Team 1: Lipsa Jena, Ji Kang, Seyoung Nam, Richie Wang, & Janice Yang
Professor Chirag Shah
IMT 574
14 March, 2021

Background & Motivation

TED, a non-profit dedicated to spreading creativity and new ideas, organizes, advertises, moderates, and publishes "Ted Talks". "Ted Talks" are 10-20 minute talks given by notable figures in specific fields to spread an idea, notion, or exciting new concept to the general population. These talks are not meant to be a substitute for college lectures or to teach the audience about how to solve specific problems but rather inspire them to take initiative, explore new ideas, and/or discover a new passion or hobby. The topics are broad, including things such as technology, entertainment, design, business, lifestyle, music, and so on.

These talks are designed to be relatable, require little to no background knowledge on the topic or domain itself, and entertain the audience while educating them. Coupled by visuals, humor, performances, and/or demonstrations, these talks get hundreds of thousands of views; with the most popular videos often getting tens of millions. Typically the speakers have accomplished many feats in philanthropy or social good in general.

Ted Talks have a phenomenal reach and have a direct impact on the audience majority of the time. An example of this is shown when Sal Khan, the founder of Khan Academy, gave his Ted Talk back in 2011. Prior to Khan's Ted Talk, Khan Academy had approximately ~7 million students but quickly exploded to over 140 million students shortly after his talk. This figure has eventually normalized past the initial attention the talk garnered and Khan Academy still maintains over 100 million students globally in 2021.

Around 2000-3000 people attend each TED conference with several tens of millions viewing the Ted Talks online. It's important to practice, edit, and rehearse the speech prior to giving it of course. The one caveat to speaking on this platform is the live-aspect. Speakers are not able to gauge how successful their talk will be or know how they should revise their speech further. With TED playing a pivotal role in millions of lives through their Ted Talks, our team wanted to research and analyze what makes a Ted Talk "successful". To start out with, our team was provided two datasets, "TED Main" and "TED Transcripts". The available features and definitions for each feature are listed below:

TED Main Dataset:

- **name**: The official name of the TED Talk. Includes the title and the speaker.
- **title:** The title of the talk
- **description**: A blurb of what the talk is about.

- **main_speaker**: The first named speaker of the talk.
- **speaker_occupation**: The occupation of the main speaker.
- **num_speaker**: The number of speakers in the talk.
- **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.
- **published_date**: The Unix timestamp for the publication of the talk on TED.com
- **comments**: The number of first level comments made on the talk.
- **tags**: The themes associated with the talk.
- **languages**: The number of languages in which the talk is available.
- **ratings**: A stringified 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.
- **url**: The URL of the talk.
- **views**: The number of views on the talk.

TED Transcripts Dataset:

- url: The URL of the talk
- transcript: The official English transcript of the talk.

