Popularity measures on TED Talks

Clinton Ali, 998218281, Sajeed Bakht, 1001527975, Kim Estelo, 1001640325, Quinton Goos, 1002542796, Mohamed Osman, 1000851562

30 November 2019

i. Responsibilities

Sajeed and Quinton explored the viability of neural network models and exploratory data analysis. Clinton, Kim and Mohamed explored clustering and topic modelling. Everyone worked on preprocessing the data. Clinton and Kim explored and fitted the logistic regression model.

0 Introduction

Popularity is defined as the state or condition of being liked, admired, or supported by many people. This report aims to define popularity in relation to TED Talks, and use it to find the underlying factors that contribute to a TED Talk being popular.

Background

The report uses two datasets. Both were scraped from ted.com by Rounak Banik[1] and contains data on all the TED Talk videos that were uploaded on the site from 27 June 2006 to 21 September 2017.

The first dataset contains records on each TED Talk, and includes features, such as the amount of views and comments, ratings, video titles, descriptions, and the name of the speakers who appeared in the video. It also includes a list of TED Talks that are related to the video. There are 17 features and 2550 records. The second dataset contains transcripts for each TED Talk, and a list of URLs that link to the transcribed videos. There are 2467 records, which is less than the amount in the first dataset since there are TED Talks that do not have transcripts. Additionally, there are a couple of duplicate entries in the transcripts which reduce the number of entries to 2464. Moreover, the population of this data are viewers from the TED website.

```
## # A tibble: 6 x 17
##   comments description duration event film_date languages main_speaker
##      <dbl> <chr>          <dbl> <chr>     <dbl>     <dbl> <chr>
## 1     4553 Sir Ken Ro~     1164 TED2~    1.14e9        60 Ken Robinson
## 2      265 "With the ~      977 TED2~    1.14e9        43 Al Gore
## 3      124 New York T~     1286 TED2~    1.14e9        26 David Pogue
## 4      200 In an emot~     1116 TED2~    1.14e9        35 Majora Cart~
## 5      593 "You've ne~     1190 TED2~    1.14e9        48 Hans Rosling
## 6      672 "Tony Robb~     1305 TED2~    1.14e9        36 Tony Robbins
## # ... with 10 more variables: name <chr>, num_speaker <dbl>,
## #   published_date <dbl>, ratings <chr>, related_talks <chr>,
## #   speaker_occupation <chr>, tags <chr>, title <chr>, url <chr>,
## #   views <dbl>
```

[1] Rounak Banik, "TED Talks," Kaggle, September 25, 2017, accessed November 25, 2018, https://www.kaggle.com/rounakbanik/ted-talks.

1 Preprocess

1.1 Transforming transcripts

A common method in representing transcripts is by using a term frequency - inverse document frequency (TF-IDF). By removing punctuations, common english stop words and transforming the text to lowercase, a ranking was produced based on how frequent the words occur (term frequency) and how important the words are (document frequency). Then, the top 10 words based on the TF-IDF ranking was selected. Method is based on “Tidy Text Mining with R”[2].

```
## Joining, by = "document"
```

```
## # A tibble: 6 x 7
##   document term      count total       tf   idf  tf_idf
##      <dbl> <chr>     <dbl> <dbl>    <dbl> <dbl>   <dbl>
## 1        1 1930s         1  1428 0.000700  4.00 0.00280
## 2        1 19th          1  1428 0.000700  3.23 0.00227
## 3        1 2065          1  1428 0.000700  7.81 0.00547
## 4        1 30s           1  1428 0.000700  4.10 0.00287
## 5        1 ability       2  1428 0.00140   1.57 0.00220
## 6        1 abstract      1  1428 0.000700  3.15 0.00220
```

1.2 Analysis

The amount of views in a video can be seen as an indicator of its popularity. However, it is simply a metric of how many people saw it, and does not inherently represent a positive or negative response.

For example, there exist Youtube videos that have a high amount of views but are also unpopular, such as “Friday”, by Rebecca Black.[3] As of 25 November 2018, it has over 127 million views and 4.5 million like-and-dislike ratings, but it also has a like-to-dislike ratio of 27.8%, which implies the video is severely unpopular. This also shows why ratings and the amount of ratings are more reliable indications of a video’s popularity.

Ratings are represented in the data as JSON objects, and a ratings column for a TED Talk is a list of ratings that were given by TED users. Ratings have a “name” key that describes a certain sentiment, such as “Funny” and “Inspiring”, or “Longwinded” and “Obnoxious”. The first goal is to extract these sentiments and convert them into binary values that represent either “positive” or “negative” sentiments. To find the sentiment of these ratings, the Sentiment Analysis toolbox was used. [4]

