Popularity measures on TED talks

Motivation

Popularity is defined as the state or condition of being liked, admired, or supported by many people.

Using data from TEDTalk videos, this project attempts to determine the underlying factors that influence the popularity of a TEDTalk without relying on views. [1] [2]

Popularity Measure

The ratings of each TEDTalk were converted into a binary feature using Sentiment Analysis, which determines how positive or negative a rating might be.

By aggregating the frequency of positive and negative ratings, it shows that there was bias in the amount of positive ratings. It was then decided that the proportion of negative ratings will be the measure of popularity that this report will work with.

Numerical Data

All of the numerical features were removed because they were highly correlated with the number of views a TEDTalk had. Each TEDTalk was separated into two categories which denote whether it is an official TEDTalk or not.

Textual Data

On ted.com, Each TEDTalk has a list of related TEDTalks, much like Youtube's Recommended list. Through this, it is possible to paint a clearer picture of how to categorize TEDTalks by consolidating the tags of the related videos.

The transcripts of each TEDTalk is cleaned for punctuatations and stopwords, and filtered into a set of most frequent terms using the ***Term Frequency Inverse Document Frequency*** (TF-IDF) algorithm. With the unstructured text data initially cleaned, it is possible to find an underlying structure using topic modelling. [3]

Topic Modelling

The method used to find topics from TEDTalk transcripts and tags is ***Latent Dirichlet Allocation*** (LDA), which is a generative probabilistic model that discovers structures from unstructured textual data. [4]

LDA clusters a document *i* into a topic *j* in *k*, with a k-dimensional Dirichlet distribution that corresponds to a bag of words or tags. It computes the prior distribution to each word belonging to a topic *j*. After computing a prior probability, it iss possible to derive the posterior distribution given new data.

Perplexity

A perplexity plot that graphs the different levels of a topic (fig. 1) is used to evaluate the k-dimensional Dirichlet distribution that LDA uses. This is a similar evlauation to an elbow plot used to evaluate a k-means clustering algorithm. It is imperative to choose the number of topics that would give interpretability. This resulted in choosing four and six topic groups for TEDTalk tags and transcripts respectively. This resulted in topic groups (fig. 2, 3) that are interpretable when visually inspected.

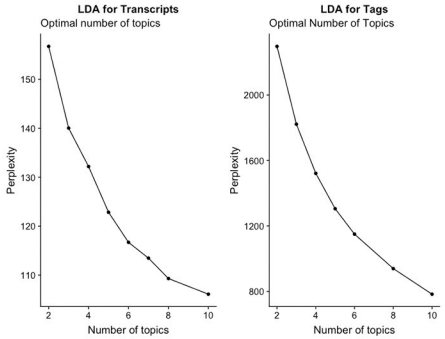

Figure 1: Perplexity plots used to determine number of topics

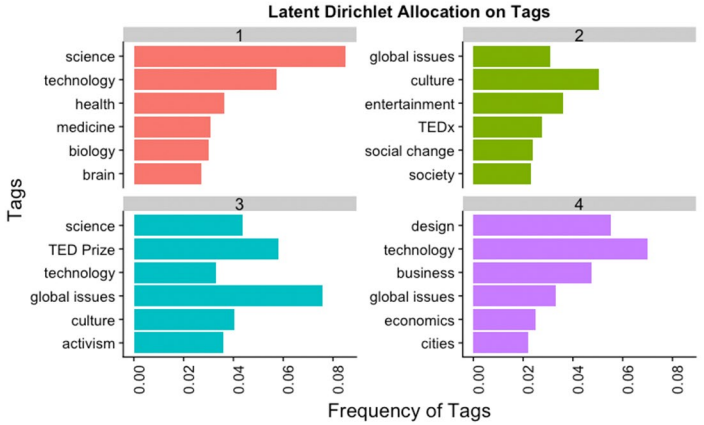

Figure 2: Top six tags for each of the four topic groups based on tags

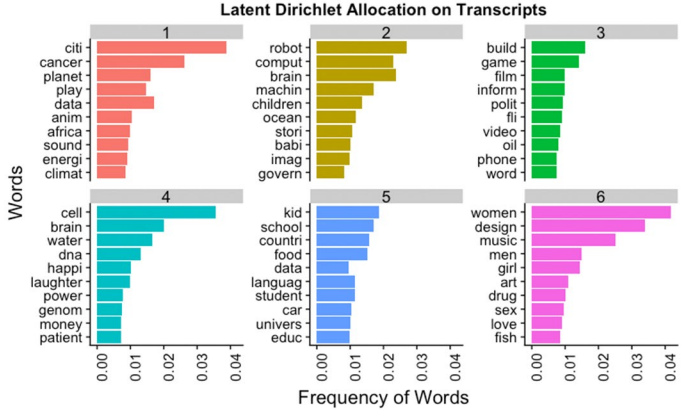

Figure 3: Top ten words for each of the six topic groups based on transcripts

1 "TED: Ideas Worth Spreading." TED, www.ted.com/.
2 Banik, Rounak. "TED Talks." Kaggle, 25 Sept. 2017, www.kaggle.com/rounakbanik/ted-talks.
3 Rajaraman, A. and Ullman, J.D. "Data Mining". Mining of Massive Datasets, 2011, pp. 1–17.
4 M., David, et al. "Latent Dirichlet Allocation." Journal of Machine Learning Research, 2003, jmlr.csail.mit.edu/papers/v3/blei03a.html.

Analysis

The overall distribution of TEDTalks for tag topic groups (fig. 4) show that they have a larger proportion of popular videos in topic groups 1, 2, and 3 than topic group 4. Similarly, for transcript topics, the overall distribution (fig. 4) shows that they have a larger propoprtion of popular videos in topic groups 3, 5, and 6 than topic groups 1, 2, and 4. Upon this result, a logistic regression model was fitted that models popularity with the officiality of the TEDTalk, tag topics, and transcript topics. The resulting coefficients of the fitted model are negative, which implies that these features suggest an unpopular TEDTalk video.

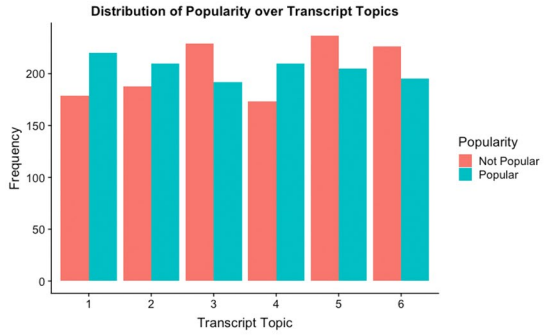

Figure 4: The popularity of topic groups represented in a bar chart

Conclusion

Based on the findings, the underlying factors that directly determine a popular TEDTalk is inconclusive. However, factors were found that could negatively affect the popularity of a TEDTalk, which can be used as a guideline on what to avoid when making popular TEDTalks for ted.com viewers.

Such factors include the TEDTalk being official, and the most prominent tags and words featured in tag topic group 4 and transcript topic groups 3, 5, and 7. Therefore, avoiding these features may increase the popularity of a TEDTalk.