**Online Educational Video Recommendation System Analysis**

**Aarushi Siri Agarwal, Vellore Institute of Technology, Chennai, India**

**Urmila Singh, Vellore Institute of Technology, Chennai, India**

**ABSTRACT**

Most online platforms which provide video content , including TEDx, usually use various recommendation systems to gather more viewers. These videos are recommended based on various criteria. they can be either based on the user behavior and history of watched videos or on the basis of generally liked videos. The aim of this project is to conduct an in-depth analysis of the education platform called TEDX. This analysis will help in deriving the current protocols and thresholds this platform follows to curate and recommend videos to new users of the platform. The end goal is to figure out the various correlations between different parameters pertaining to these videos and on this basis to derive concrete illustrative representations of said relations and also to build a framework around these facts to find the exact relation between various videos on the platform.

**KEYWORD**

online video platform , recommendation system analysis, text clustering , data visualization and analytics, PCA and k-means algorithms

1. **INTRODUCTION**

Recommender Systems (RSs) are characterized by the capability of filtering large information spaces and selecting the items that are likely to be more interesting and attractive to a user. Recommendation methods are usually classified into collaborative filtering methods, content-based methods and hybrid methods . Content-based methods, that are among popular ones , suggest items which have content characteristics similar to the ones of items a user liked in the past. For example, news recommendations consider words or terms in articles to find similarities.

A prerequisite for content-based filtering is the availability of information about relevant content features of the items. In most existing systems, such features are associated with the items as structured or unstructured metainformation. Many RSs in the movie domain, for instance, consider movie genre, director, cast, (structured information), or plot, tags and textual reviews (unstructured information).

For generations, humans have looked outwards at others for inspiration, individuals of great intellect and accomplishment have been regarded as heroes and model citizens. However, this kind of admiration has not always been shared equally amongst the diverse set of people in the world. As such, with growing awareness, one of the greatest questions of the 21st century is the question of bias, such as gender equality. By investigating user ratings of speakers in TED talks, a popular lecture series that are widely shared for approachable discussions from various experts, we aim to explore whether even experts are affected by bias and whether a platform primarily served to distribute educational content is subject to potential unconscious biases as well. We also aim to explore whether certain types of content are systematically more recommended than others.

1. **Literature Survey**

Content-based RSs create a profile of a user’s preferences, interests and tastes by considering the feedback provided by the user to some items together with the content associated to them. Feedback can be gathered either explicitly from users, by explicitly asking them to rate an item , or implicitly by analyzing her activity. For instance, in the movie domain, the features that describe an item can be genre, actors, or directors. This model may allow content-based recommender systems to naturally tackle the new item problem. Other families of content-based RSs use semantic analysis (lexicons and ontologies) to create more accurate item representations.

In the literature, a variety of content-based recommendation algorithms have been proposed. A traditional example is the “k-nearest neighbor” approach (KNN) that computes the preference of a user for an unknown item by comparing it against all the items known by the user in the catalogue. Every known item contributes to predicting the preference score according to its similarity with the unknown item. The similarity can be measured by typically using Cosine similarity. There are also works that model the probability for the user to be interested in an item using a Bayesian approach , or use other techniques adopted from IR (Information Retrieval) such as the Relevance Feedback method.

1. **Feasibility Study**

The analysis parameters for an online video recommendation system could be of any form: title, main speaker, film date, opening words related to the videos. Here, we specifically study and build our model over our own dataset. Some of the following areas of motivation:

* General application of speaker recommendation to a particular user.
* What are the mostly used words by speaker which draw in the viewers leading to high video recommendation?
* what is the possible bias in the content online and its recommendation system?
* Is the recommendation based on the speaker's popularity or is it based on herd psychology which can be seen through the number of views and comments?

#### The most feasible solution to go through the proper analysis is to check for herd psychology in the recommendation system or if the system is a random jumble of latest uploaded videos first and then gets adjusted based on the user viewing habits. To find out which of the two it is we can look at the various parameters that can point towards herd psychology and if the recommendation system is not based on our assumptions, we can say that it is recommending based on individual tests. This is of course assuming we collect the data for various users and both experienced and new to the platform.

1. **About The Dataset**

We will use the “TED Talks” dataset collected by us using a web scraper program. The dataset contains rich information from TED.com on all video and audio recordings of TED talks uploaded on the website until till date. This dataset includes information including speaker name, title of discussion, number of views, number of comments, related videos, and user ratings. The dataset also contains a list of “related talks” to watch for each TED talk according to the recommended talk links on the actual website for each talk.

1. **Feature components for analysis**

If we can somehow analyze the different features of each video on the platform and sort and compare them based on their frequency of recommendation. we can find the algorithm or technique used for these recommendations and may even be able to suggest better recommendation systems for the platform to draw in more people.

