talk contained event metadata, information describing which TED conference a given TED Talk took place. Further review revealed that some records did not correspond to a TED Talk event and or were TEDx - TEDx are independently organized TED events that focused on local communities. Having said that, the decision to remove all non TED Talk and TEDx records was motivated to improve the computational effort to analyze each TED Talk’s transcript later. Furthermore, the file date and published dates were ill formed and so were corrected. However, only records filmed earlier than 2010 were saved for analysis. With most of the cleaning done, only 1374 records remain, a 44% reduction.



**Figure 1.2 Capping Outliers (Popularity Scores)**

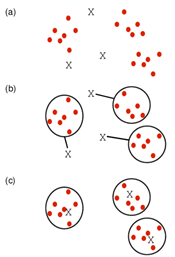
The dataset did not contain any labels for purposes of performing supervised machine learning. Labels had to be created based on the intended goal of categorizing TED Talks by popularity. For that reason, the ratings column, a dictionary based rating system which recorded the total number of people that thought a given talk was i.e. inspiring, funny, longwinded, etc. Since each word in the rating system can be categorized towards a positive or negative sentiment, a sentiment lexicon was developed so that the count for all positive or negative words can be summed. In doing so, the compilation of a “popularity score” can be calculated from the total number of positive over the total number of negative for each talk. As shown in Figure 1.2, the dataset contained popularity scores beyond the highest observation. As such, data points considered to be outliers were capped at the upper whisker. Furthermore, each numerical based popularity score were transformed into qualitative measures: popular, nominal, unpopular. All popularity scores greater or equal to the 75th percentile were classified as popular, and all popularity scores less than or equal to the 25th percentile were classified as unpopular. Records classified as nominal were deprecaded from the dataset further reducing the record count to 688.

With the dataset cleaned and labeled, features were generated for vectorization. The approach used was to identify the most important unigrams and bigrams. To select the most important unigrams, a random forest of 100 decision trees compiled each transcript, and only transcripts having a mutual information score at the 99 percentile were kept. Next, the dataset were separated into popular and unpopular sets. In doing so, the impost important bigrams can be generated to distinctly differentiate from bigrams used in popular and unpopular talks. Each bigram was generated based on the collocation of multiple words commonly co-occurring, as measured using Pointwise Mutual Information. Bigrams found to be common between the popular and unpopular talks were removed because the intention is to only use bigrams that are unique or more prevalent in either popular or unpopular talks. The ngrams generated were used as input terms when creating vectorized matrices of Count, Boolean, and Term Frequency-Inverse Document Frequency measures (consisting of about 900 features and 688 records). All vectorized datasets were then saved into a file for further and or later analysis.

## **CLUSTERING**

**K means clustering**

K means require a number (k) as in input to represent number of clusters. This algorithm then randomly assigns items to the k clusters. Calculate new centroid for each of the k clusters and the distance of all items to the k centroids. Then assign items to the closest centroid. Repeat this process until cluster assignments are stable as illustrated in Figure 2.1

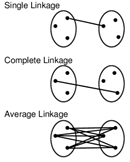


**Figure 2.1 K Mean Clustering**

*Source* [*http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html*](http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html)

#### **Cluster Linkage**

Distance between two clusters can be derived by using single, complete or average linkage methods as shown in Figure 2.2. Single linkage uses the minimum distance between two points, whereas the complete linkage used the maximum distance between two points. Average linkage calculates an average of the distance between all points between the clusters



**Figure 2.2 Cluster Linkage**

*Source* [*http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html*](http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html)

#### 

#### **Data standardization**

Center and standardize

1. Center: subtract from each value the mean of the corresponding vector

2. Standardize: divide by standard deviation

Center and scale with the scale () function

3. Center: subtract from each value the mean of the corresponding vector

4. Scale: divide centered vector by their *root mean square* (*rms*):

xrms=1n−1∑i=1nxi2−−−−−−−−−−−√xrms=1n−1∑i=1nxi2

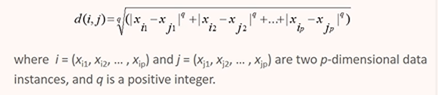
#### 

#### **Distance Methods**

Following distance methods can be used to cluster group of records or data points into multiple buckets so that a cluster of points are closer to each other and the distance between the clusters are far apart to make a clear distinction of various subgroups in the dataset. For text documents, every word is treated as an attribute or column, and every row is a text document that we want to group them.

