Group 148: Exploring trending YouTube videos and popular Websites

First Name	Last Name	Share project with ITMD 527? (Y or N)
Alisha Anna	Jose	N

Table of Contents

1. Introduction

Web analytics is gaining popularity as most of the businesses now rely on their online portals and advertising by means of online media and social networking. Hence, analyzing web traffic trends and devising strategies to gaining popularity in the online space is the top concern for most businesses. Gauging web traffic and popularity trends can reflect the changing trends in the people's interests and, at the same time it can be useful for market research.

2. Data

The data sets chosen for analysis are the following:

1. **Popular websites across the globe:** The dataset contains top 50 ranked websites from each of the 191 countries (9540 records).
   https://www.kaggle.com/bpali26/popular-websites-across-the-globe

2. **Trending YouTube video statistics:** The dataset includes several months of daily trending YouTube videos (with 200 listed trending videos per day) for USA, Great Britain, Germany, Canada and France (each in a separate file, with video categories in corresponding JSON file) (5 files with around 23500 records each, plus 5 corresponding JSON files)
   https://www.kaggle.com/datasnaek/youtube-new

3. Problems to be Solved

1. Predict range of traffic rank of popular websites.
2. Predict privacy category of websites with 'Unknown' category
3. Explore website data for identifying any hidden clusters
4. Predict YouTube video category based on title and channel title separately to know if title alone, or channel title alone can give any insight on the video category

4. KDD

4.1. Data Processing

Data Set considered for analysis:

- Popular websites (3401 records)
- Trending Youtube videos_USA (23,363 records)

Implementation:

- Python 3.6, ScikitLearn, NumPy, Pandas, RE, NLTK, Matplotlib libraries

Popular Website data

a) Redundant data

The dataset contains daily and monthly pageview/reach values and their percentages, which can be derived from the later. Hence, removed the columns to prevent redundancy.

Also, for analysis of popular websites irrespective of country, Country rank has been removed and duplicate entries (since some websites are popular in different countries) are removed.

The final dataset of unique popular websites consisted of 3401 records.

b) Missing values

Missing values were replaced by the corresponding column Means for various columns such as avg_daily_visitors, avg_daily_Pageviews, Facebook/Twitter/LinkedIn/Pinterest/GooglePlus/StumbleUp likes using Python Imputer class in ScikitLearn library

c) Normalizing Features

Many features used for analysis was scaled using Python StandardScalar class in ScikitLearn library.

d) Encoding Categorical values

Traffic rank was converted to Range values to indicate popularity level as Very High(upto 100), High(upto 1000), Intermediate(upto 10000), Low(upto 100000), Very Low(above 100000).

Labels used for analysis; Traffic rank category, Privacy category attributes are decoded to Numerical values using Python LabelEncoder class in ScikitLearn library.

e) Test-train split by N-fold Cross validation

Since the dataset for analysis was small in size (3401 records), 10-fold cross validation was used to split data into Test and Train sets using Python KFold class in ScikitLearn library.

Trending YouTube video data

a) Extract category title from JSON file
   Category title has been extracted from category_USA.json file by using Python code using Pandas and dictionaries.
b) Non-English characters removal
   The dataset contains a lot of non-English characters, accented characters and special characters, which where removed using Python Re and NLTK library. Also, irrelevant words have been removed by comparing against Stopwords available Python NLTK library.
c) Test-train split by Hold-out Evaluation
   The dataset is large (23,363 records) and hence it was split with 20 and 80 percent test and train sets, respectively

4.2. Data Mining Methods and Processes

1. ***Predict range of traffic rank of popular websites***

After required preprocessing and test-train split, the following Classification models were applied:

a) K-Nearest Neighbor Classifier
   The model was applied using Python class KNearestNeighbors in ScikitLearn library with various K-values, and different distance measures; Manhattan and Euclidean distance to identify the best model. The accuracy obtained by best models and parameters have been tabulated in the Evaluation section.

b) Space Vector Classifier
The model was applied using Python class SVC in ScikitLearn library with two kernels; linear and radial bias function. The accuracy obtained by best models and parameters have been tabulated in the Evaluation section.

c) Decision Tree Classifier
The model was applied using Python class DecisionTreeClassifier in ScikitLearn library with entropy criterion. The accuracy obtained by best models and parameters have been tabulated in the Evaluation section.

d) Random Forest Classifier
The model was applied using Python class RandomForestClassifier in ScikitLearn library with entropy criterion and various no. of estimators. The accuracy obtained by best models and parameters have been tabulated in the Evaluation section.