Let's take a dummy example and first analyze what could be the possible online educational video related parameters that are stored and on the basis of which we can analyze the data.

every online video database based on educational videos must contain the following based on common collectible parameters:

##### video publishing date

##### views

##### comments

##### duration

##### title

##### category

##### speaker profession

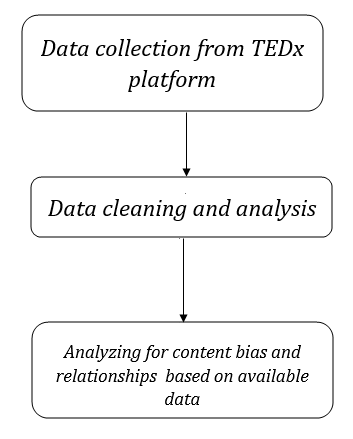
##### speaker details

based on this we can say that the video collection and correspondingly its recommendation may be biased on the following basis:

1. based on profession of speaker
2. based on speaker gender, region , religion, etc.
3. based on topics or category of video - for this we can use clustering algorithms like k means
4. based on time of upload of video content

##### To check if recommendation is biased in these aspects, we sort the videos based on recommendation frequency and correspondingly check the distribution of the videos based on the three categories. We also can see the distribution and pattern based on the number of views and comments which can play a big part in recommendation systems as it is possible that the video liked by most people will also be liked by new users of the platform and hence increase the number of platform users.

1. **Design and flow of models**



**Fig.1 design and flow of model**

for the analysis we have used the following modules and analysis parameters :

1. **module 1 : data collection**

In this module we have made the URL links text file and then downloaded the meta data in the URLs to extract video information from.After getting the metadata we extract the information for each video and store it in csv file format . the meta data includes the data about the video title, views, comments ,tags, related videos(videos usually recommended alongside the given video, duration, publishing date ,transcript and also the speaker details like speaker name and speaker profession. Not just this we can also collect the URL details , whether it has subtitles , how many people have downloaded the video and a lot more.

1. **module 2 : data cleaning and dataset analysis**

Once we have the data, we need to clean it and analyze what data we were able to collect. After this we can easily plan which parameters to analyze to find out the recommendation system used on the platform.

1. **module 3 : analysis and visualization from dataset**

The attributes from the obtained data set are compared with each other to find correlations and dependencies and then these are visualized using different types of graphs. The graphs obtained are then studied in detail to obtain further observations to see to what factor each attribute influences the recommendation system and which attributes are dependent and independent.

* 1. **based on views and comments**

The relation between the views and the comets and their effect on the recommended videos is checked . Also, the effect of more views and less comments usually signifies the possibility of hate comments or dissatisfaction with the talk and hence if it affects the video recommendation is checked.

* 1. **based on profession**

the distribution of videos based on the speaker profession is first checked. Then it's seen how the videos are recommended based on the already established idea of whether views play a central role in the recommendation system.

* 1. **based on content category or talk category**

since we were able to get the transcripts we can use TF-IDF to first get the unique and important words used in the entire collected video data and to convert this unique word dataset in to numerical form. following this we can use algorithms like PCA ( principal component algorithm) to reduce its dimensionality which can be used in clustering algorithms to find out the various clusters the uploaded videos belong to base on the content of the talks. We can check the number of categories of videos uploaded by checking the silhouette scoring for different number of clusters formed using kmeans.

**PCA algorithm**

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges, which will lead to biased results. So, transforming the data to comparable scales can prevent this problem.

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data. Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.

principal components represent the directions of the data that explain a maximal amount of variance, that is to say, the lines that capture most information of the data. The relationship between variance and information here, is that, the larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has.

**K-means**

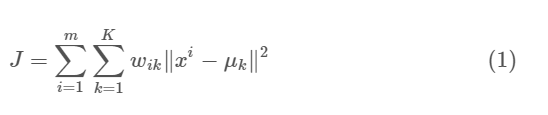
**Kmeans** algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

The way kmeans algorithm works is as follows:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids.
4. Compute the sum of the squared distance between data points and all centroids.
5. Assign each data point to the closest cluster (centroid).

Compute the centroids for the clusters by taking the average of the all-data points that belong to each cluster.

The objective function is:



**6. Risk Analysis:**

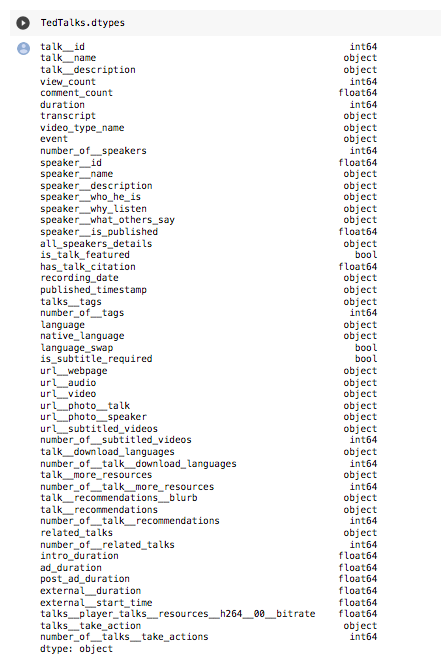
Due to the vast amount of data in online video content platforms like TEDx gathering accurate data on videos on all fields might be tough. Hence, we have taken the data on videos recommended and watched only for a particular year and not over many years, but this can affect our analysis for bias based on time of publishing hence we have not analyzed based on time of upload or publishing of video. There can be several parameters which can affect the  bias created in the data and the videos recommended. Some of few can be the profession, the type of content, the gender, the age of the audience for the content and hence we restrict ourselves to several parameters for data analysis and visualization. Also, we might not get other datasets besides TEDx platform for comparing our analysis with recommendation systems used on educational online video platforms.