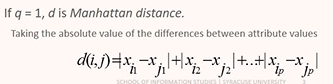
**Minkowski distance**

Minkowski distance between two points or rows can be explained using the following mathematical equation



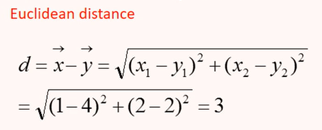
**Manhattan distance**

When q=1 in the Minkowski distance equation, then the equation derives Manhattan distance



**Euclidean distance**

When q=2 in the Minkowski distance equation, then the equation derives Euclidean distance. This is one of the most popular distance measure techniques used in a wide application.



**Cosine similarity**

Cosine similarity is the angle of measure between two points from the origin which is a different way of accessing the distance between rows or points and the equation to calculate the angle is as follows

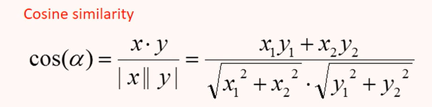
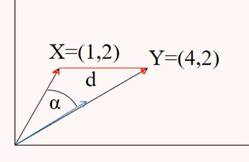


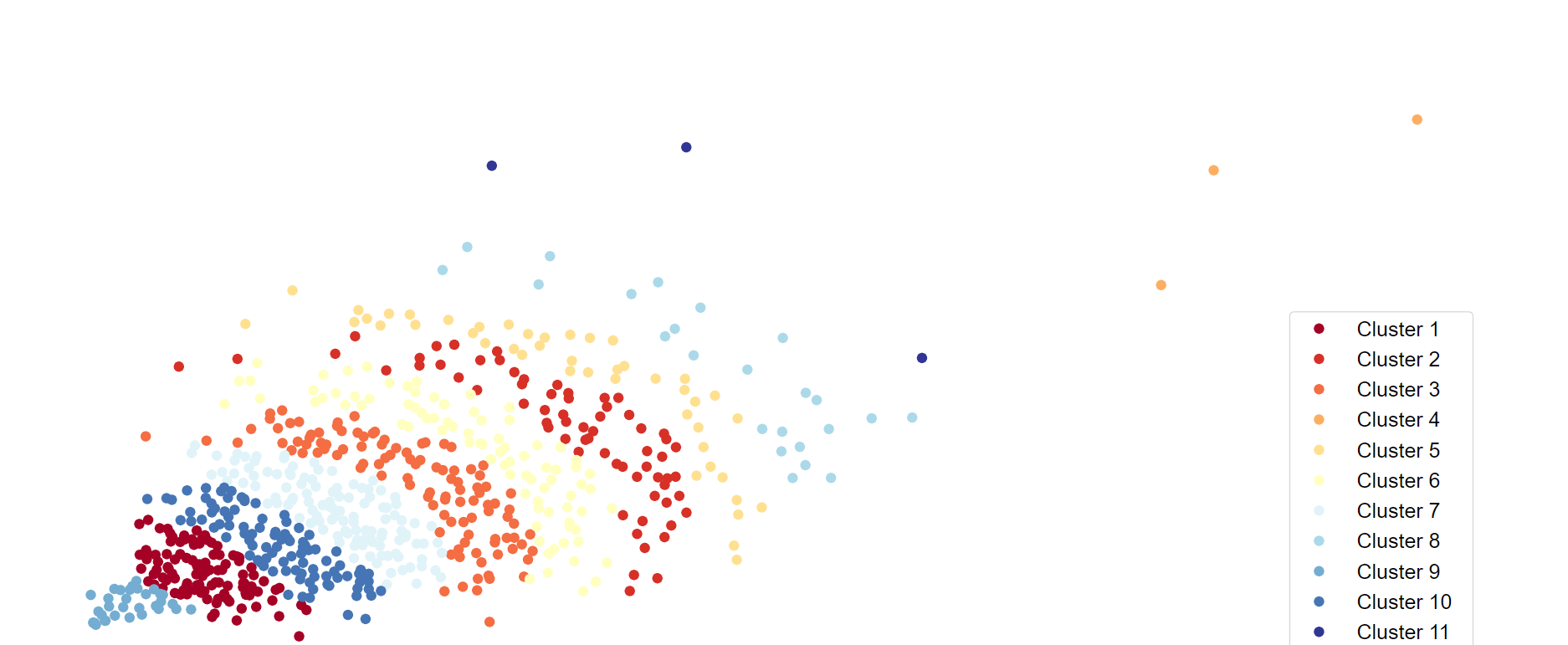
Figure 2.3 illustrates distance and angular measure between two points