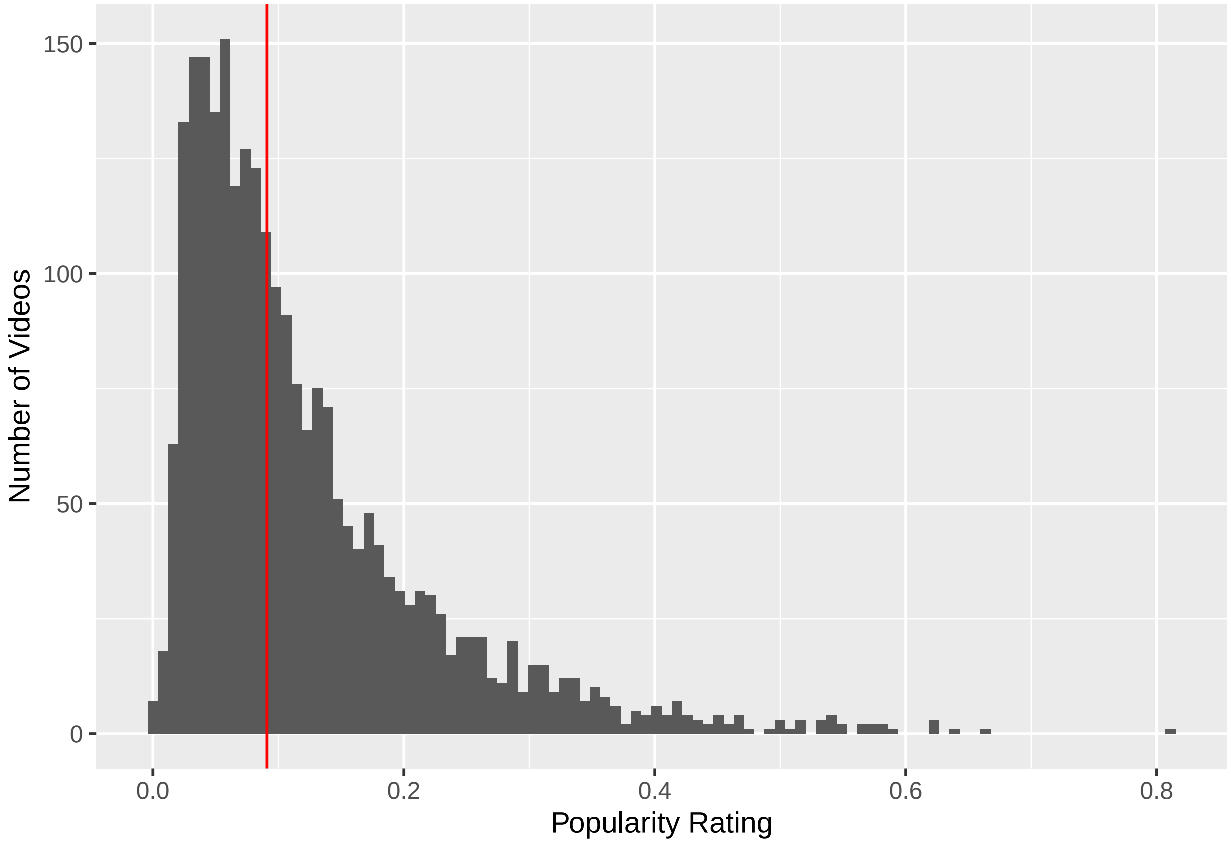
The general objective of sentiment analysis is to find the “sentiment” of a word, or a corpus, of the text. There are many sentiments a word can have, but, for the purposes of this report, the only notable extraction is whether it is positive or negative. There were manual adjustments to some of the classifications that were misclassified by the algorithm. From this, a proportion of negative-to-positive ratings was made for each video. By visual inspection it can be seen that the library misclassified “Funny” as negative, and “Longwinded” as positive. Due to such a small ammount of unique ratings it is paramount that all classifications be correct to obtain an accurate result. These two were adjusted to be properly classified as positive for “Funny” and negative for “Longwinded”.

[2] Julia Silge & David Robinson, “Tidy Text Mining With R”, Website, accessed November 25,2018, https://www.tidytextmining.com/

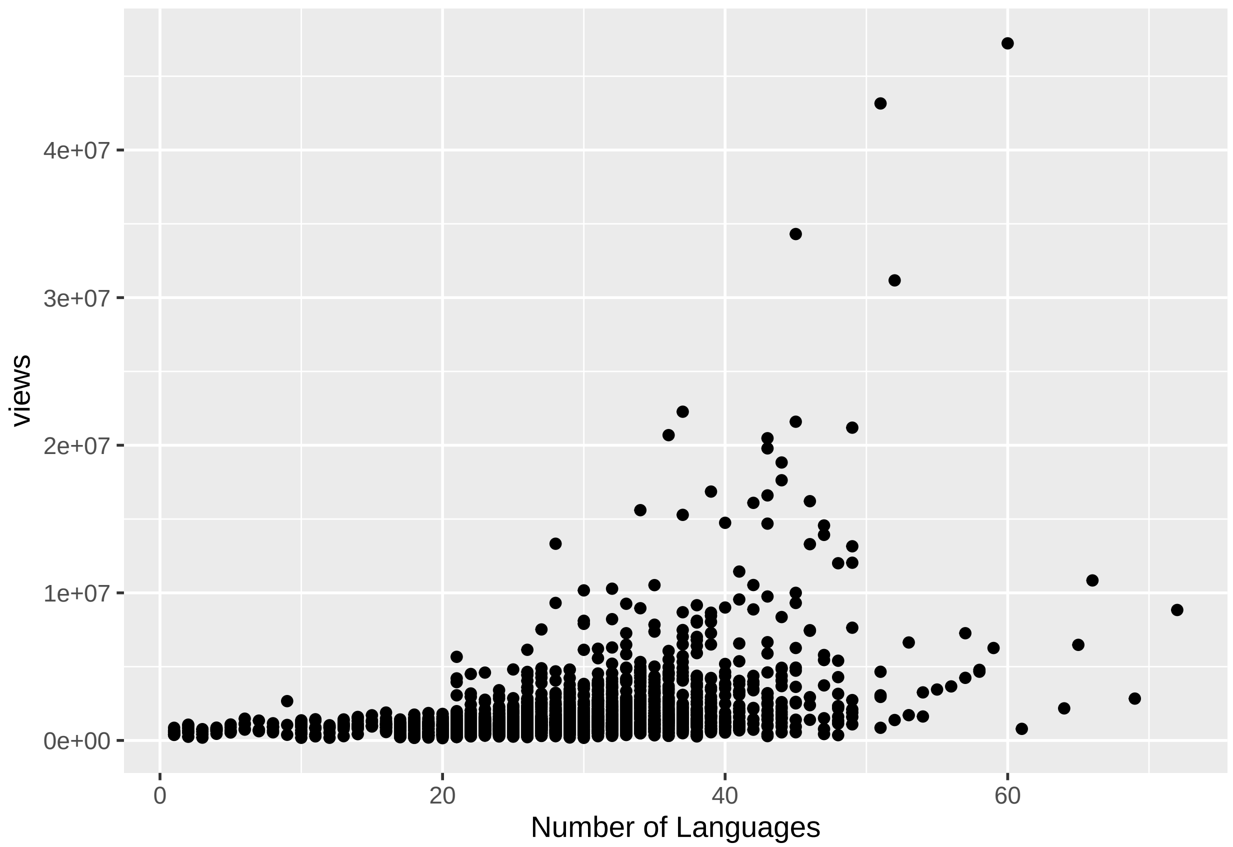
[3] Rebecca Black, “Rebecca Black - Friday,” YouTube, September 16, 2011, accessed November 25, 2018, https://www.youtube.com/watch?v=kfVsfOSbJY0.

[4] Stefan Feuerriegel and Nicolas Proellochs, The Comprehensive R Archive Network, April 09, 2018, accessed November 25, 2018, https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html#applications-in-research.

