Final Project: YouTube Trends in Various Countries

Author
Eric Dogariu Affiliation
Princeton University email
dogariu@princeton.edu

Abstract

Using trending YouTube (2) data from the Trending YouTube Video Statistics dataset (1), which used the Trending-YouTube-Scraper tool (3), the patterns of most popular YouTube videos in various countries is ascertained. This paper introduces a "popularity index", which is defined as the number of views a video has plus 4 times the number of likes minus 4 times the number of dislikes plus 2 times the number of comments, so that 1 like or dislike is equivalent to 4 views and 1 comment is equivalent to 2 views. Different unsupervised analysis models are used to determine the most frequently occurring features among the trending videos, or words or numbers from the title, description, channel title, or tags associated with a video. Then, supervised analysis is used to predict popularity using the top 40 most important features.

1 Introduction

In this final project, trending YouTube data in the United States, Canada, United Kingdom, and India during the time late 2017 to mid 2018 (1) is used to determine any interesting trends. The title, channel title, tags, and description (1) associated with each YouTube video is used in Latent Dirichlet Allocation (LDA) analysis and K-means clustering analysis, which are unsupervised (4,6,8), and then in supervised analysis to perform regression to attempt to predict the popularity score using the data. Unsupervised analysis means that there is no specific outcome data to predict with the training data, so the analysis is carried out without training labels and testing labels, and Supervised analysis means that the training data is used to predict a certain outcome (8). In this paper, all of the methods described use dimension reduction in order to detect patterns in the data (4, 5, 8). In the analysis models, a composite of all of the videos is used to create 10 "latent" or composite clusters that summarize the data (7,9). Each latent clusters contains certain features, such as certain words that appear often in the titles or descriptions. Different types of analysis are then performed, such as determining the most predictive features of each latent user and evaluating how well each model works (7,9). Furthermore, supervised analysis is then used to predict popularity from each model. Furthermore, the top 40 features used in the supervised analysis (7,9). This paper uses the code and writeup from Assignment 3, since the data and type of analysis is similar (7,9).

2 Methods

The data for each country is in a csv file, so the Canada data is in the CAvideos.csv file, the India data is in the INvideos.csv file, the UK data is in the GBvideos.csv file, and the USA data is in the USvideos.csv file, which were all downloaded from Kaggle (1). The category numbers are described in the .json files accompanying the csv files (1). The categories of data are video-id, trending-date, title, channel-title, category-id, publish-time, tags, views, likes, views, comment-count, thumbnail-link, comments-disabled, ratings-disabled, video-error-or-removed, and description (1). Firstly, I downloaded the zip file from Kaggle (1). For each country, I then loaded corresponding csv

Number	Country	Title	Channel Title	Popularity
1	Canada	YouTube Rewind: The Shape of 2017	YouTube Spotlight	145126668
2	Canada	Childish Gambino - This Is America	ChildishGambinoVEVO	111079833
1	India	YouTube Rewind: The Shape of 2017	YouTube Spotlight	132518125
2	India	Marvel Studios' Avengers: Infinity War Official Trailer	Marvel Studios' Avengers: Infinity War Official Trailer	100841285
1	UK	Nicky Jam x J. Balvin - X (EQUIS)	NickyJamTV	435415656
2	UK	Te Bote Remix - Casper, Nio Garcia,...	Flow La Movie	347510347
1	USA	Childish Gambino - This Is America	ChildishGambinoVEVO	244966023
2	USA	Ariana Grande - No Tears Left To Cry	ArianaGrandeVevo	161032050