**Figure 2.3 cosine angle measure**

**Analysis**

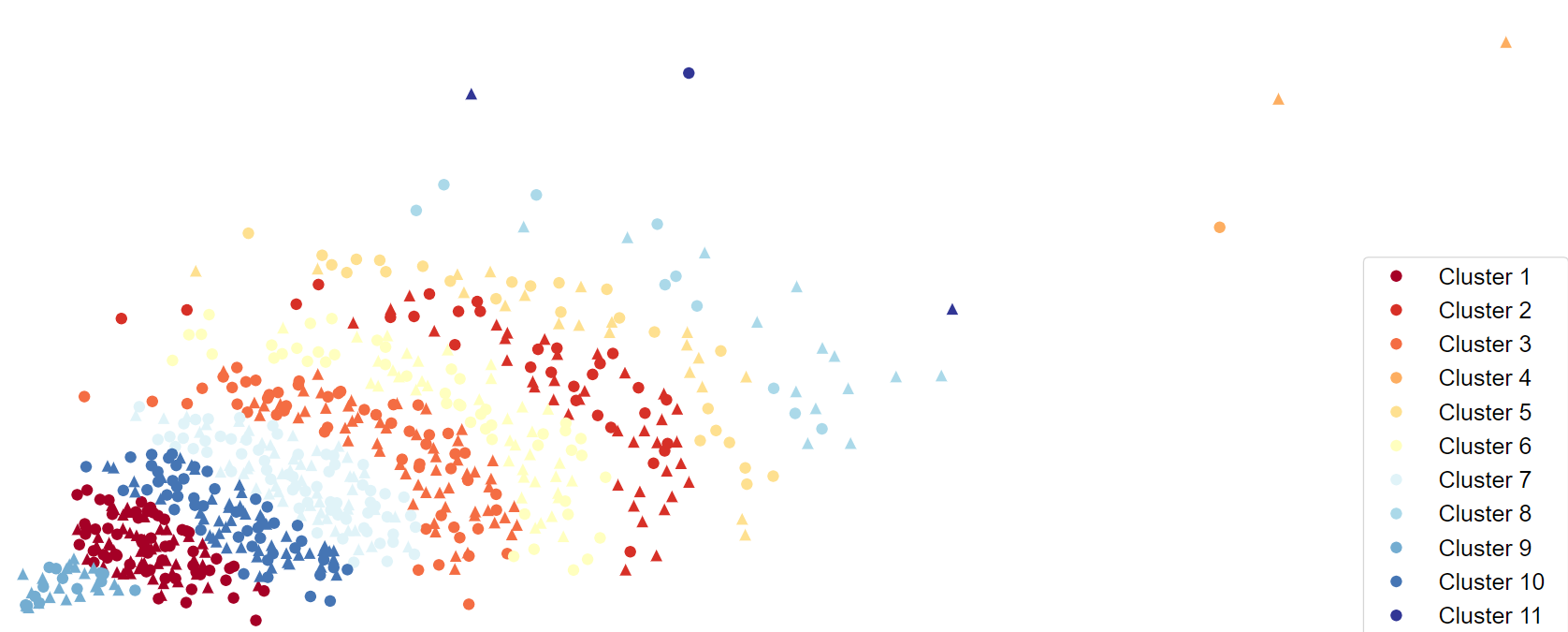
K Means Clustering performed using Word Frequencies and Euclidean Distance method and with a value of K=11. Below figure 2.4 shows the distribution of TED Talks and the clusters that are formed with the euclidean distance measures.