**7. IMPLEMENTATION**

7.1 First we import modules that can help us collect the data from the TEDx platform .

**Importing all the libraries and modules**

First import the libraries to better analyze the data set. here matplotlib is basically used for visualization and word cloud.

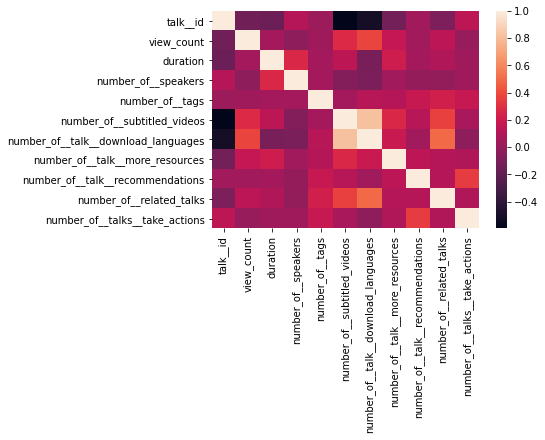
**2.Dataset analysis** 

**Before reducing dimensionality**

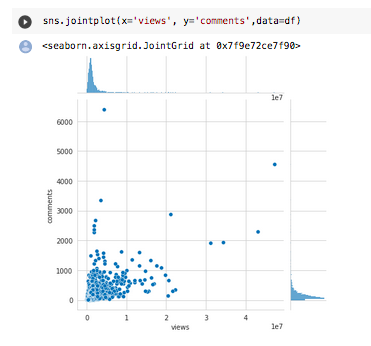


**After reducing dimensionality**

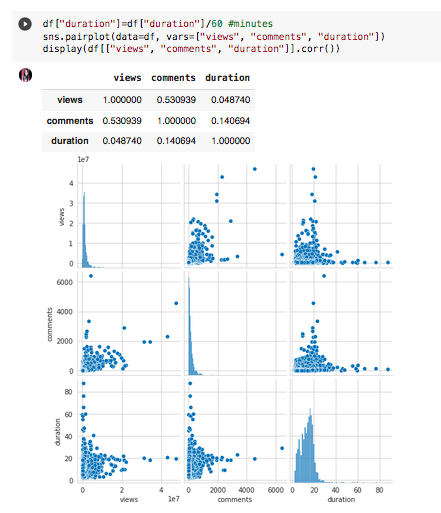
We also change the dimensionality of the dataset to include the fields we need for our analysis. Then analyze the correlation of each column or parameter in our dataset with the other parameters using heat map .

**Heat map**

1. **correlation between view and comment count**

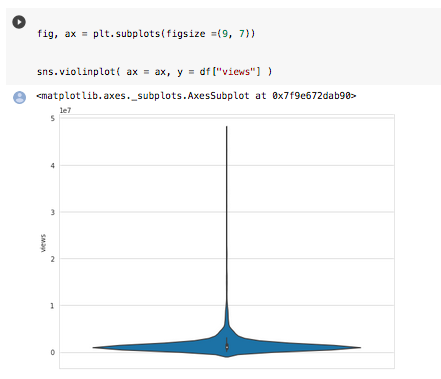
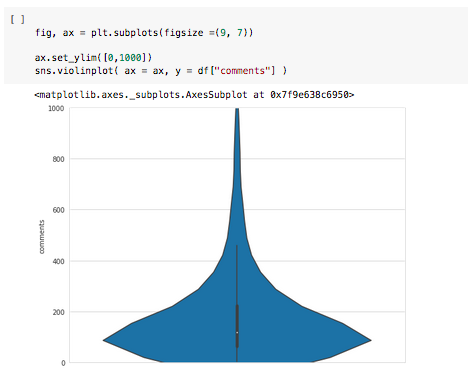
it is assumed that there is always a high possibility of views and comments being used in recommendation of a video. Those parameters and correlations are checked first.

It is seen that there is a strong correlation between the views and the comments as expected . Another common recommendation factor can be the correlation with the time duration as more the time duration less likely that the user be patient and listen to the whole talk without getting bored. hence, we also analyze based on the talk duration.