Table 1: Top 2 Distinct Most Popular Videos

file, using Pandas (10), and removed the trending-date, video-id, publish-time, and thumbnail-link columns. Then, I removed the rows that had comments-disabled, ratings-disabled, or video-error-or-removed values that were true, and then I deleted the comments-disabled, ratings-disabled, or video-error-or-removed columns. Then, I used the formula popularity = views + 4*(likes-dislikes) + 2*comments to replace the views, likes, dislikes, and comment-count columns with a popularity column. Then, I listed the top 2 most popular distinct videos for each country in Table 1. Then, I used the TfidfVectorizer (4,5) function to select the top feature or word for each video, with a custom stop-word list that I created, which includes small, commonly used words, such as "the" and "a", and commonly occuring words on YouTube that weren't very descriptive, that I didn't want included as categories. Then, the answers in the dataframe were replaced with the one word found from the TfidfVectorizer, and then one hot encoding was performed on the new dataframe, using pd.get-dummies from Pandas (10), as in the precept code and Homework 3 (7,9). Then, the data was manually split, with 80 percent of the data being training data and 20 percent of the data being testing data. After this, LDA (Latent Dirichlet Allocation) (4,5) was used, with 10 components, as done in the precept code (7,9). This was done by ranking the features or question responses by important in percentiles, then taking and showing the top 10 features for each latent cluster in the code (7,9). The model was scored using overall log-likelihood, average log-likelihood for each sample, using the score() function from the Gaussian Mixture model from scikit-learn (4,5), AIC analysis (4,5), and BIC analysis (4,5). The AIC analysis and BIC analysis were performed using the GaussianMixture model (4,5). The scoring metrics of log-likelihood (4,5), average log-likelihood per sample (4,5), AIC analysis (4,5), and BIC analysis (4,5) were also evaluated at 10 components for each model. The same analysis, as for LDA, was repeated for K-means clustering (4,5). Scatter plots for the K-means clustering were produced using ta modified version of the PCA scatter plot code from precept (7,9), with Pylab (10) being used, shown in Figures 1a, 1b, 2a, 2b, 3a, 3b, 4a, and 4b. In addtion, to silhouette score (4,5) was calculated for K-means clustering. The results of the unsupervised analysis are shown in Table 2. For the supervised analysis, the GradientBoostingRegressor (4,5) and RandomForestRegressor (4,5) were used for the popularity category. The SelectKBest method from scikit-learn (4,5) was used for feature selection to select the best 40 features for predicting popularity. These features are shown in Figures 5, 6, 7, and 8. The regressors were scored using the mean-squared error (MSE) (4,5), explained variance score (4,5), and R2 (4,5) scoring metrics. .

3 Results

Table 1, Table 2, and Table 3 use abbreviations for the columns and rows. They are LDA-Latent Dirichlet Allocation, KM-K-means, llGM-Log-likelihood from Gaussian Mixture, LLPS-Log-likelyhood per sample from Gaussian Mixture, Silscore-Silhouette score, AIC-Akaike Information Criterion, AICPS- Akaike Information Criterion Per Sample, BIC-Bayesian Information Criterion, BICPS-Bayesian Information Criterion Per Sample, NF-Number of Features Selected, MSE-Mean Squared Error, R2- R2 score, EVS-Explained Variance Score, GBR-GradientBoostingRegressor, and RFR-RandomForestRegressor.

Copy to clipboard

4 Discussion and Conclusion

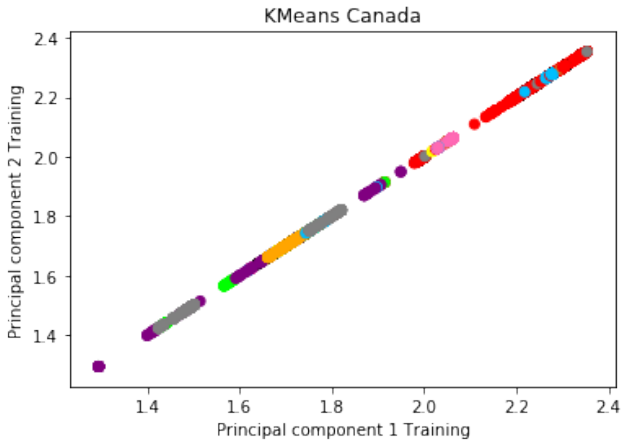
From the data, several patterns emerge. Firstly, from Table 2, it can be seen that the log-likelihood, AIC (6), and BIC (6) values for the K-means clustering are larger in magnitude than those for the