****

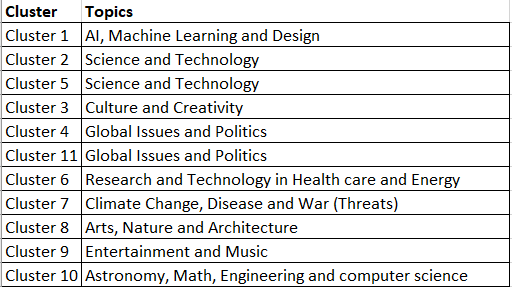
**Figure 2.4 TED Talk Clusters**

**Results**

In Figure 2.5, Clusters are marked against popular and unpopular TED talks. It appears that popular and unpopular talks are well distributed across all topics with some bias to topics related to Technology and neuroscience. Also Table 2.1 shows the topics that are arrived from these clusters based on the majority of the TED Talks appearing in each cluster in a specific areas of interest. Even though the clustering analysis did not yield a perfect topic focussed on each cluster. It grouped TED talks which are closely associated to each other based on the dimension of word frequencies appearing on these talks.

****Popular Talks  Unpopular Talks

**Figure 2.5 TED Talk Clusters for Popular and Unpopular Talks**

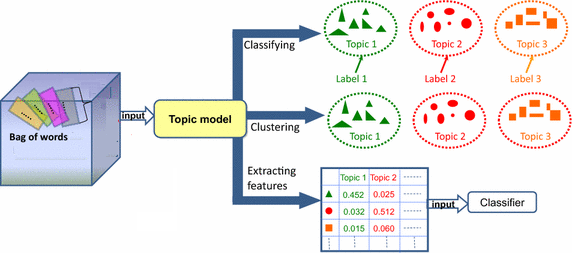


**Table 2.1 Clustering Topics**

## **TOPIC MODELING**

**Analysis**

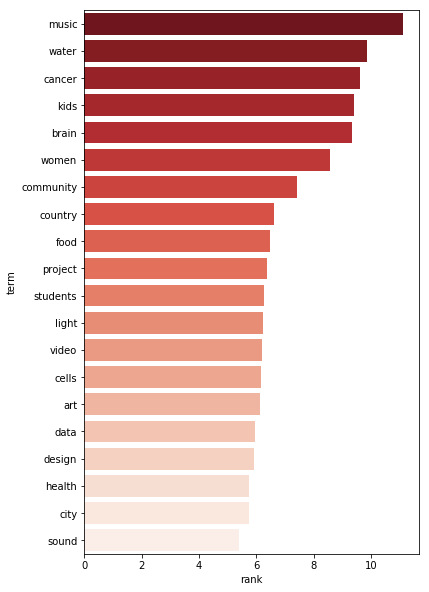
Topic modeling is a type of statistical model for discovering the abstract “topics” that occur in a collection of documents.



The analysis starts with reading in the ted\_clean csv file and saving it as a dataframe. The dataframe is split according to the label. Each dataframe is vectorized using TfidfVectorizer. The popular words from each dataframe are analyzed. Some words repeat. The LatentDirichletAllocation algorithm is run for 10 topics. This results in several repeated and insignificant words. The MNF algorithm is run to identify topics. The output is 10 clear topics for each dataframe.

**Results**

**Popular Top Ranked Words**

****

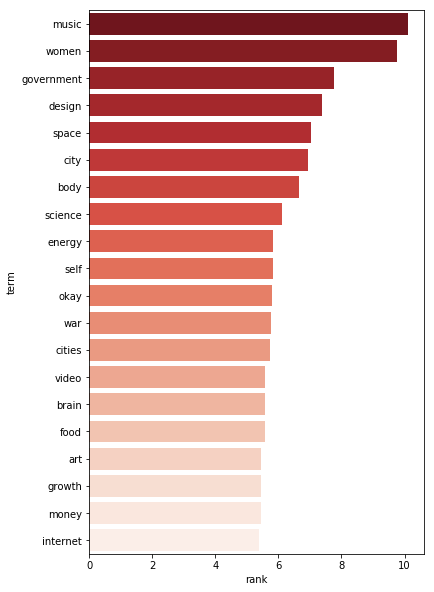
**Figure 3.1 Top ranked words(popular)**

**Unpopular Top Ranked Words**

Repeated words are music, women, brain, food, video, art, design, and city.

