Final Project Report - TED Dataset - December 17, 2017

Nrithya Saravanabhavan, Liz Boccolini, Xiqiao Tong, Rohith Putcha, Akash Pandey

Introduction

Our final project required us to analyze data mined from Ted Talks. The data consisted of two files: one file with information about 2,550 TED Talks and one file with the transcripts to 2,471 of those TED Talks.

Objective

With the dataset we hope to answer the following business questions.

1. Can we predict the number of views and comments for a new video?
2. Can we cluster videos based on popularity?
3. Based on views, comments, and ratings what kind of impact did a video have on viewers?
4. Can we cluster videos based on tags for a video recommendation system?

What is TED?

According to its website, TED "is a nonprofit devoted to spreading ideas, usually in the form of short, powerful talks (18 minutes or less)" (ted.com). To date, TED has sponsored over 2,600 talks covering 425 topics in 115 languages ("TED Talks").

Data Curation

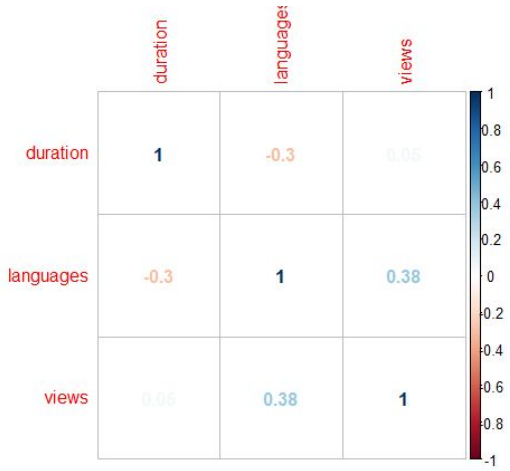
Few fields in the dataset required processing before we could use it for analyses.

- We used R, UNIX, and Excel to manipulate the data into a manageable format
- The date fields were in unix timestamps and had to be converted to human readable dates
- The ratings field had data on all 14 ratings and their counts in a single field and had to split into multiple fields to be able to use them in the analyses
- Additionally, we also created a tag dictionary with frequencies based on the video tags

Question 1: Can we predict the number of views and comments for a new video?

Using Logistic Regression

We wanted to know if we could develop a model to predict the number of comments and the number of views a video would receive. Below is a plot depicting the correlations between all three variables:

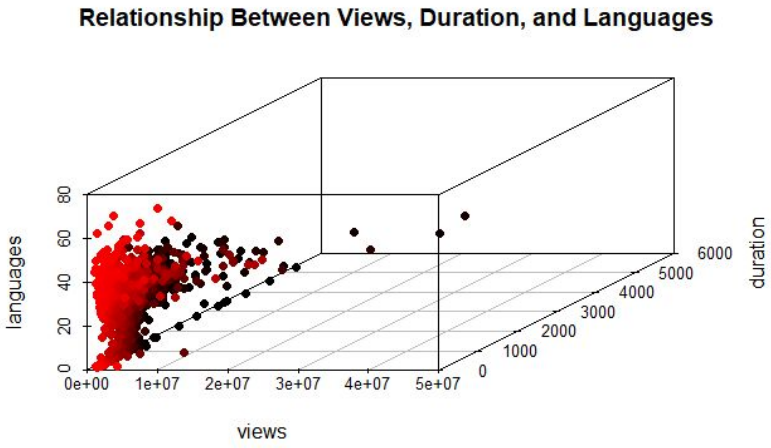


As shown, the highest correlation is between views and languages at 0.38. The correlation between views and duration is 0.05. We can see from this chart that the correlation coefficients are not very high between these variables. The weak correlation between views and duration might be due to the fact that there is not much variability in the length of the videos. (There seem to be only less than 10 videos that are over 30 minutes long.)

To predict the number of views, we used duration of the talk and the number of languages in which the talk was available. Our model was

y = -2339034.6 + 11741.1(duration) + 112232.3(languages)

Below is a scatter plot depicting the relationship between all three variables.



As shown, most videos are clustered between 0 and 10,000,000 views. Because TED Talks are generally not longer than 18 minutes, it is no surprise that most of the videos are clustered between 0 and 1,000 seconds (about 16.67 minutes).