2. Predict privacy category of websites with 'Unknown' category

Data mining processes applied for Privacy prediction is the same as above.

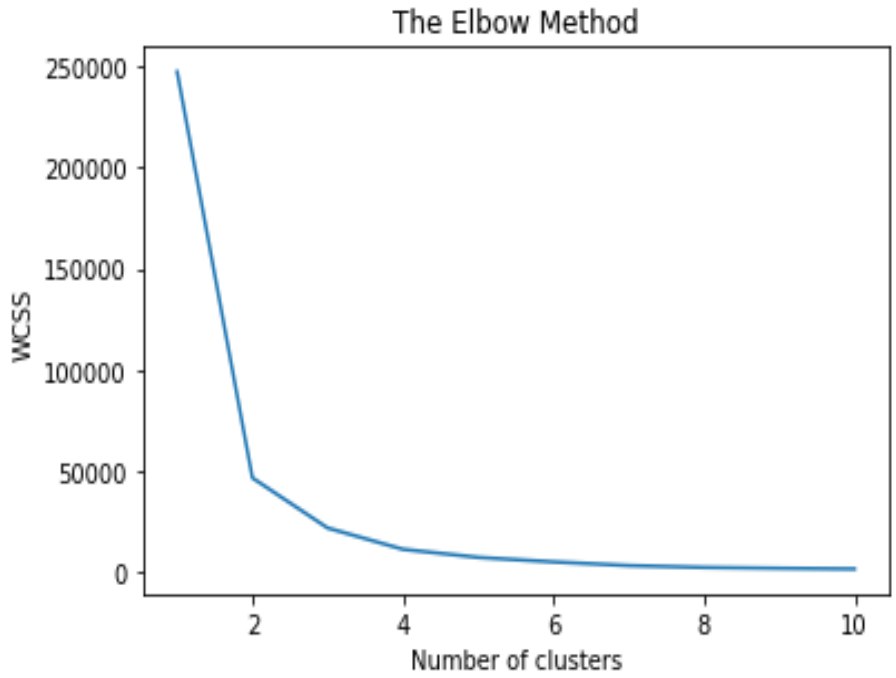
3. Explore website data for identifying any hidden clusters

In order to identify any possible hidden structures, the following Clustering techniques were run:

a) K-Means Clustering
This is implemented using Python's KMeans class in ScikitLibrary. Here, the two pitfalls associated with K-Means which can affect the model performance were addressed as follows:

i) Best initial cluster: Set init value for KMeans() to 'k-means++' to model using the best initial cluster.

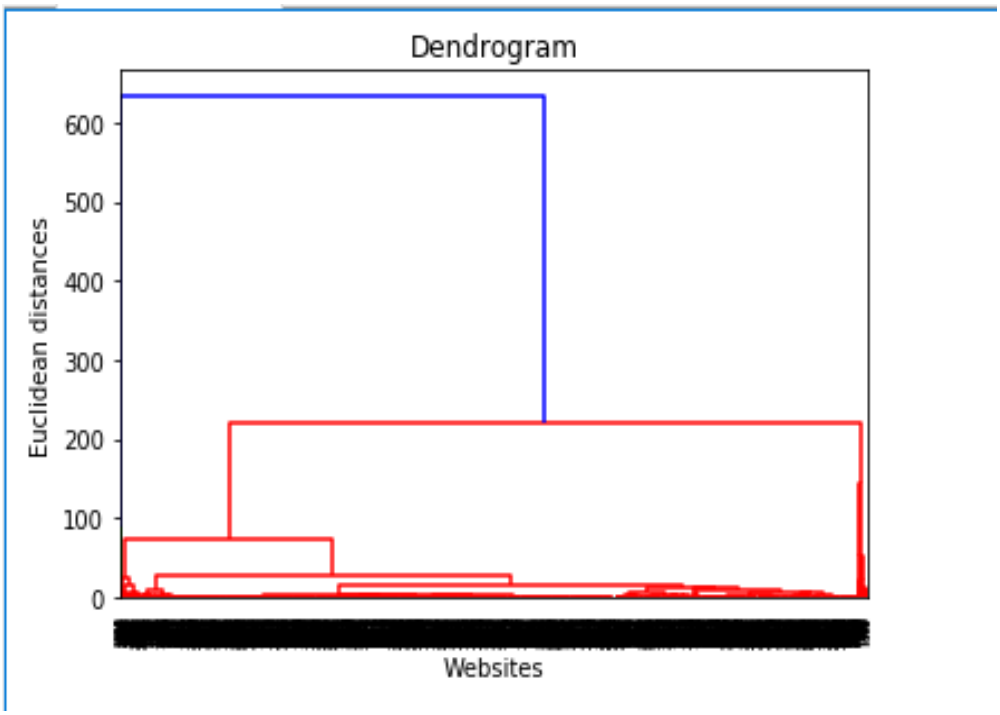
ii) Best K-value: Used Elbow Method to obtain the optimal K-value from 1 to 10 as follows: As per the figure below, we can see Within Cluster Sum of Squares(WCSS) decreases from k=3 onwards, and we chose to model K-Means model with optimal value K=3.