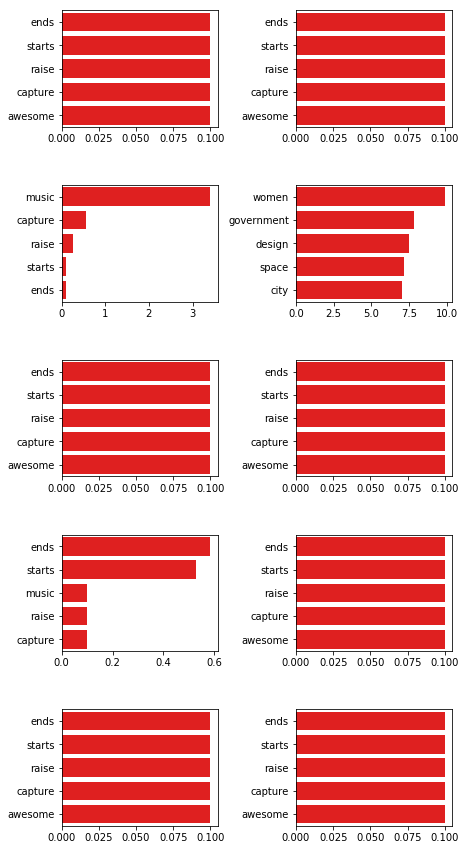
Unique popular words are water, cancer, kids, community, country, project, students, light, cells, data, health, and sound.

Unique unpopular words are government, space, body, science, energy, self, okay, war, video, brain, growth, money, and the internet.

****

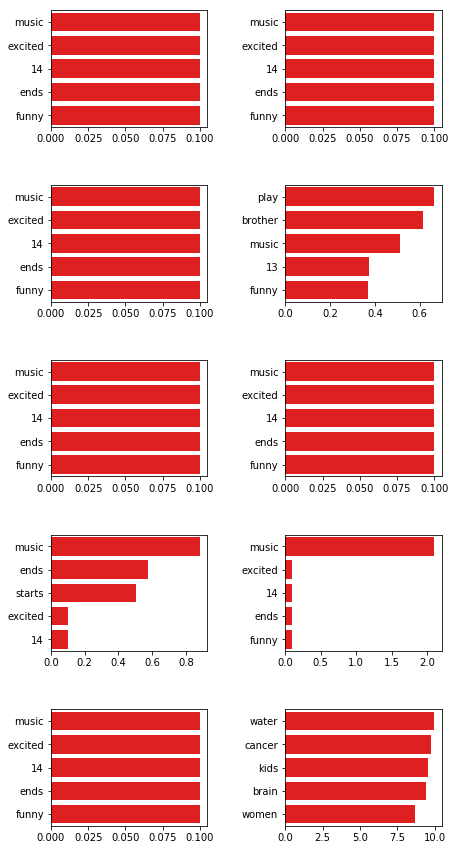
**Figure 3.2 Top ranked words(Unpopular)**