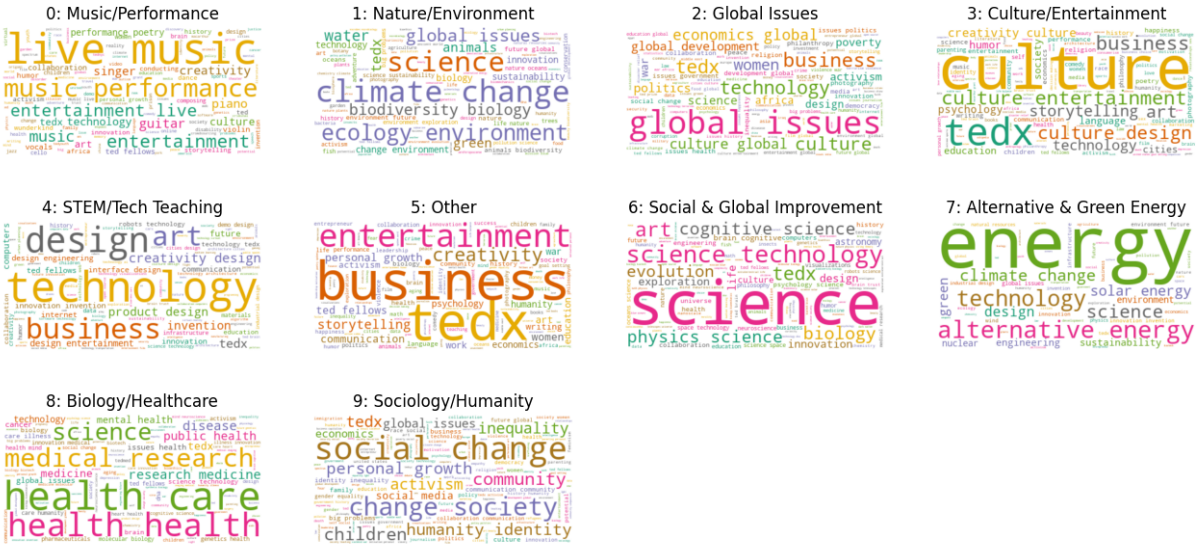
Clustering

Examining a handful of Ted Talks, we experimented with how we should proceed with our analysis. One key insight we've found is that not all Ted Talks are the same; that is, what constitutes a "successful" Ted Talk may be too general of a question. Ted Talks on a niche topic such as quantum computing would get fewer views than a talk on making good lifestyle choices due to their target demographics. The specific reason(s) why one Ted Talk performs well is not the same for every single Ted Talk. Factors that make an art/music performance Ted Talk successful are not the same factors that make a Ted Talk on climate change successful. For this reason, our team decided to cluster Ted Talks into separate categories then attempted to find out the specific features and characteristics that make each type of Ted Talk "successful".

We decided to use KMeans as the clustering algorithm. KMeans is a divisive technique that works in a top-down mode to group the objects based on features into K number of groups. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

We made several experiments before we decided on the clustering feature(s). As most of the columns in the datasets are text, we transformed those data into numerical forms using CountVectorizer and Linguistics Inquiry and Word Count (LIWC) which will be explained in more detail shortly. We applied both tools on columns such as title, description, tags, and

transcript, and we conducted clustering using different combinations of the transformed data. After evaluating the result, we found that using CountVectorizer on “tags” generated the best clusters as we can find similarities in most of the clusters. We also visualized the tags using word clouds to better understand the primary topics in each category. Finally, we were able to conclude a category name for each cluster.



Analysis

For all of our analyses on the clusters, we’ve defined our own “response variable”, ‘success’, which is based on the number of views the Ted Talk videos get. This is considered true if the given Ted Talk has a higher-than-average number of views for that given cluster or “category” and false otherwise. It is important to note that we did not utilize features that’d be unavailable to the upcoming Ted Talk speaker - such as ratings and number of comments in our analysis. We did utilize ratings as a means to confirm our findings at the end.

As previously mentioned we’ve used the LIWC tool in an attempt to cluster, but we’ve also utilized the tool for our analyses. Utilizing 3 separate dictionaries, LIWC identifies and classifies words into categories then calculates quantitative metrics for the text. These include things such as “sadness”, “overall affect”, “verb”, “past/present/future focused”, “word count”, “words per second”, “authority”, “clout”, and so on, gaining quantitative metrics for our qualitative data allowed us to properly analyze certain clusters. If curious, the full list of the 95 available LIWC metrics can be found here.

Cluster 0: Music/Performance

There were 86 records for this cluster. We’ve transformed the transcript into a quantitative representation by removing stop words like “a”, “and”, “but”, “how”, and “or”, then count vectorized them to train a logistic regression model from sklearn library’s “linear_model” module. The response variable was a simple binary yes or no to if the given Ted Talk is

successful or not. Running the count vectorized transcript through the model yielded a 30-run-average accuracy rate of ~0.471. For a binary classification problem, a baseline rate of 0.5 accuracy was expected which shows this approach was not effective for this cluster..

Processing just the transcript proved to be inefficient for this cluster so we utilized LIWC metrics - specifically word count, pronoun, dic (dictionary words), ppron (personal pronouns), and cogprog (cognitive processes). LIWC provides an overabundance of metrics. Removing the metrics with a low correlation resulted in approximately 80 total features. Next, we utilized a sequential neural network from keras.models and dense layers from keras.layers. First, we partitioned the columns into independent variables (X), and our response variable (y), then split into training-testing sets using a 0.3 ratio for test set size.