In this report, a popular video is defined as having a proportion that is below the median proportion of all the videos in the dataset. If the proportion is equal to or greater than the median, then the video is not popular. Each video was classified using this rule, and the distribution of the proportions was plotted in the following graph. The median was chosen as a measure of centrality because the proportion of negative ratings are not distributed normally.



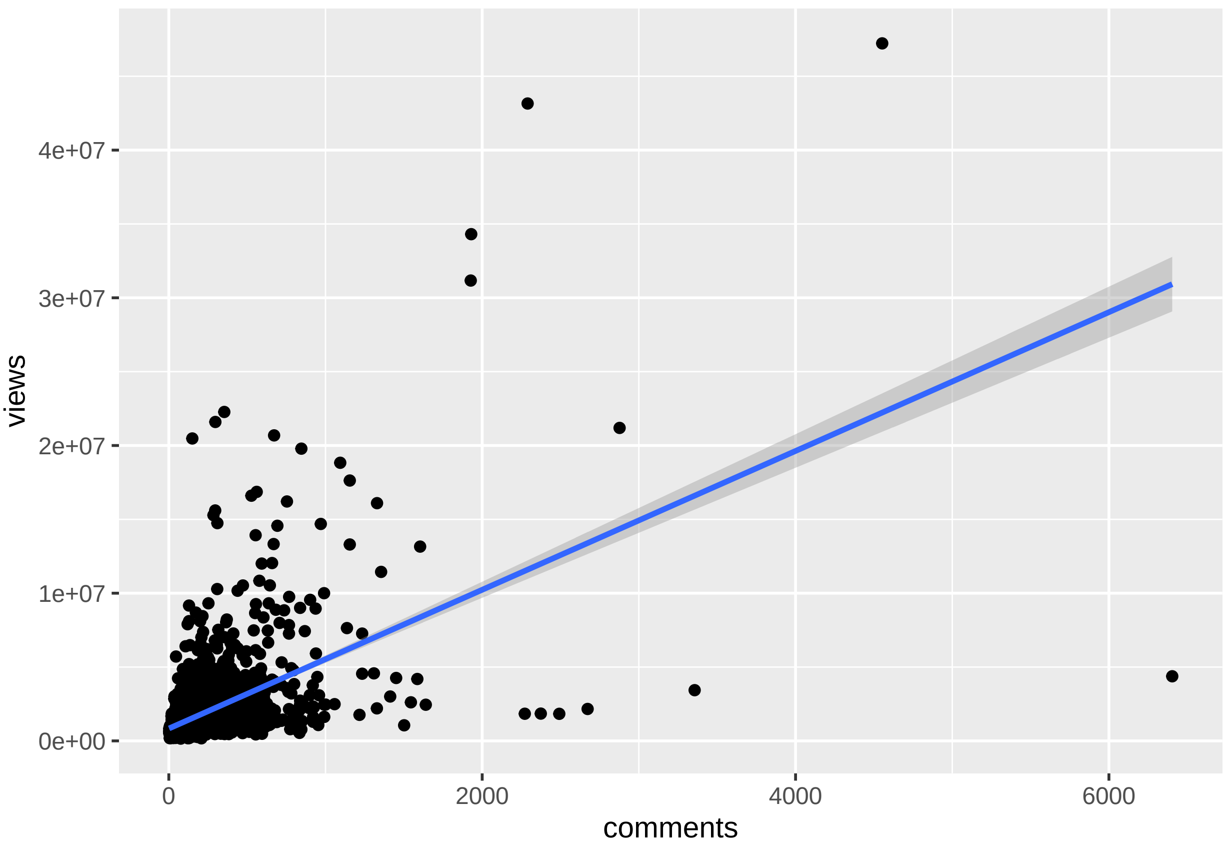
The observations to the left of the red line are popular while the ones to the right are unpopular.



Upon inspection an upward trend of views as the number of langauges increases can be seen. This is due to the fact that as a video attains more and more views, they then get transcribed in more languages. This is an indication that using languages is not the best measure to figure out if a video is popular because multicollinearity exists. Also, the popularity measure introduced later is to predict popularity on 2017 data. And most videos in 2017 start off with being transcribed in only 1 langauge.

It can be seen that in 2017 videos has a very low amount of languages compared to the other variables. So it can be concluded that languages are correlated with views, and will not be a *fair predictor variable.*

Comments are also correlated with views. It's an unfair predictor, and later further reasoning as to why comments are not used as a predictor. But for a sneak peek, trying to understand what makes a video popular, it's likely that a model will overfit if comments are included as a part of it due to comment uniqueness. And there's nothing special about saying "well if a video has alot of comments then it will probably be popular."



Lastly, transcript.csv was obtained from [5], these are which transcripts for each video. This is because it is of interest to find the finer themes and words that make videos popular.

Transcripts are included in this report to see if there are any particular words that were said in a TED Talk that make it popular.

1.3 Film date

When looking at a video, it is important to acknowledge the key considerations when attempting to make a popular video on TED. The hypothetical thought experiment can be seen as the following:

Imagine a publisher who has the ability to make a TED Talk. The metrics the publisher can personally control are tags, the transcript, the speaker, and the event type. However, the publisher cannot directly influence user-provided metrics such as views and comments. Therefore, the process excludes variables that is out of the publisher's control, as well as data that is irrelevant to this report, such as URLs. The amount of languages is also discarded since it is linked to the amount of views it has, making it a user-proivded metric. Film date is also discarded because there is no information to infer from it; publish date is the only relevant date to look at.

1.4 Official Ted Events

TED talks are hosted by Events. These events are sometimes hosted by ted themselves, or by independent TED approved groups.So far, there are 85 unique events such as "TED2006" and "Elizabeth G Anderson School".

[5] Rounak Banik, "TED Talks," Kaggle, September 25, 2017, accessed November 25, 2018, https://www.kaggle.com/rounakbanik/ted-talks.

A new variable,is_official, was preprocessed to indicate whether the talk took place during an official ted event. An official ted event is either TED Global or TED. TEDx and all other are considered independent.

It is important to note that all the videos below the red line is considered popular. It seems that unofficial events tends to popular more often.

1.5 Related Talks Tags

Each observation has a related_talks column, which is a JSON list, indicating the ids and titles of talks similar to the observation. The titles were extracted, and then used to search for their tags, which were appended, to create a related_talk_tag column.