**LDA Popular Topics Top 5 Words**

****

**Figure 3.3 Topics(popular)**

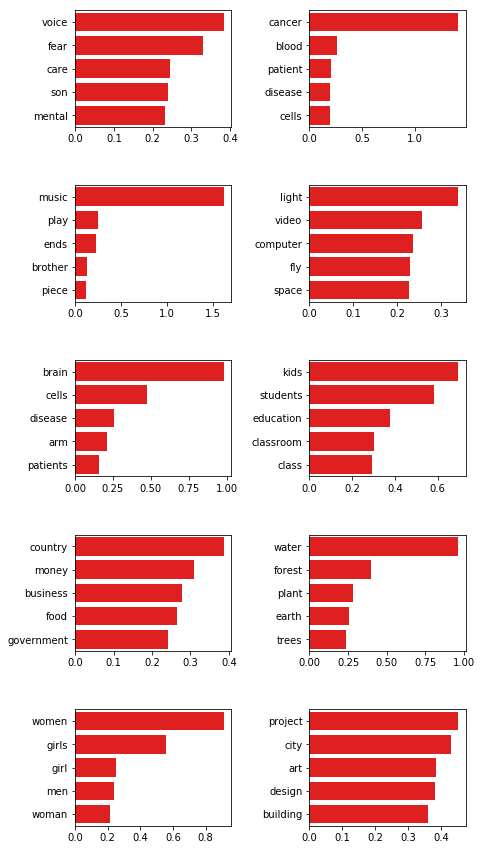
**LDA Unpopular Topics Top 5 Words**

****

**Figure 3.4 Topics(Unpopular)**

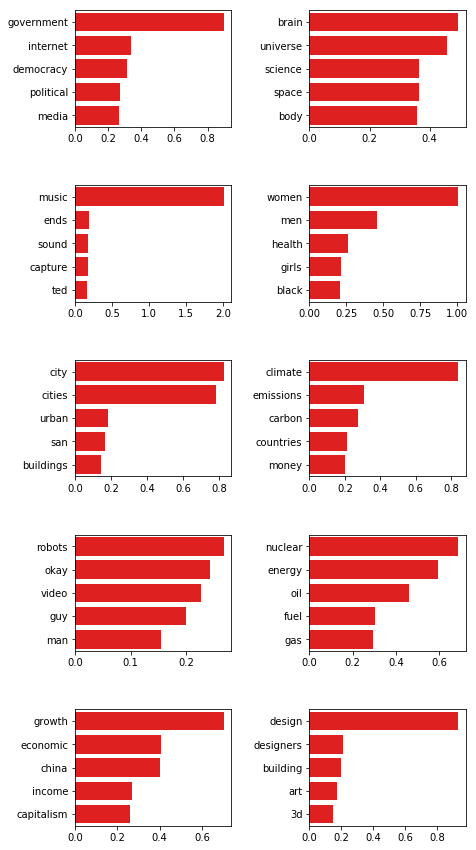
The LDA algorithm was ineffective at differentiating the topics. The first couple topics repeat in both dataframes and insignificant words/numbers are picked up.