For this specific cluster, we utilized an input layer of size 80, the same number of features, with a rectified linear unit (relu) activation function. Next, 3 hidden layers, all using the relu activation function, reduce the number of outputs to 65, 27, then 9 respectively. Finally, the output layer outputs a single value using the sigmoid function which is then rounded up or down using Python's built-in round() function to signify true or false. Initial runs have shown accuracy rates as low as 0.5, only marginally better than our logistic regression approach, and as high as 0.83. Taking a 30-run-average yielded an average accuracy of 0.691 in predicting if a Music/Performance Ted Talk will be successful or not.

Cluster 1: Nature/Environment

We identified cluster 1 to include talks that are related to environment and nature. There are 75 talks under this category, and 73 talks have the tag environment, which showed that the clustering method worked well.

To build a predictive model for this category, we tried various combinations of features and evaluated which feature(s) predicted the result most accurately. Again, we used CountVectorizer and LIWC to transform text information into numerical values.

We compared results when using sequential neural network and logistic regression. When using the LIWC metrics from transcript with sequential neural network, the model's accuracy rate to predict whether a video is successful or not was at 53.4%. Unlike other clusters, we found that this method did not work well for environmental talks. Therefore, we tried to run logistic regression on data that used CountVectorizer on title and combined with duration and language, but the accuracy rate was only 48%, which is even lower than using LIWC on transcript. However, as we looked into the transformed data from CountVectorizer, we realized that there are many stop words such as "it", "the", "by", and "a", and after we removed those stop words, the accuracy rate improved to 90%.

Cluster 2: Global Issues

Our clustering method showed that cluster two was the “Global Issues” category. In this category, there were 420 Ted Talks and 209 of them had “Global Issues” as one of the tags in the video. The next common tags were culture (124), business (88), and technology (80). We felt that calling this category the Global Issues category was an appropriate representation of the titles.

Originally, we tried to use the original data from the Ted talks to identify and predict its success. Again, we are defining success as performing above average for this category. When using a similar method to apply count vectorizer to the title, duration, and language, the results were not ideal and only around 0.45. This result wasn’t great, especially since there were only two outcomes. Next, we tried to incorporate the LIWC metrics to assist the regression model. After experimenting with different metrics and different combinations, the model that included 'social', 'family', 'friend', 'percept', 'see', 'hear', 'feel', 'power', 'reward', 'focuspast', 'focuspresent', 'focusfuture', 'leisure', and 'home'. These variables in a logistical regression model yielded a 0.7118 accuracy rate in predicting the success of a Ted Talk.

Cluster 3: Culture/Entertainment

In Cluster 3, it was found to be related to Culture and Entertainment is because of the high prevalence of these two tags within the category. Of the 278 records in this category, all of the records had the tag ‘culture’ and 53 of them had ‘entertainment’, making it the second most prevalent tag in the category.

Originally, the cluster was predicted using different combinations of the LIWC metrics. Some of the combinations we tried were personal identity identifiers (I, you, we, heshe), words that describe a group or a person (social, family, friends, power, etc.), and all LIWC measures. They achieved an accuracy of 0.664, 0.672, and 0.680 respectively.

Next, to improve our accuracy rate, we tried to use a neural network model to predict the success of the Ted Talks in this category. We started with the input 95 nodes and went directly to 1 output, which yielded an accuracy rate of 0.57. We eventually improved the model to 95 inputs, 45 nodes/1 hidden layer, and 1 output node. This structure resulted in an average of 0.771 over 30 runs. This result was significantly better than all of the different combinations that we have tried using logistic regression.

Cluster 4: STEM/Tech

Through our clustering, we’ve determined that Ted Talks in this cluster are predominantly STEM or Tech related talks with 461 records total. Running a multiple linear regression model and printing out the corresponding p-values of the LIWC metrics shows that the following were statistically significant at the 0.05 level in relation to views: WPS (words per second), pronoun,

you, shehe, auxverb, interrog, verb, focuspresent, focusfuture, friend, sad, power, reward, motion, leisure, and AllPunc.