2 Topic Modelling

2.1 Dimensionality reduction

There are 12860 features on the preprocessed data. A viable way to visualize the data is by performing Principal Component Analysis (PCA). On figure A, the first two principal components show a clear separation of clusters. However, these two components amount to less than 5% of the entire data. Therefore, performing PCA to reduce the dimensionality isn't a viable choice.

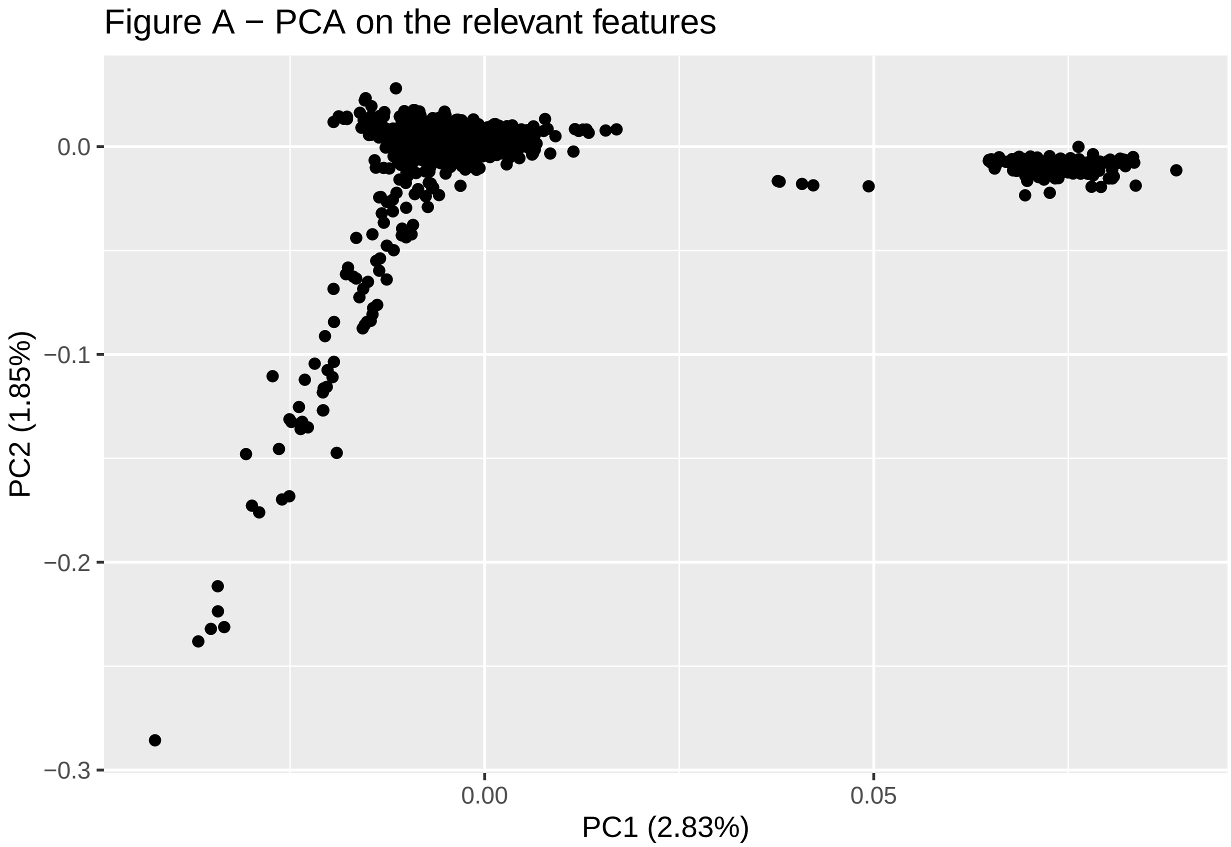
2.2 Creating bag of tags and words

The list of tags and related tags on a given video is converted into a bag of tags model, where each column is tied to a tag and each entry in the matrix is tied to the count of words on a given row. Also, the list of words from the TF-IDF rating done previously is then converted into a bag of words model similar to the tags.

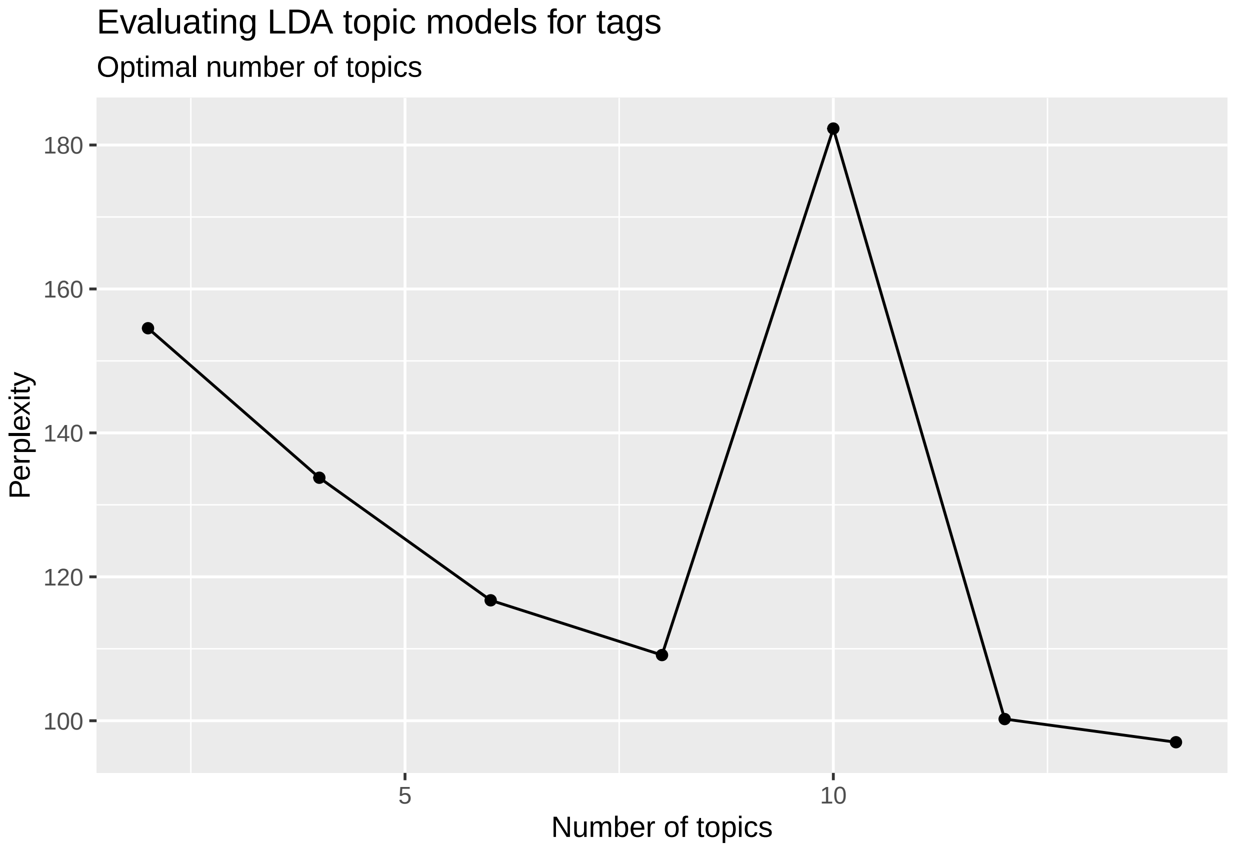
In order to combat the huge dimensions of the data, it is possible to reduce the bag of words and bag of tags into clusters. K-means clustering isn't a good method of unsupervised learning because it uses euclidean distance to cluster into k groups. In the bag of words and tags, each observation is tied to a count of words or tags in each document. Measuring the euclidean distance of words and tags does not make sense because two words may have a short distance but entirely different meaning. One of the proposed models to cluster topics is called Latent Dirichlet Allocation (LDA) which comes from topic modelling and is entirely diffrent from cluster analysis. This method is generative probabilistic model that discover structure from unstructured data.