**MNF Popular Topics Top 5 Words**

****

**Figure 3.5 MNF Topics(popular)**

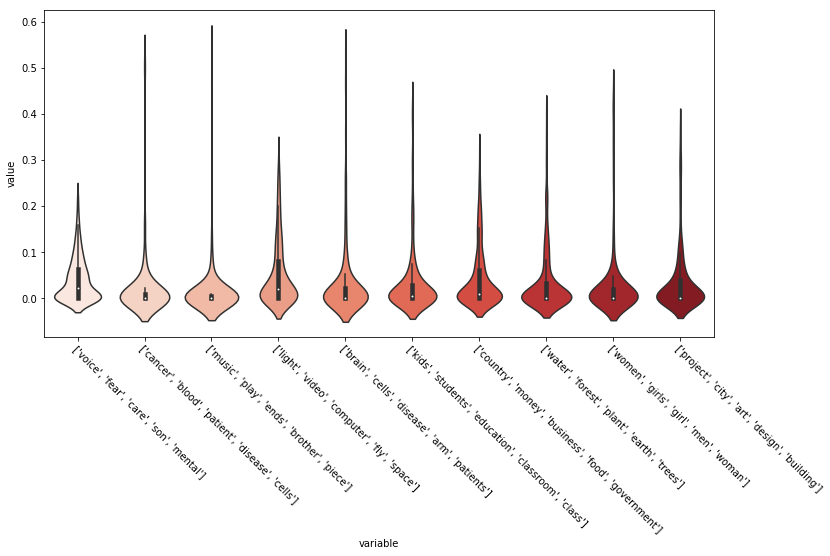
**MNF Unpopular Topics Top 5 Words**

****

**Figure 3.6 MNF Topics(Unpopular)**

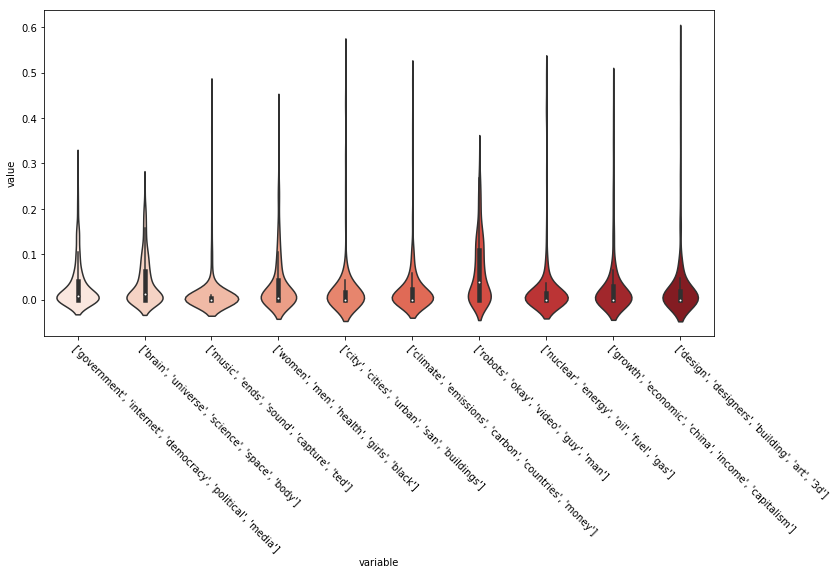
The MNF algorithm identifies 10 clear topics for each dataframe. The top words that were repeated in the first analysis are differentiated.

**Popular Topics Distribution**

****

**Figure 3.7 Topic distribution(popular)**

**Unpopular Topics Distribution**

****

**Figure 3.8 Topic distribution(Unpopular)**

Violin plots are similar to box plots, except that they also show the probability density of the data at different values.

## **SUPERVISED CLASSIFICATIONS**

Five different models were created using three supervised learning algorithms:

i. Multinomial Naïve Bayes

ii. Binomial Naïve Bayes

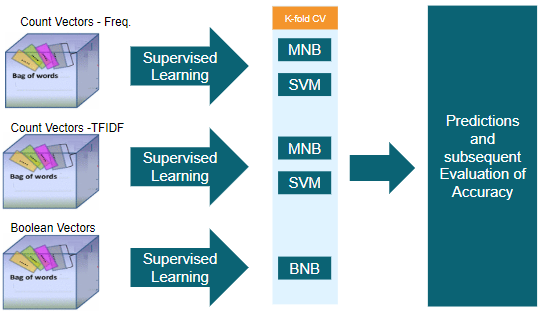
iii. Support Vector Machines

Data was vectorized in the following ways for creating the models:

i. Sparse matrix with frequency counts

ii. Sparse matrix with TFIDF values

iii. Sparse matrix with Boolean values



Following models were created with five and ten-fold cross-validations respectively:

i. Multinomial Naïve Bayes Model with Frequency Count (MNB-FC)

ii. Multinomial Naïve Bayes Model with TFIDF Values (MNB-TFIDF)

iii. Binomial Naïve Bayes Model with Boolean Values (BNB)

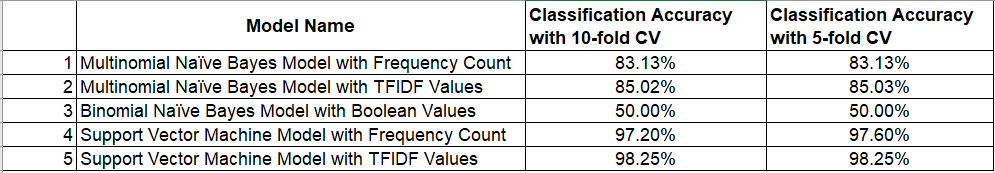
iv. Support Vector Machine Model with Frequency Count (SVM-FC)

v. Support Vector Machine Model with TFIDF Values (SVM-TFIDF)

K fold cross-validation method was chosen to avoid the risk of overfitting. Initially, models were generated using 10-fold cross-validation. Subsequently, a new set of models were created using five-fold cross-validation to see models with comparable accuracies could be generated with half the computation time.