Taking these features, splitting them into training-testing sets, and then fitting them into a logistic regression model, yielded an 30-run-average accuracy rate of ~0.63 - a substantial result! However, we took a step further by utilizing every LIWC metric available to us. The LIWC metrics had mild correlation with the number of views; and consequently, if a video is successful or not. Utilizing a similar approach to cluster 0, we utilized a sequential neural network. The input layer accepted an input size of 95, the number of LIWC features available, with 4 hidden layers with dimensions of the following sizes respectively: 80, 55, 30, and 9. All utilizing a relu activation function. The output layer used a sigmoid activation function with a single output value that was rounded to signify 1 or 0 - success or not.

We've repeated both the logistic regression and sequential neural network analysis 30 times each, averaging. We should note, the neural network result does have variability - running the 30-runs and averaging multiple times have yielded average values ranging from 0.65 to 0.75. However, we've found that it settled around an average accuracy rate of 0.7321.

Cluster 5: Others

Cluster 5 is the only cluster that doesn't follow any theme or specific topics. It contains various topics like personal growth, entertainment, creativity, business, story-telling, etc. Even though it contains topics from various categories which do not fit into any of our clusters, we were still able to calculate accuracy of 79% in this cluster.

Cluster 5 follows the same method as cluster 1. It gets the best results when CountVectorizer is applied to the column title to transform it into numeric form. Then, we use logistic regression on a few of the feature columns and the transformed title.

Exploratory variables for this model are:

- CountVectorizer transformed title
- Duration
- Languages

The response variable for this model is:

- Success

We then ran the logistic regression model 30 times, achieving 79% accuracy in success.

Cluster 6: Social & Global Improvement

This cluster was one of the more populated one with 316 rows. We've noticed some of the language used is very similar to clusters 4, and 9, and took a similar approach. Some statistically significant LIWC metrics were number, cause, leisure, focusfuture, death, and male. Falling in line with expectations on improvement, many talks present quantitative data, have a cause & action theme, and has a focus on the future - hence "focusfuture". 'death' and 'male' are hard to explain without manually combing through the transcripts.

We took two approaches, a logistic regression model as well as a sequential neural network. The independent variables, X, were the LIWC metrics mentioned and the dependent variable, y, was 'success'. Following a standard training-testing split with a test set ratio of 0.3, we processed the data respectively. For the logistic regression model, we added a constant to the X and proceeded onward. The 30-run-average accuracy rate was 0.653 - above a 0.5 rate which is satisfactory for this project.

For the neural network, using the same features, with 4 hidden layers using a relu activation function and an output layer using a sigmoid activation function, yielded an average accuracy rate of 0.633. Not quite as efficient. Curious, we utilized all the LIWC metrics available, which is 95 total, and repeated the experiment. The hidden layers had dimensions of 80, 55, 30, and 9 respectively and yielded a 30-run-average accuracy rate of 0.7000 - a better result than our logistic model.

Cluster 7: Alternative and Green energy

Cluster 7 was one of the minor clusters with just 33 rows. We found out that it had TedTalk videos about alternative resources and green energy. As the cluster size was so small, we could anticipate the features looking at the columns. As expected, the columns which we picked were providing the best accuracy.

Exploratory variables for this model are:

- Duration
- Languages
- film date

The response variable for this model is:

- Success

We then ran the logistic regression model 30 times, achieving an average of 71% accuracy in success. There were few outliers here, where we noticed that few videos with famous speakers like Elon musk had a higher number of views. As the data set was so small, we could not do more analysis on this.

Cluster 8: Biology/Healthcare

Cluster 8 has 162 records, mainly encompassing biology and healthcare issues. Considering that biology-related contents themselves are difficult to digest, we assume that videos with higher views must have contents worth watching again or contents distinguished from competitors. Thus, we decided to utilize a combination of LIWC metrics from transcripts.

The problem is there are too many variables in LIWC metrics. To determine features, we chose four variables that had a relatively high correlation with the number of views. Selected variables are “Authentic”, “Dic (Dictionary words)”, “affect (Affective processes)”, and “posemo (Positive emotion)”. With those four features, we run KNN, Logistic Regression, and SVC models thirty times each and calculated the average accuracy rate.

It turned out that the range of accuracy rate for different classification models was quite similar, mostly ranging from 0.5758 to 0.9091. And the best model with the highest average accuracy rate was a Support Vector Model with the polynomial kernel function, marking 0.8293.