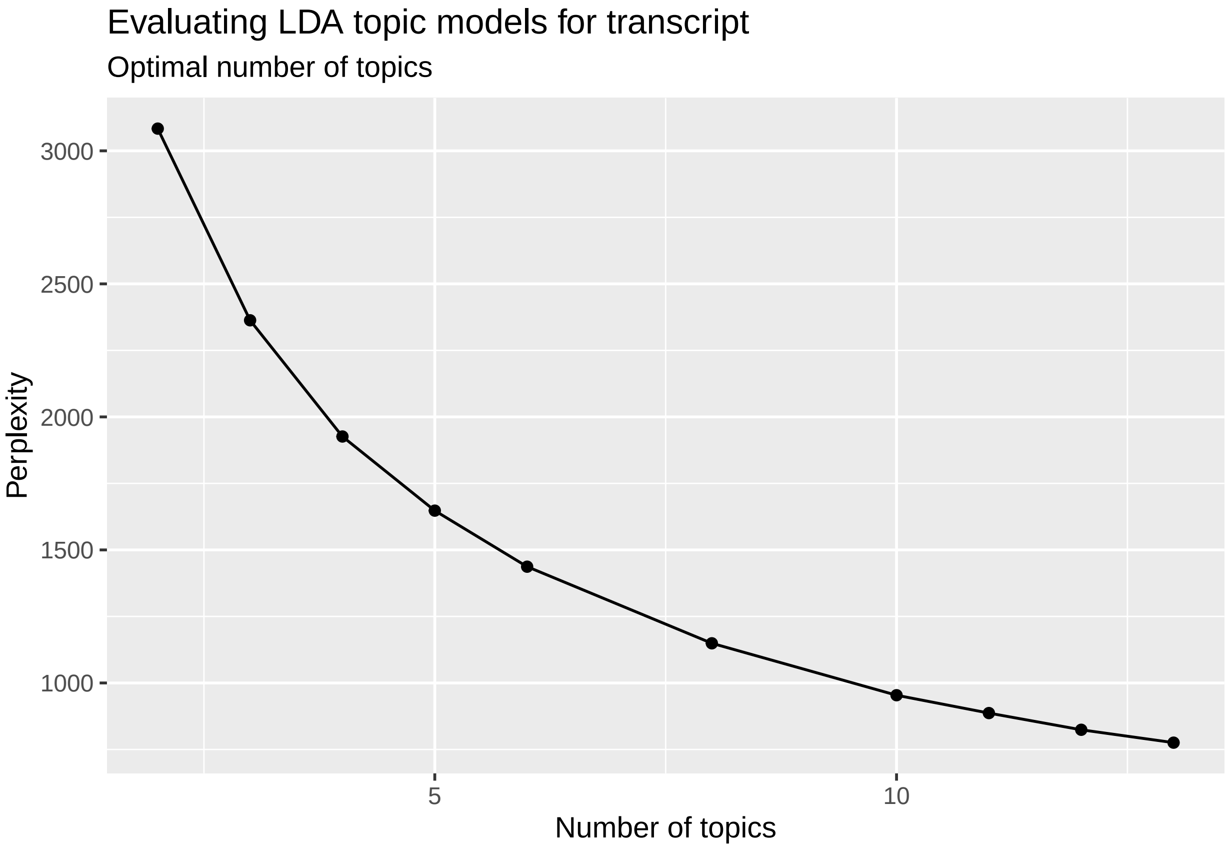
LDA clusters a document i into a topic $j \in k$ with a k-dimensional dirichlet distribution that corresponds to the bag words or tags. It computes prior distribution to each words belonging to a topic j. After computing a prior probability, it is then possible to derive the posterior distribution given new data. The full derivation of the LDA can be read on http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf by Blei, Ng and Jordan.