Model	Country	Num	SilScore	LLGM	LLPSGM	AIC	AICPS	BIC	BICPS
KM	Canada	10	0.92	42.10	0.0053	-673219.30	-84.037	-668613.83	-83.46
KM	India	10	0.84	37.33	0.0052	-532082.82	-74.48	-527552.83	-73.85
KM	UK	10	0.93	44.93	0.0059	-683707.55	-89.68	-679134.72	-89.08
KM	USA	10	0.94	42.02	0.0052	-674808.30	-83.87	-670199.96	-83.30
LDA	Canada	10	N/A	29.49	0.0037	-471208.92	-58.82	-466603.45	-58.25
LDA	India	10	N/A	29.63	0.0041	-422051.89	-59.078	-417521.90	-58.44
LDA	UK	10	N/A	28.61	0.0038	-434855.89	-57.038	-430283.05	-56.44
LDA	USA	10	N/A	27.31	0.0034	-438188.07	-54.46	-433579.73	-53.89

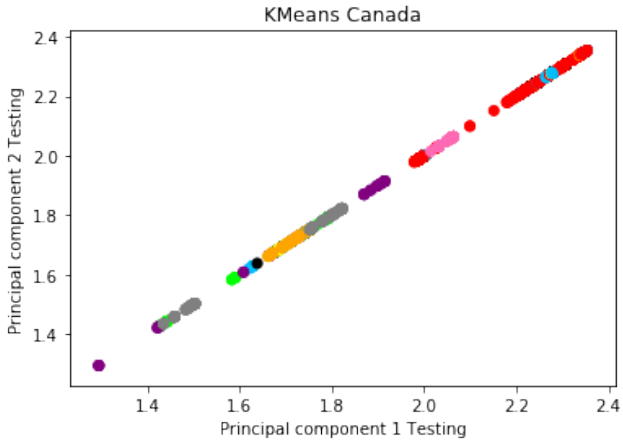
Table 2: Unsupervised Analysis Results

Regressor	M	NF	Country	MSE	R2	EVS
GBR	KM	40	Canada	18476979351729.96	-0.0021	4.14e-05
RFR	KM	40	Canada	18477673029296.61	-0.0021	3.87e-05
GBR	LDA	40	Canada	18476918932960.39	-0.0021	4.35e-05
RFR	LDA	40	Canada	18473944551390.35	-0.0019	4.38e-05
GBR	KM	40	India	9874449425716.26	-0.0080	6.41e-05
RFR	KM	40	India	9869487114204.77	-0.0075	6.43-05
GBR	LDA	40	India	9874439119625.94	-0.0080	6.50e-05
RFR	LDA	40	India	9871376967319.38	-0.0077	6.71e-05
GBR	KM	40	UK	733878513079887.20	-0.04457	0.00050
RFR	KM	40	UK	734093158341149.00	-0.0449	0.00048
GBR	LDA	40	UK	733696751912559.50	-0.04431	0.00093
RFR	LDA	40	UK	734063825080816.80	-0.0448	0.00037
GBR	KM	40	USA	214674949552417.00	-0.0664	-7.13e-05
RFR	KM	40	USA	214605333146567.28	-0.0660	-7.70e-05
GBR	LDA	40	USA	214684580296477.50	-0.0664	-3.81e-06
RFR	LDA	40	USA	214626181042848.03	-0.0661	-0.00034