Cluster 9: Sociology/Humanity

For cluster 9, we basically took the same steps for cluster 8. The only difference was the features we chose from the LIWC metrics for transcripts. They were “pronoun (Total pronouns)”, “ipron (Impersonal pronouns)”, “anx (Anxiety)”, “AllPunc (All Punctuation)”, “Quote (Quotation mark)”. Based on five features and a binary vector classifying successful videos, we again run the same set of classification models thirty times each.

The best accuracy rate that we received was 0.9688 in the Support Vector Model with the Radial Basis Function kernel, while the best average accuracy rate came from the Support Vector Model with the linear kernel, with 0.7854.

Correlations between Ratings and views

After finding out predictive factors for the successful TedTalk videos of each category, we became curious about one more thing. We wanted to know what kinds of content people prefer watching in each sub-category of Ted Talks. To investigate, we run a correlation function between word ratings and the number of views. The correlation table below table represents different preferences for each category. To be specific, people love consuming “Persuasive” and “Fascinating” contents for healthcare issues, while “Informative” and “Fascinating” contents when it comes to nature themes. Also, we have found out that, regardless of video category, people evaluate successful contents as “Informative”, “Inspiring”, and “Fascinating”. Those three adjectives are key to success in TedTalks.

	Long-winded	**Infor-mative**	**Inspiring**	Beau-tiful	Inge-nious	Confu-sing	OK	Jaw-dropping	Uncon-vincing	Obno-xious	Persua-sive	Funny	**Fasci-nating**	Coura-geous
Music	0.22	**0.53**	**0.63**	0.60	0.68	0.43	0.32	0.50	0.17	0.10	0.63	0.66	**0.74**	0.49
Nature	0.33	**0.60**	**0.42**	0.35	0.18	0.31	0.37	0.43	0.19	0.41	0.28	0.55	**0.46**	0.22
Global Issues	0.52	**0.81**	**0.82**	0.72	0.56	0.58	0.67	0.76	0.33	0.40	0.60	0.53	**0.83**	0.73
Culture	0.52	**0.83**	**0.81**	0.47	0.76	0.58	0.79	0.68	0.36	0.35	0.81	0.75	**0.85**	0.56
STEM	0.12	**0.51**	**0.47**	0.47	0.65	0.16	0.46	0.43	0.12	0.10	0.34	0.55	**0.68**	0.40
Social Improv.	0.34	**0.80**	**0.70**	0.55	0.47	0.28	0.58	0.53	0.10	0.13	0.57	0.44	**0.78**	0.35
Green Energy	-0.03	**0.26**	**0.78**	0.44	0.56	0.08	0.26	0.41	-0.12	0.19	0.02	0.16	**0.84**	0.42
Healthcare	0.40	**0.74**	**0.65**	0.56	0.54	0.42	0.73	0.48	0.40	0.21	0.78	0.70	**0.74**	0.34
Humanity	0.74	**0.88**	**0.92**	0.83	0.86	0.63	0.74	0.93	0.17	0.22	0.79	0.91	**0.94**	0.83
Others	0.56	**0.81**	**0.85**	0.58	0.66	0.48	0.74	0.59	0.44	0.36	0.80	0.34	**0.89**	0.63

Conclusion

We wanted to find out predictive factors for the number of views in TedTalk videos and conducted initial research on a given dataset. Soon we realized that lower views don't necessarily mean a bad TedTalk with inferior contents. Obviously, some videos with niche topics are difficult to generate the same amount of attention compared to those with fancy or already popular topics.

To address the topic preference issue, we clustered the TedTalk dataset into ten groups by video tag keywords and applied different machine learning models to each category to achieve the highest average accuracy rate.

Throughout these experiments, what we have found is that no single model fits every topic. Each category required different features and models to get the best accuracy rate. We also learned that qualitative features such as transcripts and titles are more effective than quantitative ones to predict successful TedTalk videos. Lastly, TedTalk videos with high view numbers are evaluated and recognized as "Informative", "Inspiring", and "Fascinating", regardless of topics.

References

Bajarin, Tim. “Why TED Matters.” Time Magazine. 24 Mar. 2014. <https://time.com/34784/why-ted-matters/>

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin. DOI: 10.15781/T29G6Z