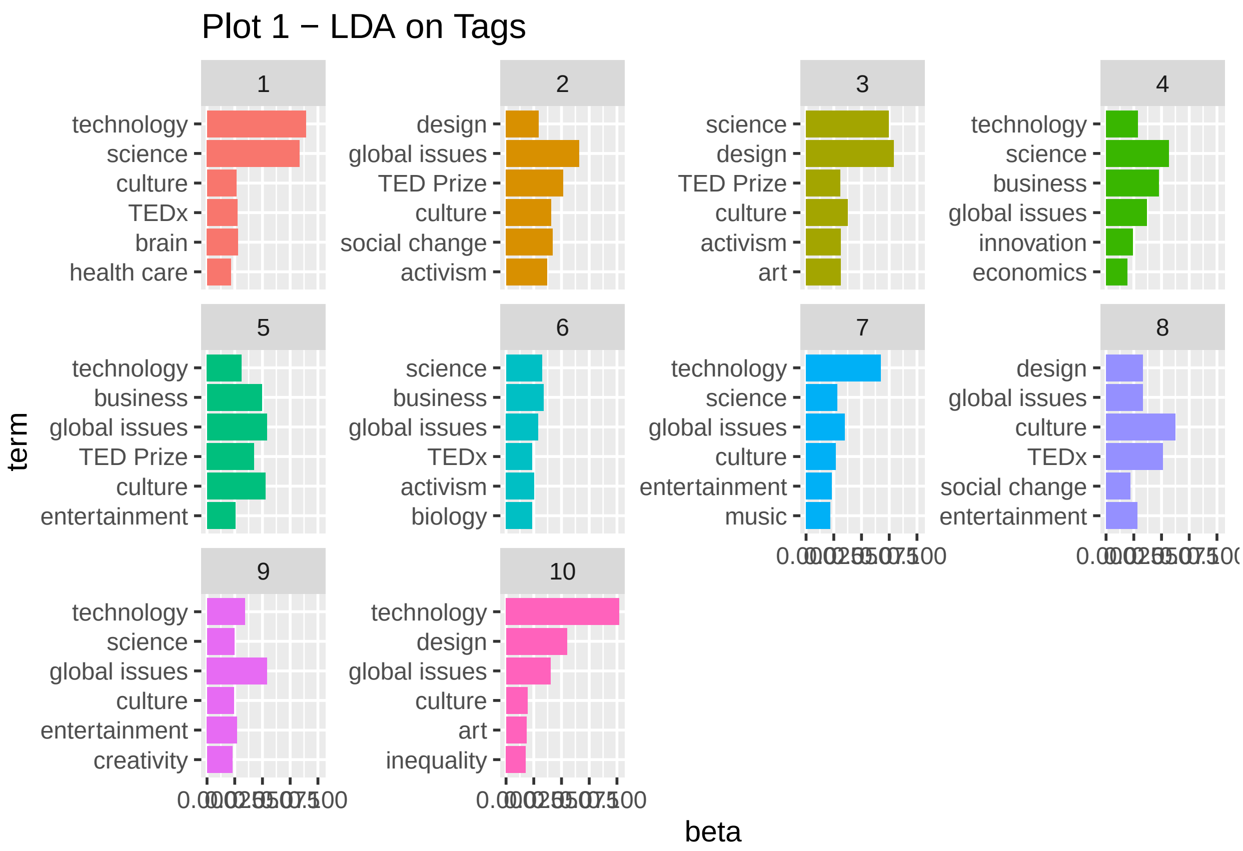
Some aspects of LDA are driven by intuition, and perplexity is a stastical measure on how well a probability model predicts a sample. For a given k topics the perplexity will measure how well k topics captures the entire data. Based on the perplexity plots below, it was observed that there was a local maxima at k=10 for tags and an inversely proportional relationship. Then by inspecting plot 1 and 2, it shows that 10 topics gives clarity to both tags and trancripts. For example, in plot 3 it shows that topic 1 and 10 are tags related to brains and technology respectively. Also, in plot 4 it shows that topic 5 and 10 are transcripts related to brains and women. This confirms that the separation of topics are viable.

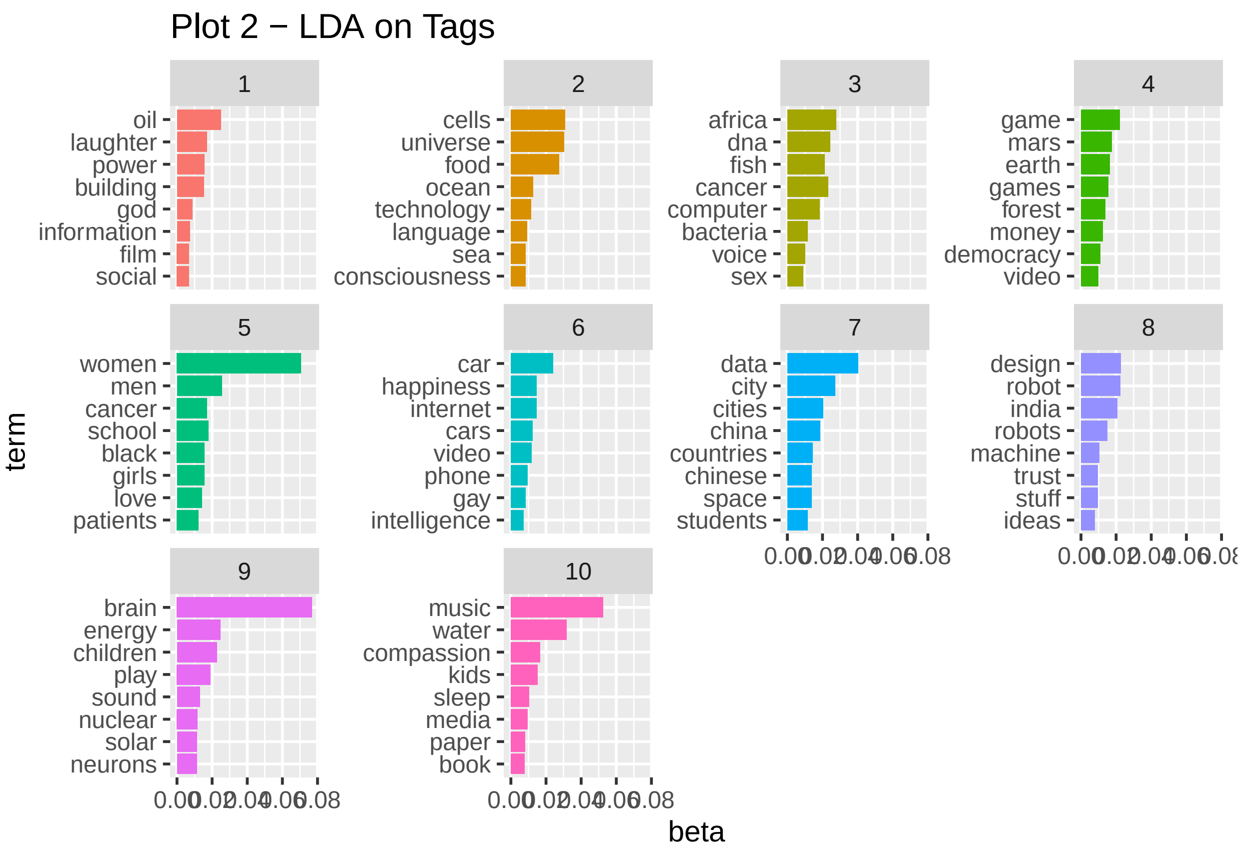


```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
```