Table 3: Popularity Supervised Analysis Results

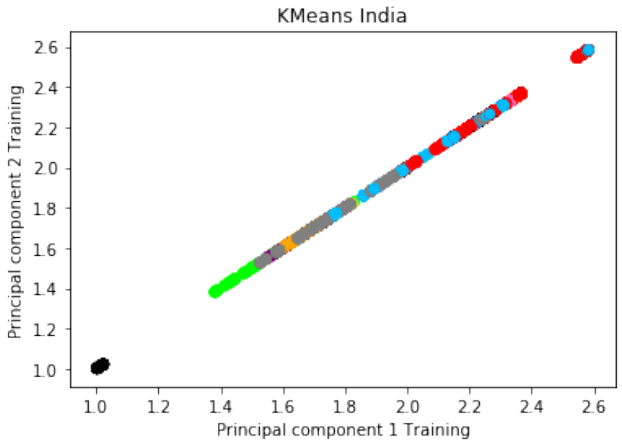


(a) Figure 1a

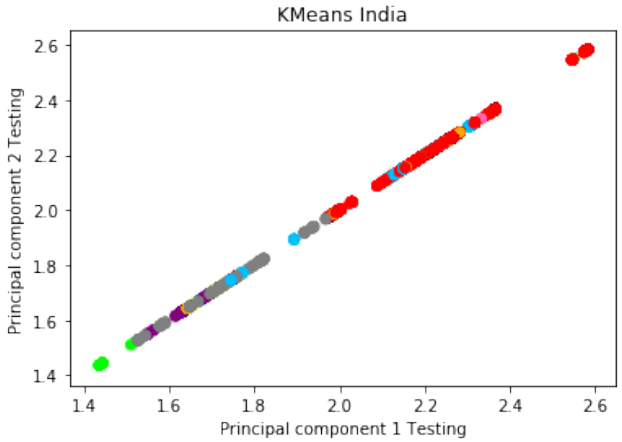


(b) Figure 1b

LDA analysis, showing that K-means clustering did a better job at grouping the data than LDA analysis did. Looking at the silhouette scores for the K-means clustering, the scores for the USA, Canada, and UK were all above 0.9 while the score for India was only 0.84, showing that there are more distinctive clusters of similar videos with similar features in the USA, Canada, and the UK than in India. This trend of India being the weakest country for predicting video clusters holds across the log-likelihood, AIC, and BIC values. For LDA, across log-likelihood, AIC, and BIC, the predictive power of assigning latent clusters was similar for all of the countries, but was the weakest, by a slight margin, for the USA. Looking at Table 1, it can be seen that both Canada and India had "YouTube Rewind" as their most popular video. Furthermore, Canada had "This is America" as it's second most popular video and the USA had "This is America" as the most popular video. None of the top 2 videos in the UK showed up in the top 2 videos in the other countries shown, showing that the UK is the least similar to the other countries from the perspective of trending YouTube videos. Looking at Figures 1a, 1b, 2a, 2b, 3a, 3b, 4a, and 4b, it can be seen that, generally, the distinctive clusters in the training data also show up in the testing data, proving that the K-means (4,5) clustering technique is indeed detecting specific clusters of data. Looking at Table 3, the GradientBoostingRegressor (4,5) and RandomForestRegressor (4,5) had very similar performance across countries and methods. Furthermore, the mean-squared-error, explained variance score, and R2 values were very similar for both LDA and K-means clustering. Based on the table, especially on the MSE scores, the popularity was able to be predicted the best for India, second best for Canada, third best for the USA, and worst for the UK. Looking at Figure 5, the top 40 most predictive features for popularity for Canada were all Asian-alphabet words, which perhaps signifies that Asian-language videos are more popular than English language videos, on average across all videos, in Canada. Looking at Figure 6, the top 40 most predictive features for popularity for India were all Indian-alphabet words, which perhaps



(a) Figure 2a



(b) Figure 2b

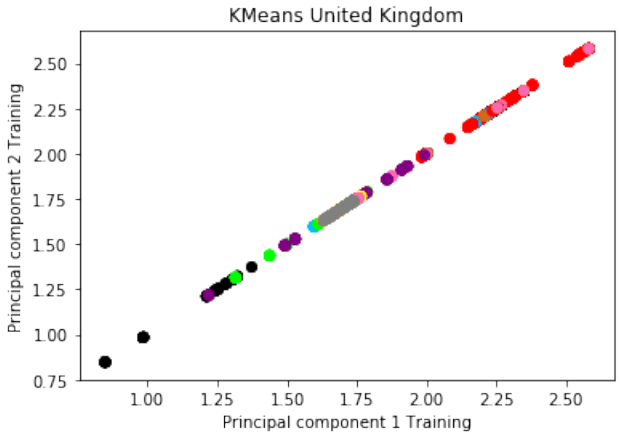
signifies that Indian-language videos are more popular, on average across all videos, than English-language videos in India. Looking at Figure 7, unlike for Canada, the top 40 most predictive features were mostly English words, with a few Asian-alphabet words at well. Some of the English words occurring in the list are ”worldstarhiphop”, ”women”, ”zedd”, ”wild”, and ”xbox”, which helps to glean some insight into the topics that are popular in the UK. In the USA, most of the top 40 predictive features were also English words, as with the UK, but with some Asian-alphabet words as well. Some of the words that occur in the list for the USA are shared with those in the UK list, such as ”zedd”, ”women”, ”wish”, ”wwe”, ”zendaya”, and ”years”. Furthermore, the words ”wireless”, ”wrinkle”, and ”xgames” also occur in the USA list. Since the USA and UK lists are very similar to one another, it can be concluded that the USA and UK likely share more pop culture with each other than either of the countries shares with Canada or India. In conclusion, this paper shows that similarities can be found in trending YouTube data across countries, and that both K-means clustering and LDA analysis are able to uncover patterns in trending YouTube data.