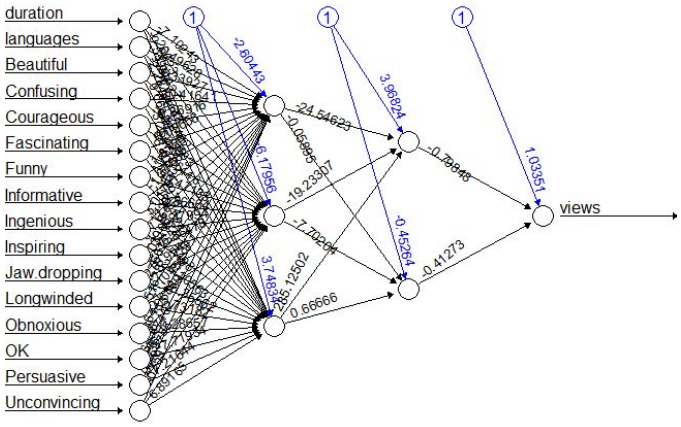
To predict the number of comments we use views, comments, languages, and duration. The strongest correlation in this model is between views and comments (0.53). The correlation between comments and languages is weaker at 0.32, and the correlation between comments and duration is the weakest at 0.14.

Our model for predicting the number of comments was

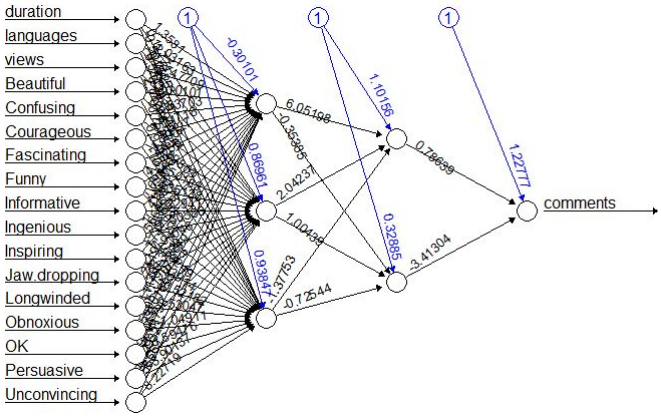
y = -.01693 + 0.1351(duration) - 0.0005035(views) + 5.991(languages)

Using Artificial Neural Networks

In addition to creating linear models, we explored predicting both outcomes using artificial neural networks. However, we added the number of times each talk received each of TED's 14 ratings (beautiful, courageous, persuasive, etc.) to our previous data set. Our training set consisted of 75% of the entire data set (1,912 data points). We found that the MSE was extremely high for predicting both views and comments; introducing more numeric variables (from outside the given data set) might improve the model.



MSE: 6283596155909



MSE: 4068261144209

Question 2: Can we cluster videos based on popularity?

Popularity of a video can be determined with the number of its views, comments, and languages available. We tried to do some unsupervised clustering using K-means to see if we can determine the popularity of a video. The dataset was split 75% for training and 25% for testing.

Kmeans –Trial 1

Features : View + Comments + Language

Clusters : 2 **BSS/TSS ratio** : 57.4%

```
K-means clustering with 2 clusters of sizes 2501, 49

Cluster means:
  comments languages    views
1 177.9208  27.04078  1433505
2 887.8367  41.89796 15213507

Clustering vector:
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
   2    1    1    1    2    2    1    1    1    1    1    1    1    1    1    1    1    1    1    1

 981  982  983  984  985  986  987  988  989  990  991  992  993  994  995  996  997  998  999 1000
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
 [ reached getOption("max.print") -- omitted 1550 entries ]

Within cluster sum of squares by cluster:
[1] 3.360316e+15 3.425818e+15
 (between_SS / total_SS =  57.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

Kmeans –Trial 2

Features : View + Comments + Language

Clusters : 3 **BSS/TSS ratio** :76.5%

```
K-means clustering with 3 clusters of sizes 199, 25, 2326

Cluster means:
  comments languages    views
1  465.201  36.71859  5381650
2 1114.080  43.60000 20507549
3  158.236  26.34781  1181007

Clustering vector:
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
   2    3    3    3    1    2    1    3    3    3    3    3    3    1    3    3    3    3    3    3

 981  982  983  984  985  986  987  988  989  990  991  992  993  994  995  996  997  998  999 1000
   3    3    3    3    1    1    3    3    3    3    1    3    3    3    3    3    3    3    3    3
 [ reached getOption("max.print") -- omitted 1550 entries ]