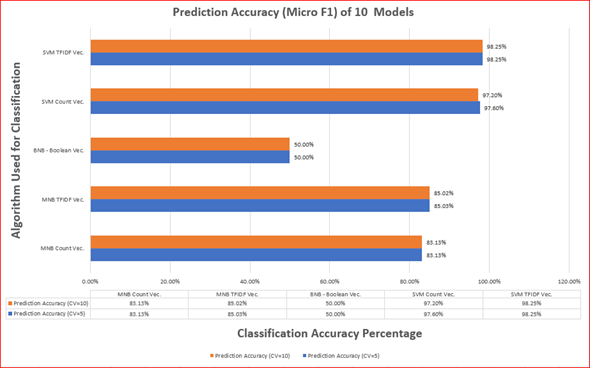
## **COMPARISON OF CLASSIFICATION MODELS**

The table below provides the classification accuracies obtained using the models described above.

****

**Table 4.1 Accuracy of Models**

Note that no major difference was found between the accuracies of models generated using five and ten-fold cross-validations. In fact, they were nearly identical. The bar graph below shows the same.



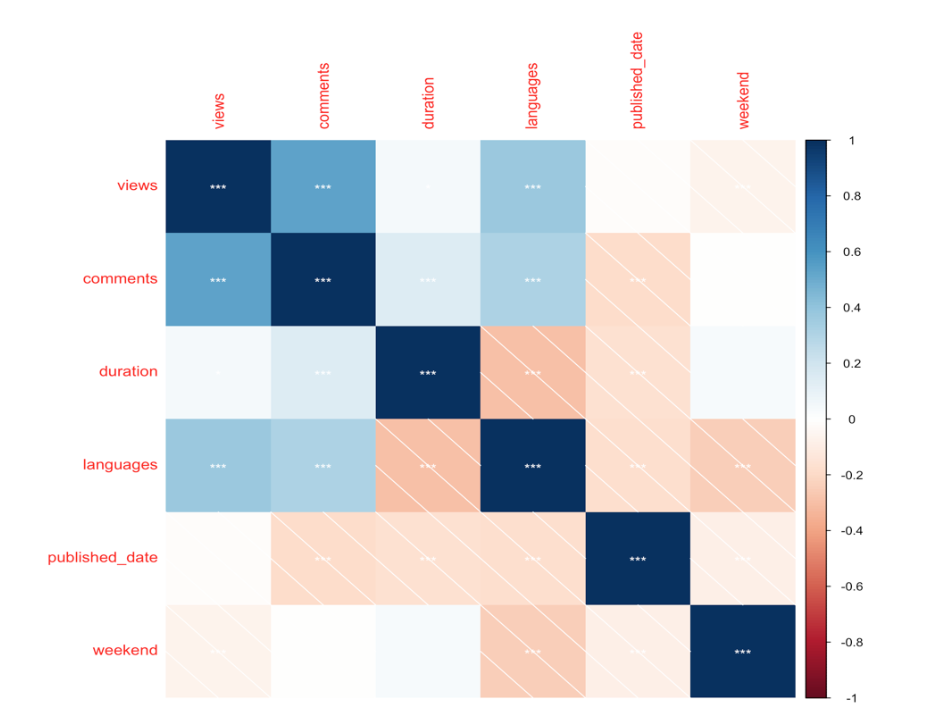
**Figure 4.1 Accuracy of Classification Models**

## **CONCLUSION**

What makes [TED Talks](https://www.ted.com/) so great is that they are given by some of the top people in almost every field imaginable and they are restricted to be very short. Most speakers only have ten to fifteen minutes to talk. Every TED Talk is free to watch and download. TED Talks wants to find, organize, and share ‘ideas worth spreading’ for decades to come. For that reason, this analysis modeled and predicted the most popular talks from the past decade. TED can focus on posting the most interesting talks. Having said that, it is recommended that TED talks should be posted on Fridays, and TED can be confident that subjects related to music, cancer research, education, and etc. may garner decent views. While these insights have provided an invaluable summary about TED Talks produced within the last 10 years, continued analysis must be revisited to monitor changing trends.

The correlations of a Popular TED Talk are shown in **Figure 5.1** and its features are listed are as follows

* High number of comments.
* Translations in many languages.
* It shouldn’t be too short .
* Higher number of tags, ideally between 3–8!
* It would be uploaded on a weekday, preferably a Friday!



**Figure 5.1 Correlations but not the causation of popular TED Talks**