The model obtained has three clusters (0,1,2) and the following was observed:
The clustered seem to give little insight on any hidden structures. This is because the cluster 2 consisted of very few websites which come in 1-5 traffic rank range, the cluster 1 consisted of websites with traffic rank in the range 5-20, and all other websites with ranks between 20-100000 and above comprised the cluster 0.

b) Hierarchical Clustering

The dataset was further explored using Python's sch class in ScikitLearn library for Hierarchical clustering with 'ward' method. As we can see below, the dendrogram give little insight on any hidden structures.



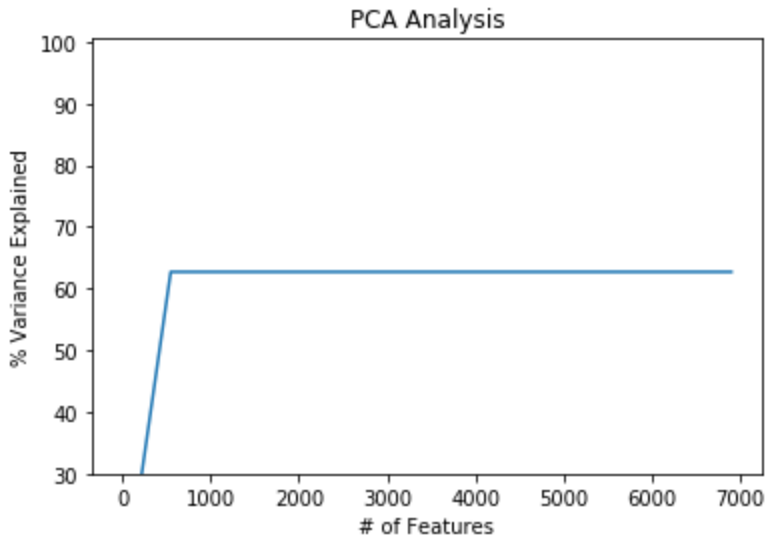
4. ***Predict YouTube video category based on title and channel title separately to know if title alone, or channel title alone, can give any insight on the video category***

   Bag of words model was used to predict Video category based on Video title and Channel Title separately, so that we can identify, whether the title alone, or channel title alone can be used to provide recommendations.

   The Bag of Words models are created as follows:
   a) Count Vector and
   b) TF-IDF vector
   c) TF-IDF vector and applying Dimensionality Reduction using Principal Component Analysis(PCA)