Within cluster sum of squares by cluster:
[1] 8.247889e+14 1.968490e+15 9.516347e+14
 (between_SS / total_SS =  76.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

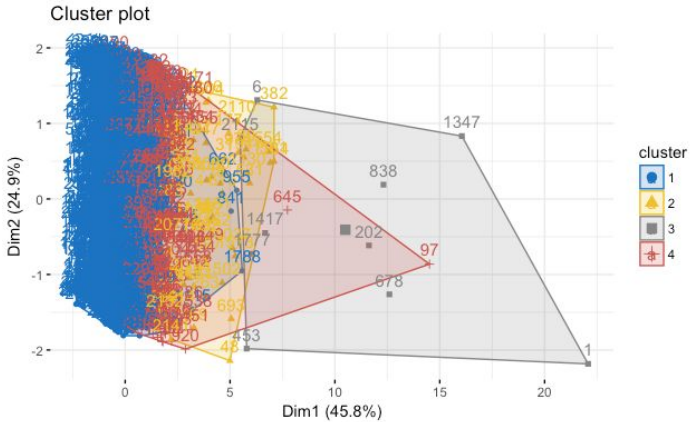
Kmeans –Trial 2

Features : View + Comments + Language

Clusters : 4 **BSS/TSS ratio** : 84.7%

```
K-means clustering with 4 clusters of sizes 318, 10, 2167, 55

Cluster means:
   comments languages    views
1  374.5943  34.50629  3616759
2 1589.6000  46.10000 28186533
3  146.3189  25.86294  1066983
4  661.7091  40.05455 10663848

Clustering vector:
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
   2    1    3    3    4    2    1    3    1    1    3    3    3    1    3    1    3    3    3    3
  21   22   23   24   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40

 981  982  983  984  985  986  987  988  989  990  991  992  993  994  995  996  997  998  999 1000
   3    3    3    1    1    1    3    3    3    3    1    3    1    3    3    3    3    3    3    3
 [ reached getOption("max.print") -- omitted 1550 entries ]

Within cluster sum of squares by cluster:
[1] 3.992767e+14 9.466419e+14 5.276874e+14 5.669711e+14
 (between_SS / total_SS =  84.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

As seen in the model above Trial 3 gives the best clustering model, which has a BSS/TSS ratio of 84.7 % (0.847). BSS/TSS ratio measures the goodness of the clustering that k-means has developed, here 84.7 % illustrates a good fit. The visualization with 4 clusters are presented below.



The sizes of the 4 clusters are respectively 318, 10, 2167, 55. The majority of the spots are highly concentrated on the left side of the plot, which indicates that most of the videos share similar popularity according to the number of its views, comments, and languages available. On the other hand, a few spots are spread apart in the center and on the left of the plot, which demonstrates there are small number videos that have different popularity compared to the majority.

Question 3: Based on views comments and ratings, what kind of impact did a video have on viewers?

TED talks have a complex rating system. A viewer can choose one, more than one, or all of the 14 options associated with rating (Beautiful, Informative, Fascinating, Jaw-dropping, OK, Inspiring, Courageous, Persuasive, Ingenious, Unconvincing, Obnoxious, Funny, Confusing, Long-winded). Services like YouTube and Netflix moved from “5-star” to a “thumbs-up” system which resulted in more votes and in turn better recommendations. In order to strike a middle ground, we created a 4-level metric to understand the ratings.

RATINGS	METRIC	IMPACT
Beautiful, Fascinating, Jaw-dropping, Inspiring, Ingenious	4	High Impact
Informative, Courageous, Persuasive, Funny	3	Good Impact
OK, Long-winded	2	Neutral
Confusing, Obnoxious, Unconvincing	1	Bad Impact

Based on these metric we tried to classify the videos using KNN and Random Forest into the 4 levels of viewer impact. This model can be used to truly understand what people like and don't like about a video for better content.