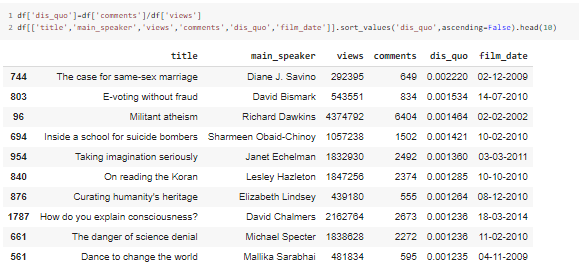


**visualizing views , comment, duration**

On analysis it was found that the assumption was wrong as the time duration parameter does not strongly correlate with the views and comments. After finding how strongly correlated the views and comments are . it is considerable that they might be strong parameters used for recommendation systems . to check this their distribution is seen based on the sorted videos list. For this we use violent plot. The violent plot a method of plotting numeric data. It is similar to a box plot, with the addition of a rotated kernel density plot on each side. Violin plots are similar to box plots, except that they also show the probability density of the data at different values, usually smoothed by a kernel density estimator.

**violent plot for views and comments distribution**

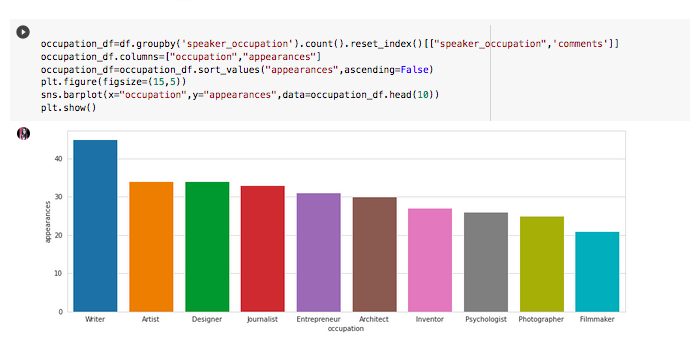
The distribution shows that the number of comments and views are more for the highly recommended videos.



**Visualizing based on difference in views and comments**

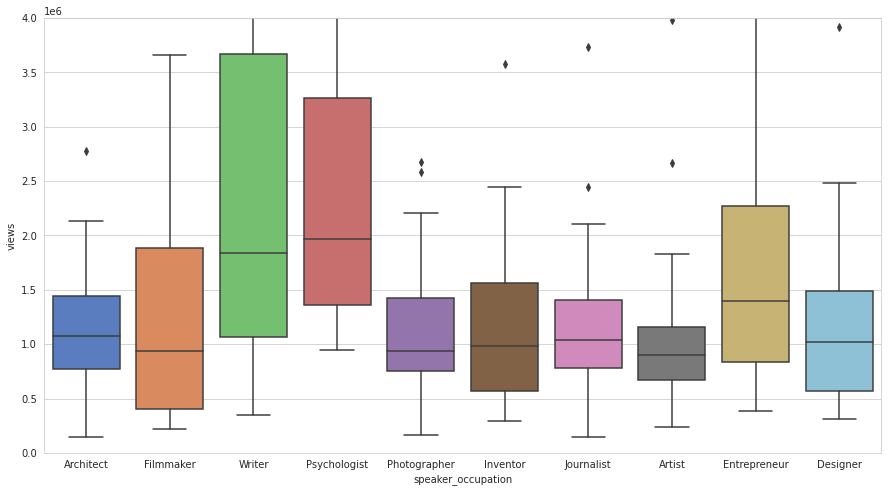
The distance quotient shows the difference between the views and the comment higher the distance less recommended is the videos . This is because the videos are sorted based on their recommendation and it's seen that the higher distance quotient videos have a lower serial number or id in the sorted list of videos.

1. **based on the speaker profession**

Often, we can see that the online platforms don't have equal content in different categories or if they do, they might be biased based on the type of content that is recommended . one parameter where such bias can be seen is in the speaker profession. 

**visualizing bar plot for profession**

Based on the figure above we can tell that there is no bias in the content online based on the speaker profession as there is nearly equal amount of content in each speaker profession(around 20-40 talks each).

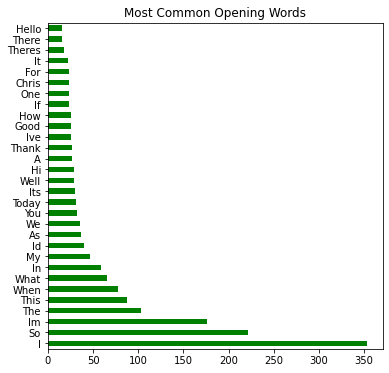
****

**Views for videos of different profession**

Just analyzing based on appearance gives us the count of videos in each profession but we also need to know its distribution based on the number of views since we had confirmed that views play a central role in recommendation of a video. The table shows that the professions with more videos like writers have the highest number of views but also that the profession like artists despite having a high number of videos does not have a high number of counts and is not recommended as much as videos on talks by psychologists. This might be due to people tending to watch talk by psychologists more than artists. This signals the possibility of a herd consciousness followed by recommendation as well. Since people tend to watch more talks given by psychologists than artists over time, talks on the former receive more views and comments and hence get recommended more.