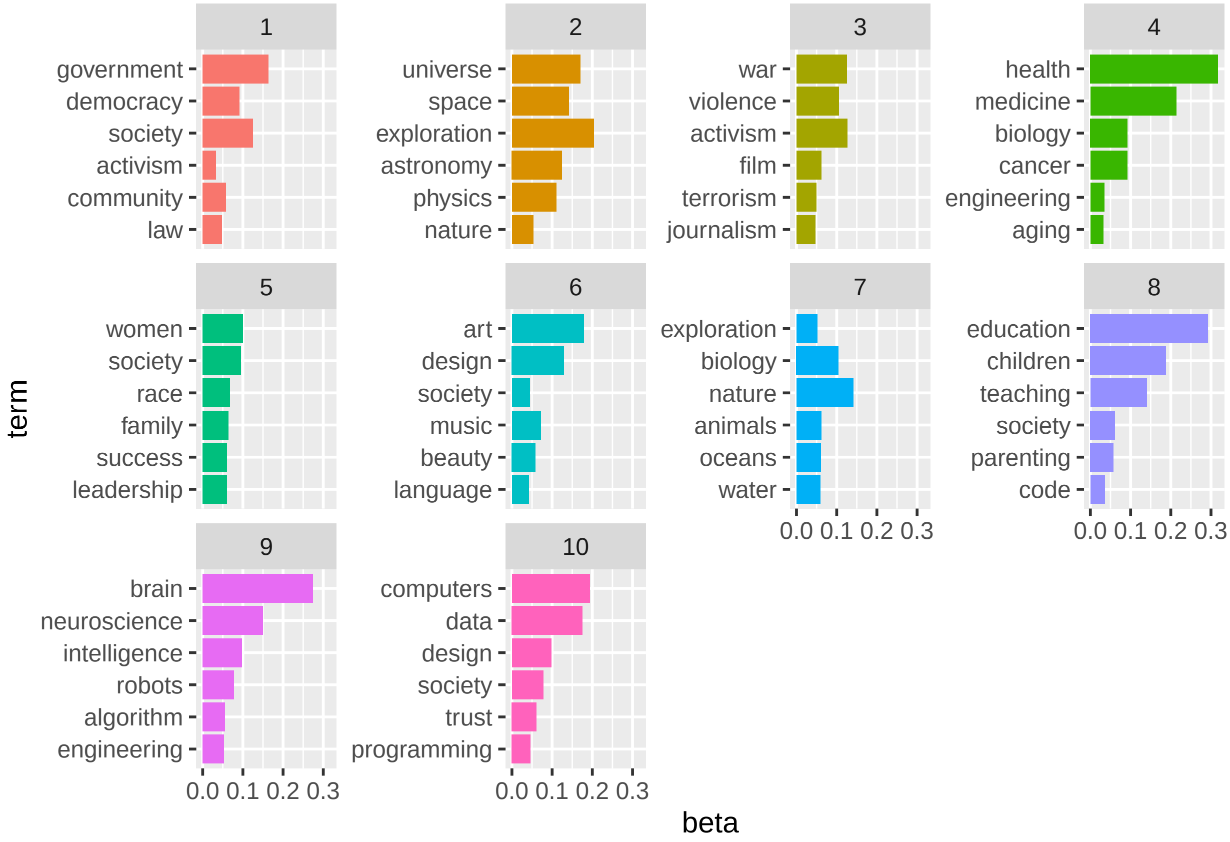

Evaluating LDA topic models for tags
Optimal number of topics
180
160
140
120
100
Perplexity
5
10
Number of topics