KNN –Trial 1

Model : Metric ~ (Ratings + View + Comments + Duration + Language)

K Nearest Neighbors : 3 **Training and Test Split** : Random sampling of 75:25

Accuracy : 50%

	testLabels				
model	1	2	3	4	Row Total
1	0	0	4	2	6
2	0	0	0	1	1
3	3	5	112	160	280
4	4	1	144	218	367
Column Total	7	6	260	381	654

KNN – Trial 2

Model : Metric ~ (Ratings + View + Comments + Duration + Language)

K Nearest Neighbors : 5 **Training and Test Split** : Random sampling of 75:25

Accuracy : 55%

	testLabels				
model	1	2	3	4	Row Total
1	0	0	1	0	1
2	0	0	1	0	1
3	2	0	101	129	232
4	3	5	157	245	410
Column Total	5	5	260	374	644

Since 55% accuracy is not much to go on we built a model using Random Forest to compensate for any bias in the model.

Random Forest – Trial 1

Model: Metric ~ (Ratings + View + Comments + Duration + Language)

Training and Test split: 60:40 **Number of trees:** 500

No. of variables tried at each split: 4 **Accuracy:** 86%

	test$metric				
model2	1	2	3	4	Row Total
1	1	0	0	0	1
2	0	0	1	0	1
3	7	1	327	29	364
4	5	2	101	546	654
Column Total	13	3	429	575	1020

Random Forest – Trial 2

Model: Metric ~ (Ratings + View + Comments + Duration + Language)

Training and Test split: 60:40 **Number of trees:** 500

No. of variables tried at each split: 4 **Accuracy:** 87%

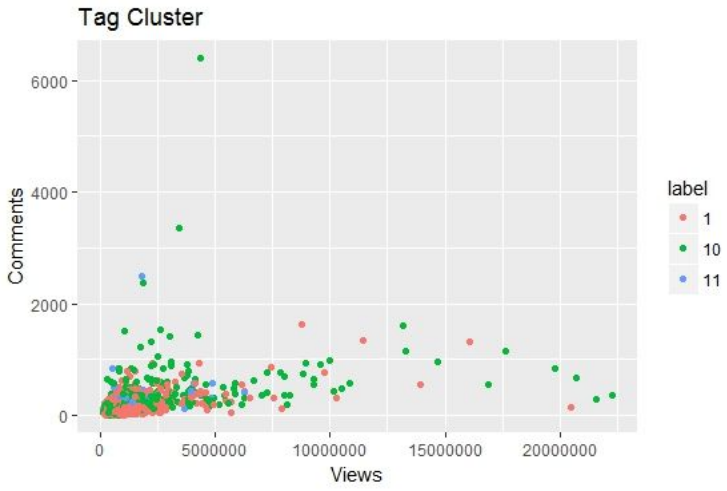
	test$metric				
model2	1	2	3	4	Row Total
1	1	0	0	0	1
2	0	0	1	0	1
3	4	3	363	45	415
4	4	3	71	525	603
Column Total	9	6	435	570	1020

We can see that Random Forest is able to generate better accuracy for this model.

Question 4: Can we cluster videos based on tags for a video recommendation system?

In the dataset, there are 147 unique tags to describe the content or the theme of a video. Each video has one or more of these tags associated with it. The viewer gets a list of recommended videos based on the video he/she just watched. We aim to build a cluster of videos based on these tags in an effort to recommend videos centered with the viewer's choice of theme/tags. Since there are 147 tags and each of these tags can be put together in any number of combinations, the number of logical clusters could be huge. Considering the scope of this project, we choose 2 of the top most tags-"technology" and "culture" to extract the videos and form labels representing the combination. This list can be developed for n number of combinations. Based on the labels we can plot a scatter plot to understand the clusters.

Tags of the scatterplot below: Technology (01), Culture (10), Technology (11)

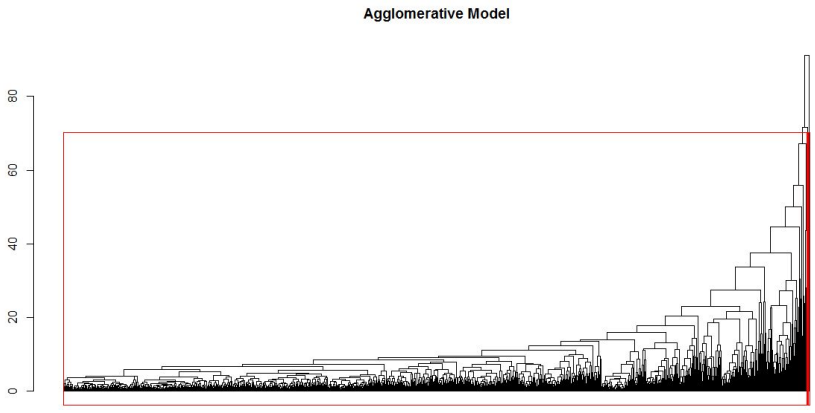


Since the clusters are not well defined we can not use K-Means clustering method to determine the clusters. We hence use agglomerative and divisive clustering methodologies. From the dendrograms we can determine which data point was associated which cluster, but it is hard to associate each cluster with its specific tag label. Hence we compute an approximate accuracy.

Agglomerative Model

Model : Tag Labels ~ Views+Comments+Duration+Languages+Ratings

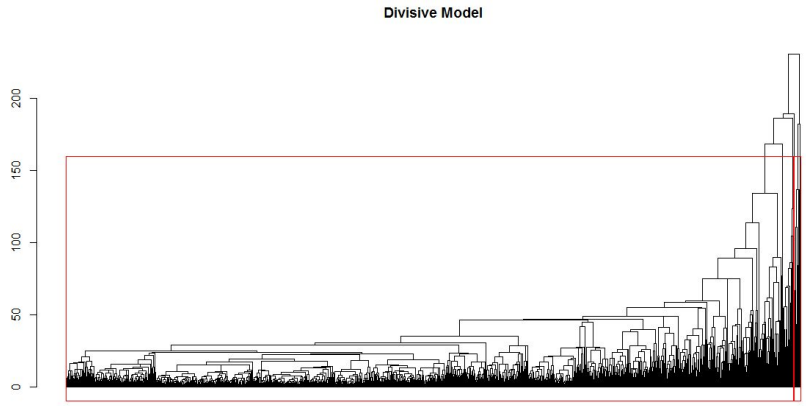