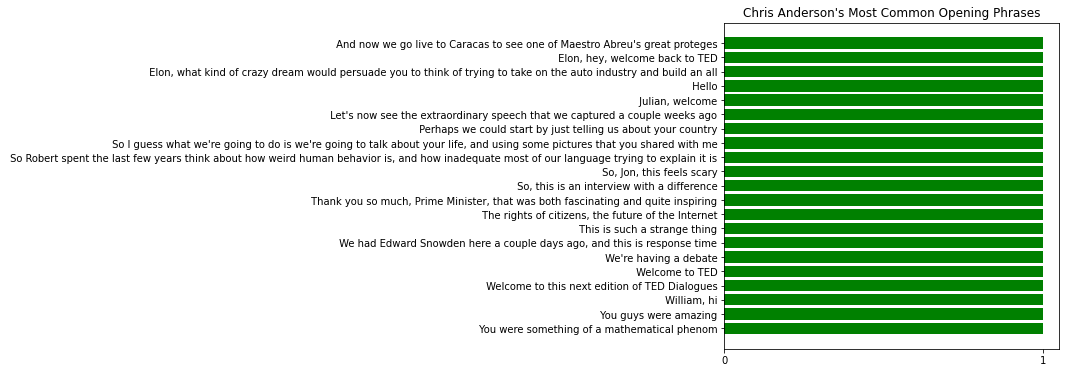
We can see in the above data that has been analysed that in almost all the scenarios, there has been a bias created on the videos with maximum recommendation and the type of content curated by the performer for its audience. For example, a psychologist who has shown up the second least number of times has got the views and likes(recommendation) equal to a writer’s, who has shown up the maximum number of times. Hence, it can be concluded there are a lot of factors affecting the analysis as well.

1. **based on speaker**

Most of the recommended videos show similar opening styles of the speakers. Also, it is general knowledge that the TEDx talks are known for their speaker openings . Hence, we have tried finding what kind of openings are getting more views. better the opening more its views and hence more recommended the video is. 

The most common words used in speaker openings can be seen Right away, I can see a number of "classes" of openings. There are the opening thanks, the time-of-day greeting ("good morning!"), and more than one person who starts with Number One ("I am..."). My eventual goal is to classify all of these openings, even if there's a class of "unique" openings at the end. First, let's just see how many different ways there are to start a speech.

Part of being a great speaker is innovating your openings and creating an engaging hook. Yet, more basic than that is simply greeting your audience. Clearly this graph shows that you should not forget to greet your audience first. Also grab the audience with something that will cause them to pay attention. Phrases like "I have a confession to make", "Let me tell you a story", and "I have a question" are all great candidates to get you started.

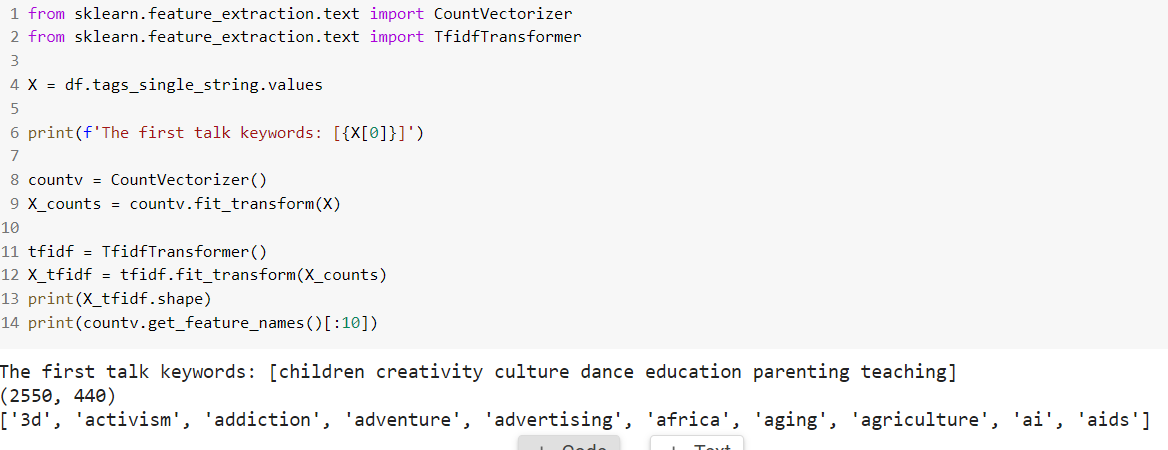


He's never repeated himself twice in 21 speeches. That's not really that surprising.