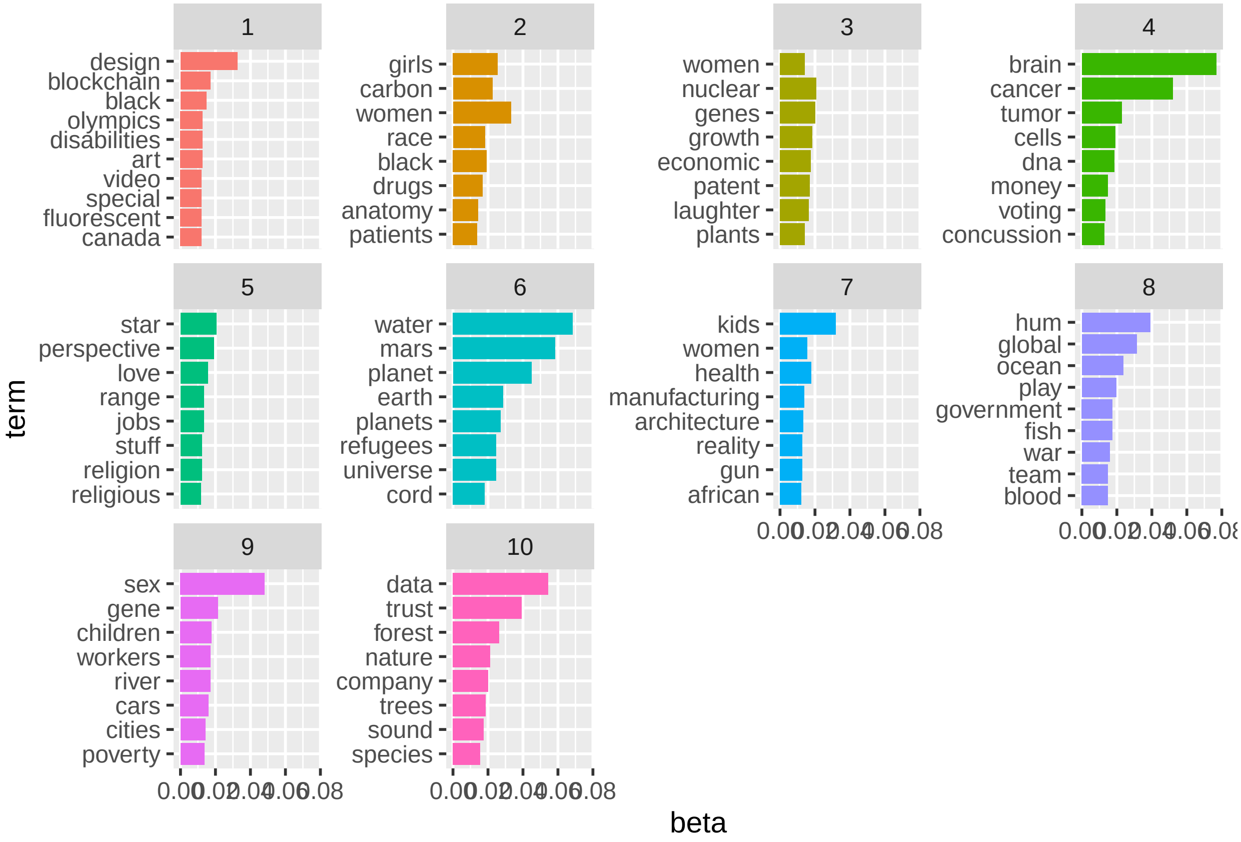

Evaluating LDA topic models for transcript
Optimal number of topics
Perplexity
3000
2500
2000
1500
1000
5
10
Number of topics


Plot 1 − LDA on Tags
1
technology
science
culture
TEDx
brain
health care
2
design
global issues
TED Prize
culture
social change
activism
3
science
design
TED Prize
culture
activism
art
4
technology
science
business
global issues
innovation
economics
5
technology
business
global issues
TED Prize
culture
entertainment
6
science
business
global issues
TEDx
activism
biology
7
technology
science
global issues
culture
entertainment
music
8
design
global issues
culture
TEDx
social change
entertainment
9
technology
science
global issues
culture
entertainment
creativity
10
technology
design
global issues
culture
art
inequality
term
0.000
0.025
0.050
0.075
0.100
beta



Running LDA on the data for 2016 also gives a good separation of topics. Choosing the same number of topics as the training data will give consistency. In plot1, topic 1 and 4 are tags related to governments and health respectively. In plot2, topic 4 and 6 are transcripts related to brain cancer and water in planets respectively.


1
government
democracy
society
activism
community
law
2
universe
space
exploration
astronomy
physics
nature
3
war
violence
activism
film
terrorism
journalism
4
health
medicine
biology
cancer
engineering
aging
5
women
society
race
family
success
leadership
6
art
design
society
music
beauty
language
7
exploration
biology
nature
animals
oceans
water
8
education
children
teaching
society
parenting
code
9
brain
neuroscience
intelligence
robots
algorithm
engineering
10
computers
data
design
society
trust
programming
0.0 0.1 0.2 0.3
term
beta



3 Classifier

3.1 Logistic Regression

After grouping the tags and transcripts into respective topics, it is now possible to build a model. Both the transcripts and the TED Talk topic features range from 0 to 10, and the popular/unpopular and official/unofficial TED Talk features are binary. The model to predict the popularity of the talks is logistic regression. We are not using an accuracy score because it has an imbalanced dataset, making it a bad metric. For example, TED 2017 might have more positive ratings, making it an imbalanced dataset. Data is not normal, so we don't take into account the expected cost of misclassification. There are many approaches to loss functions, but the chosen loss function for this model is 0-1 loss because it is the easiest to optimize. The logistic regression model has a log loss, but it approximates to a 0-1 loss. By Occam's razor, we use the simpler one. The training set for the model is the TED Talk data up to 2016, and the test set is the data from 2017. The labels for both sets are the popular measures that were defined earlier in the report. The goal for this model is to measure whether or not the historical TED data was a good predictor for future data. The model was fitted on the training data and was tested on the testing data.

```
##
## Call:
## glm(formula = ratings_pos ~ trans_topic + tag_topic + is_official,
##     family = "binomial", data = traindata)
##
## Deviance Residuals:
##    Min       1Q   Median       3Q      Max
## -1.315  -1.172  -1.051   1.173   1.309
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.29181    0.14061   2.075  0.03796 *
## trans_topic  0.01150    0.01797   0.640  0.52235
## tag_topic   -0.07797    0.02991  -2.607  0.00914 **
## is_official -0.21822    0.08592  -2.540  0.01109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3145.5  on 2268  degrees of freedom
## Residual deviance: 3131.4  on 2265  degrees of freedom
## AIC: 3139.4
##
## Number of Fisher Scoring iterations: 3
```

```
## [1] 0.5897436
```

Another model was trained only on the TED data from 2016. The goal for this model is to see if the immediate past data is a good predictor for future data.

```
##
## Call:
## glm(formula = ratings_pos ~ trans_topic + tag_topic + is_official,
##     family = "binomial", data = ted2016)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6902  -1.1044  -0.7587   1.1412   1.6063
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.11089    0.38440   0.288  0.77299
## trans_topic -0.12992    0.04682  -2.775  0.00552 **
## tag_topic    0.11733    0.04988   2.352  0.01865 *
## is_official -0.02723    0.27502  -0.099  0.92112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 336.87  on 242  degrees of freedom
## Residual deviance: 323.96  on 239  degrees of freedom
## AIC: 331.96
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] 0.5333333
```

For this run, both models achieved an accuracy just above 54.87%, meaning it was only slightly better than flipping a fair coin. Therefore, there is no definitive conclusion as to whether or not historical and immediate past data are good predictors for future data. Finally, the likelihood ratio test gave 23.5, which is greater than our chi-squared critical value (alpha = 0.05,df = 3), which is 7.814. Therefore, the null hypthesis is rejected and it is concluded that the model has significant features.

4 Closing

4.1 Next Steps

Logistic regression may not be the best predictor for predicting TED Talk data. So, one would look at other models that would produce better accuracies. Stratified k-fold cross-validating would be a good method to determine whether or not the model is overfitting. The LDA model would also be changed from a 1-gram model to a k-gram model. The transcript would also be cleaned to take out words that can be characterized as stopwords, such as "yeah". Other numerical popularity measures that are correlated to views could be considered.

4.2 Conclusion

Based on our findings, the tags, transcripts and type of event associated with the video are good starting points to determine popularity.