No of clusters : 3 **Height :** 70 **Accuracy :** 56%



Divisive Model

Model : Tag Labels ~ Views+Comments+Duration+Languages+Ratings

No of clusters : 3 **Height :** 160 **Accuracy :** 57%



Both clustering methods do not produce high accuracy, but if more data was used to define the clusters better we maybe able to produce better accuracy for clustering the videos based on tags.

Conclusion

After viewing the TED Talks data with our group, we decided that the first step was to curate and manage the data into a better format. We did this by using R, Unix, and Excel to both separate the columns with multiple data points and to store the data into another file to use in R Studio. Next, we needed to find out what the data could be used for and we used came up with important business questions to work towards solving. These involved predicting the number of views and comments for a new video, popularity analysis, predicting the impact of the video on viewers, and analyzing the tags to form a recommendation system. We used many algorithms to mold our curated data into useable information to solve these questions including Multivariable Logistic Regression, Artificial Neural Networks, K-means, KNN, Random Forest, Agglomerative Clustering, and Divisive Clustering. We were able to cluster popular videos with an accuracy of 85%, classify videos based on impact with an accuracy of 87%, and cluster videos based on tag with an accuracy of 57%. Overall, the information obtained from the analysis can be used to aid TED in understanding which videos will perform better and generate more viewership.

Reference

Kabacoff, Robert I. "Scatterplots." *Quick-R*, 2017, https://www.statmethods.net/graphs/scatterplot.html.

"TED." TED: Ideas Worth Spreading, https://www.ted.com/.

"TED Talks." TED: Ideas Worth Spreading, https://www.ted.com/talks.

"Visualize Correlation Matrix Using Correlogram." STHDA: Statistical Tools for High-Throughput Data Analysis, http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram.

Cluster Analysis in R: Practical Guide. (2017). http://www.sthda.com/english/articles/25-cluster-analysis-in-r-practical-guide/111-types-of-clustering-methods-overview-and-quick-start-r-code/

ML for DS Fall 2017 Resources kmeans.r. https://content.sakai.rutgers.edu/access/content/group/2b7fe5da-092c-4f14-bc24-012422586a14/Code/kmeans.r Retrieved on Dec 2017.

What does total ss and between ss mean in k-means clustering? (2014). https://stats.stackexchange.com/questions/82776/what-does-total-ss-and-between-ss-mean-in-k-means-clustering/82779

Convert Unix epoch to date object. Retrieved from https://stackoverflow.com/questions/13456241/convert-unix-epoch-to-date-object

How to do a non greedy match in grep. Retrieved from https://stackoverflow.com/questions/3027518/how-to-do-a-non-greedy-match-in-grep

Escaping apostrophes in regex. Retrieved from https://stackoverflow.com/questions/7400000/escaping-apostrophes-in-regex

Pattern matching and replacement. Retrieved from

https://stat.ethz.ch/R-manual/R-devel/library/base/html/grep.html