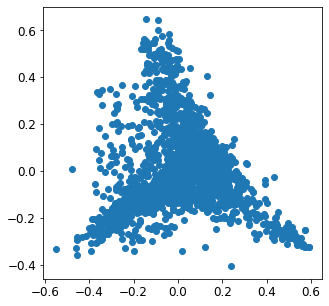
In fact, he seems to not really be making speeches at all. These are more like introductions for other speakers. An interesting find for non-regular TED talk viewers, but not very insightful for our purposes.

1. **based on content category**

since the videos are not divided into the type of content, they talk about the tags from video metadata collected earlier is used to extract the unique and important words used in the talks this is done by getting the TF-IDF score for words individual videos and then taking a select number of the words. then use variance and similarity between these words to get the different content category clusters for their classification. Once classified it is possible to determine any bias based on content category for the videos online. Also, later on we make a tag cloud of these words to see the most used words in a particular category. To analyze if the videos are recommended equally for all categories and if there are enough videos in all categories a common tag cloud can be made and the word size which correspond to its occurrence in the dataset can be seen as representation of equal contribution or talks in all categories.

 transformed keyword matrix (X\_tfidf) has 2550 observations, matching our original data frame, and 440 features, each of which correspond to a unique word. There are a total of 440 unique words across all the talk keyword sets.

A common first step in unsupervised learning is to reduce the dimensions of your data. This will allow us to plot the data in 2 dimensions to visualize it and get a sense of what is going on. Principal component analysis (PCA) is a method which finds the axes of highest variance in the data and rotates the data so that the first dimension corresponds to the greatest variance, the second dimension corresponds to the second greatest variance, and so on. Running PCA on the data it’s possible to visualize the distribution of talks in this dimensionally reduced PC space.

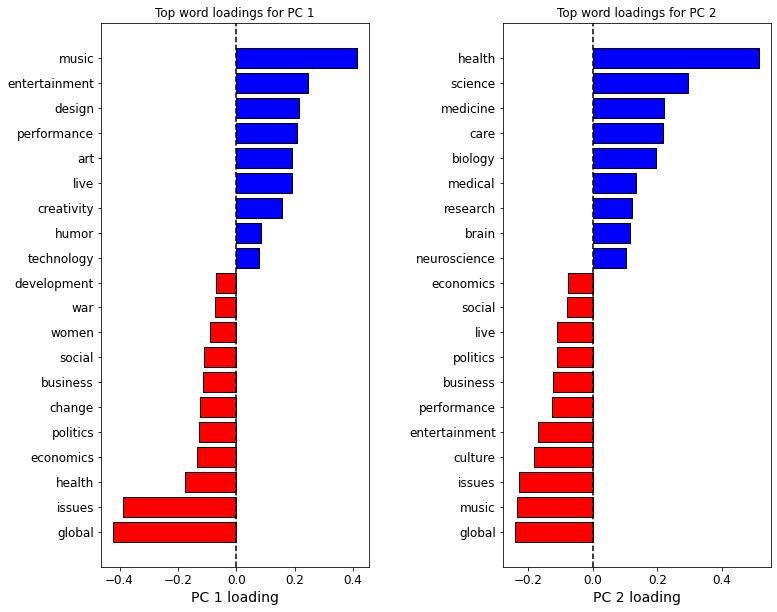
****

**all plotted values for unique words in videos**

It looks like the data points form a central cluster with 3 vertices.

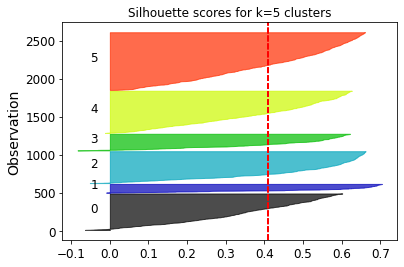
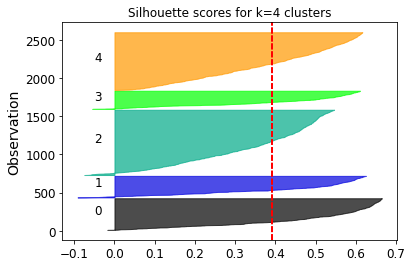
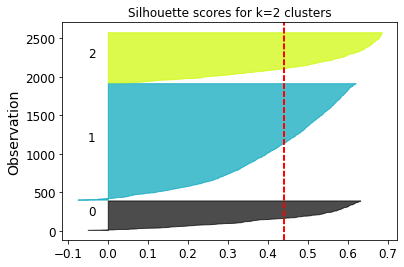
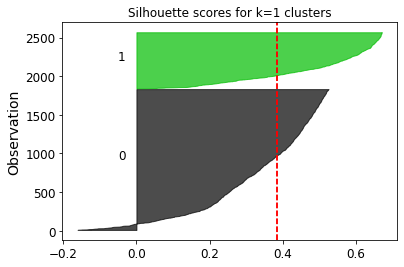
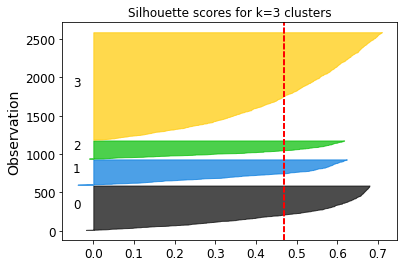
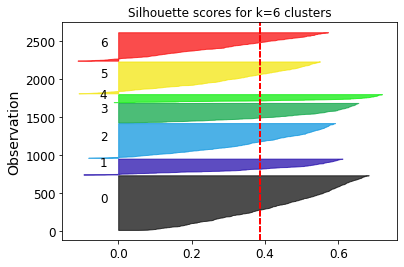
The PC1 and PC2 represent the new dimensions which were found by PCA to be a linear combination of the individual words. The loadings of each PC correspond to how much each word in our original vocabulary contributes to each principal component.