Acknowledgments

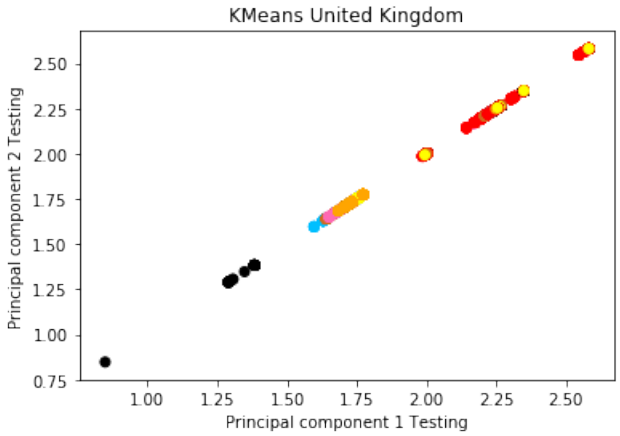
(1) @dataset title=Trending YouTube Video Statistics, author=Mitchell J, author username =DataSnaek, website = https://www.kaggle.com/datasnaek/youtube-new

(2) @API title=YouTube API, author=YouTube, website = https://developers.google.com/youtube/v3/

(3) @program title=Trending-YouTube-Scraper, author=Mitchell J, author username =DataSnaek, website = https://github.com/DataSnaek/Trending-YouTube-Scraper



(a) Figure 3a



(b) Figure 3b

(4) @articlescikit-learn, title=Scikit-learn: Machine Learning in Python, author=Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., journal=Journal of Machine Learning Research, volume=12, pages=2825–2830, year=2011, website = https://scikit-learn.org/stable/ and https://scikit-learn.org/stable/user-guide.html

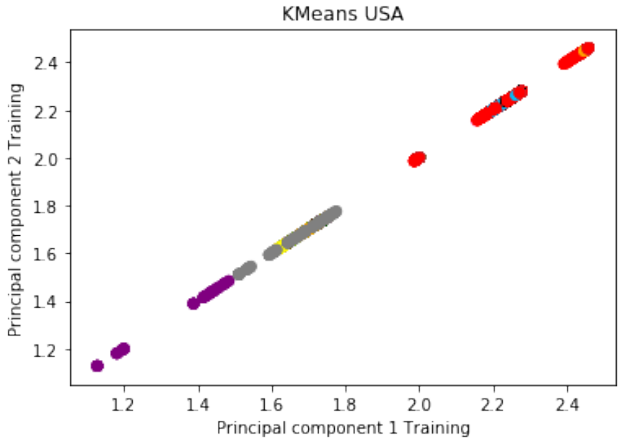
(5) @inproceedingssklearn-api, author = Lars Buitinck and Gilles Louppe and Mathieu Blondel and Fabian Pedregosa and Andreas Mueller and Olivier Grisel and Vlad Niculae and Peter Prettenhofer and Alexandre Gramfort and Jaques Grobler and Robert Layton and Jake VanderPlas and Arnaud Joly and Brian Holt and Gaël Varoquaux, title = API design for machine learning software: experiences from the scikit-learn project, booktitle = ECML PKDD Workshop: Languages for Data Mining and Machine Learning, year = 2013, pages = 108–122, website = https://scikit-learn.org/stable/ and https://scikit-learn.org/stable/user-guide.html

(6) @bookauthor = Kevin P. Murphy, title = Machine Learning : A Probabilistic Perspective, publisher = MIT Press, year = 2012, online publisher = ProQuest Ebook Centra website = https://ebookcentral.proquest.com/lib/princeton/detail.action?docID=3339490