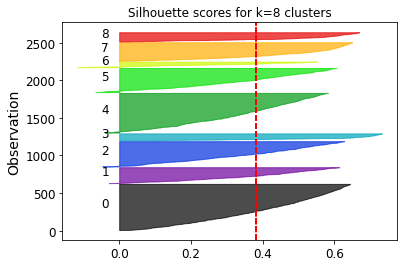
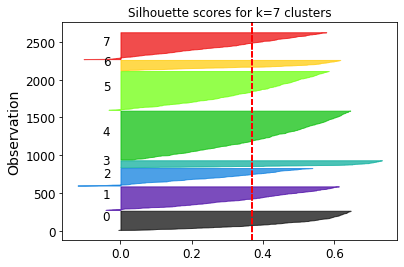
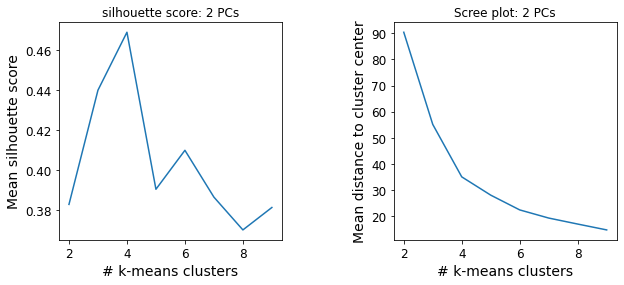
the highest and lowest loadings in PC 1 and PC 2 can be visualized using horizontal bar plot to understand what these first two principal components represent in the data.

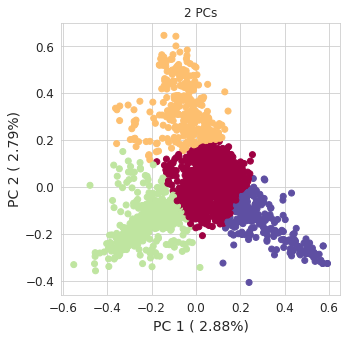
****

The bar plot shows that there are three possible clusters that can be formed. One pertaining to politics , second pertaining to art and third pertaining to medical field. Also, in both PC1 and PC2 around the middle where x axis value is 0, we see topics related to technology and science.

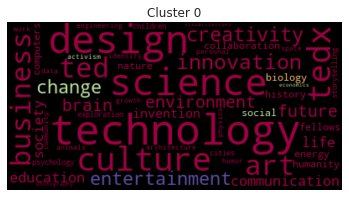
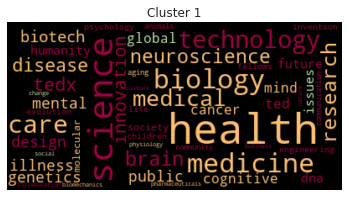
To cauterize and categorize the videos an unsupervised clustering algorithm is required . this paper uses kmeans clustering for this purpose.K-means clustering is an unsupervised learning algorithm that finds k clusters in a data set by minimizing the Euclidean distance of each point to its assigned cluster center.

But there is a problem of how to set the number of clusters, k, in the K-means algorithm. One way is to find the average silhouette score across all observations and choose k such that the silhouette score is maximized. The silhouette score essentially weighs the inter vs. intra-cluster distances. A high score indicates more unambiguous cluster assignments, and low scores indicate ambiguous or incorrect cluster assignments. The silhouette score is defined on [-1, 1]. To get the number of clusters to form different values of number of clusters are used and silhouette scores are analyzed.

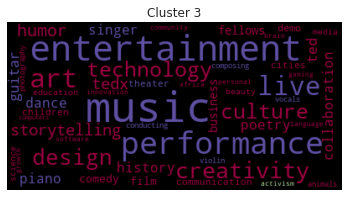
The silhouette score peaks at 4 clusters. Choosing k is not an exact science and could also be informed by our prior knowledge of the expected number of clusters based on PC1 and PC2 top words analysis.

****

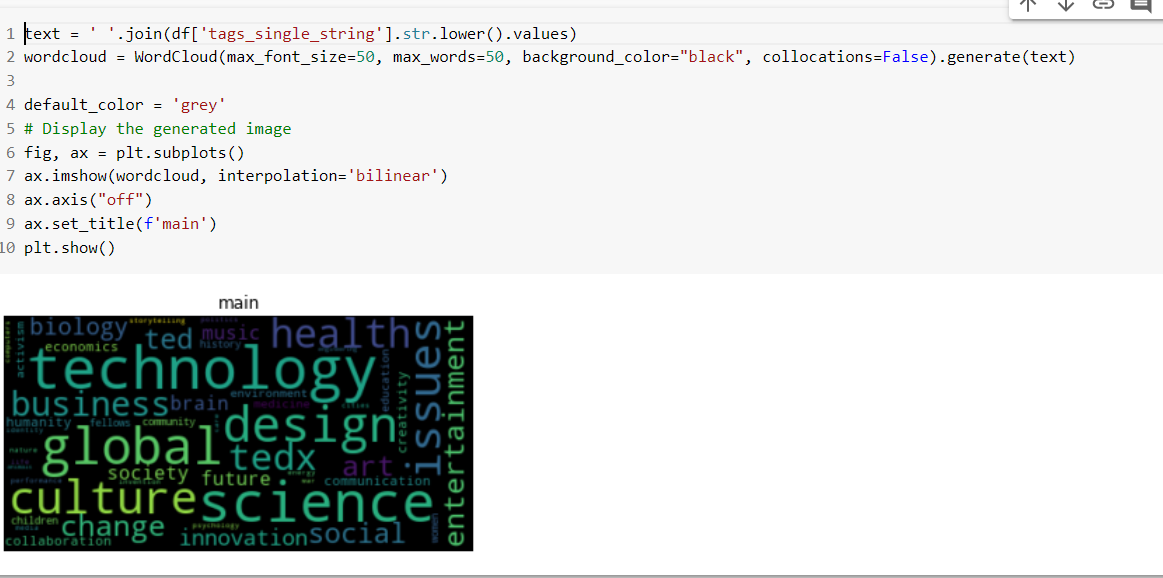
**Kmeans clustering for k = 4 on PCA output**

K-means has essentially divided the observations based on physical location in PC space. Now word clouds for each cluster can be used to see what topics each cluster roughly show.





In the word clouds above each word is colored by the cluster in which it is most frequently found. the cluster 0 (colored maroon), which resides in the bottom left region of our PC space plot, contains words like global, issues, and politics, which matches with our understanding of PC 1 from earlier. Cluster 1 (colored peach) resides in the center of PC space and contains words like technology, design, and culture. And clusters 2 (colored green) and 3 (colored purple) reside in the top and bottom right of PC space, respectively. As expected from our understanding of PC space, cluster 2 contains science, and specifically biomedical, related words, and cluster 3 contains words pertaining to art, music and entertainment.



The code and Tag cloud above shows the most repetitive words across all videos on the platform and as assumed the major words pertaining to each cluster are present and in equal size too. Showing that they are equally represented in the platform.

**CONCLUSION**

No matter how different the subjects of the talks are, the common theme of spreading ideas and inspiring people seems to be an adhesive force between them.

TED speeches seem to place a lot of emphasis on knowledge, insight, the present and of course, the people. After analyzing all the important parameters we can conclude that the TEDX platform in recommending Videos have one major flaw and that is herd psychology. This means that rather than recommending videos according to individual tastes and history, the TEDX platform looks at videos which have the most views, comments etc. and on the basis of this recommends videos to each user.

This method thus leads to many users being recommended videos which are not of their interest and thus these videos lose viewership and popularity.

The easiest way of rectifying this situation is to look at individual preferences and accordingly suggest the videos resulting in increasing popularity of that video as well as the platform.

**REFERENCES:**

1) D. N. Beede, T. A. Julian, D. Laongdon, G. McKittrick, B. Khan, M. Doms. 2011. Women in STEM: A Gender Gap to Innovation. Economics and Statistics Administration Issue Brief, 04-11.

(2) A. Boring. 2017. Gender biases in student evaluations of teaching. Journal of Public Economics, 145, 27-41

(3) L. Farrell, L. McHugh. 2017. Examining gender-STEM bias among STEM and non-STEM students using the Implicit Relational Assessment Procedure (IRAP). Journal of Contextual

behavioral Science, 6(1), 80-90.

(4) D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network. In Proc. KDD 2003.

(5) R. Zhou, S. Khemmarat, L. Gao, J. Wan, J. Zhang, Y. Yin, J. Yu. 2016. Boosting video popularity through keyword suggestion and recommendation systems. Neurocompu ting, 205, 529-541

(6) H. Shiokawa, Y. Fujiwara, M. Onizuka. Fast Algorithm for Modularity-Based Graph Clustering. In AAAI, 2013.

(7) Newman, M. E. J., and Girvan, M. 2004. Finding and Evaluating Community Structure in Networks. Physical Review E - PHYS REV E 69:026113.

(8) Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding Community Structure in Very Large Networks. Physical Review E - PHYS REV E 70:066111.