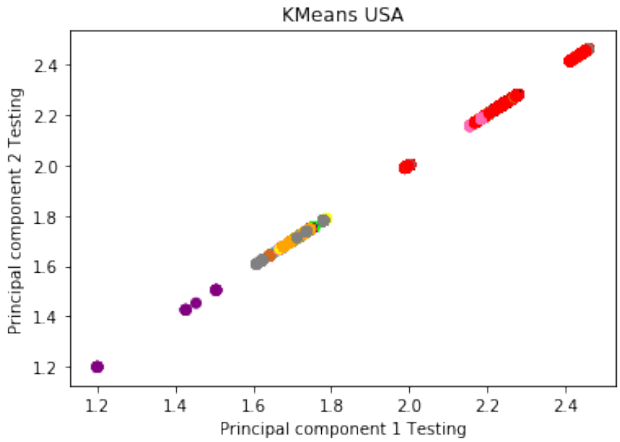
(7) @homework title=Previous Homework Code

(8) @class title=Lecture Notes

(9) @class title=Precept Material

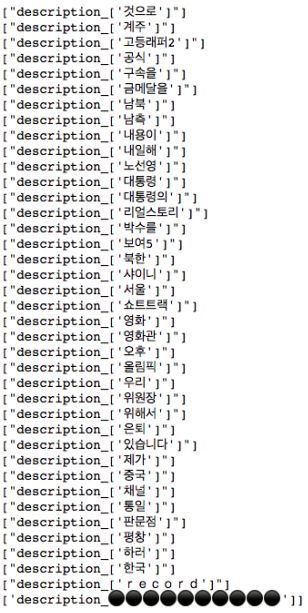


(a) Figure 4a

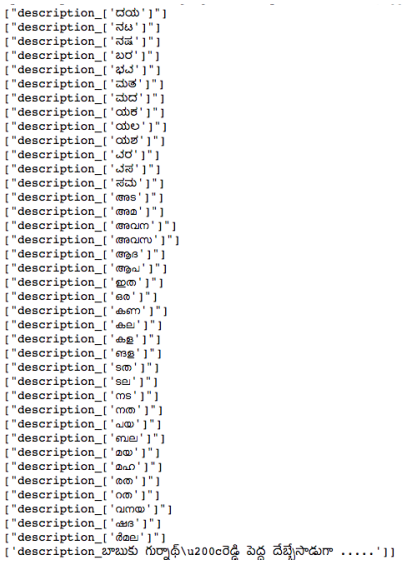


(b) Figure 4b

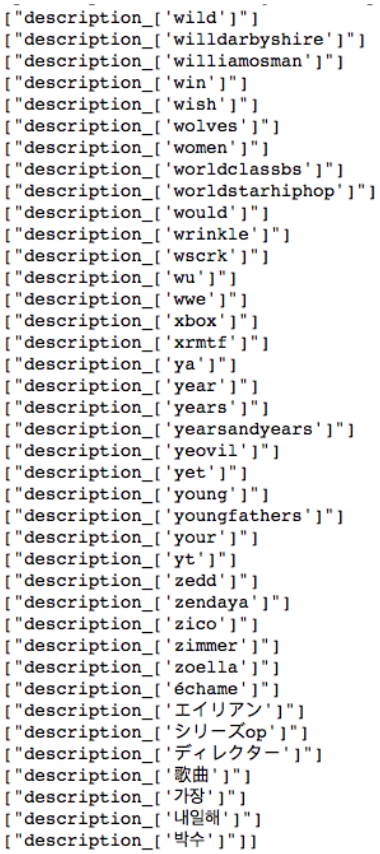
(10) @Misc, author = Eric Jones and Travis Oliphant and Pearu Peterson and others, title = SciPy: Open source scientific tools for Python, year = 2001–, url = ”http://www.scipy.org/”, note = [Online; accessed ¡today¿]



(a) Figure 5: Top 40 Features Canada



(b) Figure 6: Top 40 Features India



(c) Figure 7: Top 40 Features UK

(d) Figure 8: Top 40 Features USA