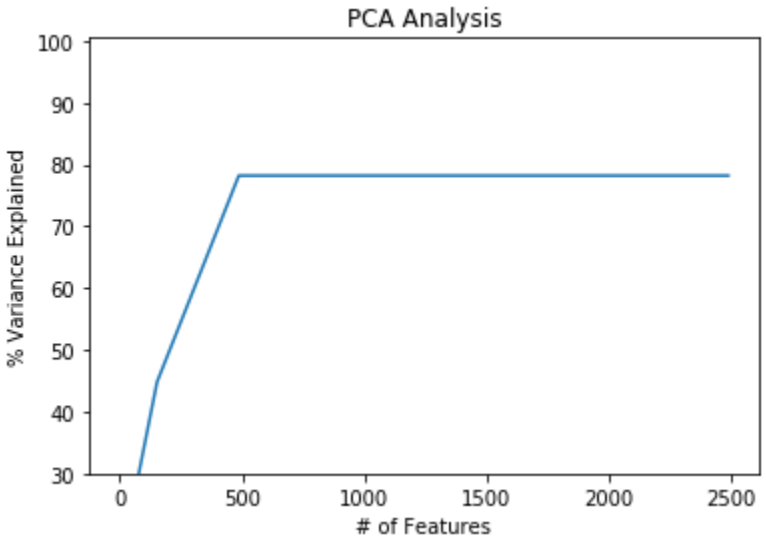
   Optimal no. of Principal Components is identified by plotting the cumulative sum of variances over the no. of Principal Components as follows:

   Title:



We can see, the first 550 Principal components can explain roughly 62% of the variance within the dataset. We gain very little by employing any feature after 550. Hence, we choose 550 principal components as independent variables for our classifiers on the Bag of words model based on Channel title.

Channel title:



We can see, the first 550 Principal components can explain roughly 78% of the variance within the dataset. We gain very little by employing any feature after 700. Hence, we choose 700 principal components as independent variables for our classifiers on the Bag of words model based on video title.

Naïve Bayes, Decision Tree and Random Forest Classifiers were applied on the Bag of words model. The prediction accuracy in each of the above cases are tabulated in Evaluation section.

5. Evaluations and Results

The models are evaluated based on accuracy score metrics available in Python class accuracy_score in ScikitLearn library, tabulated as below for each problem

5.1. Evaluation Methods

1. Predict range of traffic rank of popular websites

Classifier	N	Measure	pred_accuracy
KNN	k=5	distance=manhattan	0.810588
KNN	k=5	distance=euclidean	0.797647
KNN	k=11	distance=manhattan	0.803235
KNN	k=11	distance=euclidean	0.785
SVC		kernel=linear	0.636176
SVC		kernel=radialBiasFn	0.591176
Decision Tree		criterion=entropy	0.764118
Random Forest	estimators: 10		0.802353
Random Forest	estimators: 20		0.819118

2. Predict privacy category of websites with 'Unknown' category

Classifier	N	Measure	pred_accuracy
KNN	k=5	distance=manhattan	0.651643
KNN	k=5	distance=euclidean	0.63783
KNN	k=11	distance=manhattan	0.675315
KNN	k=11	distance=euclidean	0.664632
SVC		kernel=linear	0.661486
SVC		kernel=radialBiasFn	0.664245
Decision Tree		criterion=entropy	0.642212
Random Forest	estimators: 10		0.715583
Random Forest	estimators: 20		0.721109

3. Explore website data for identifying any hidden clusters
   The clusters were unable to give any interesting insights.

4. Predict YouTube video category based on title and channel title separately to know if title alone, or channel title alone can give any insight on the video category

Title:

VSM	Naïve Bayes	Decision Tree	Random Forest
Count	90.6892	97.2602	97.196
TF-IDF	91.1815	97.1104	97.2816

Channel Title:

VSM	Naïve Bayes	Decision Tree	Random Forest
Count	87.9922	94.8202	94.7773
TF-IDF	88.0565	94.8202	94.7559

TF-IDF Based Bag of Words model Accuracy after applying PCA based Dimensionality Reduction:

TF-IDF	No. of PC	Naïve Bayes	Decision Tree	Random Forest
Title	700	37.6926	86.7508	96.7037
Channel title	550	38.4417	50.235	75.9631

5.2. Results and Findings

- Prediction accuracy was best with KNN and Random Forest Model for traffic rank prediction on popular websites data.

⇒ This can be used to identify the level of popularity, can work on improving it, target for advertising strategies.

- Prediction accuracy for Privacy category was better with Random Forest Model, but more features that can attribute to privacy is required to make the best model.

⇒ More data on factors affecting privacy is required to make the best model for privacy prediction.

- Clustering could provide little insights about any possible hidden structures.

- Prediction accuracy for Video category was promising with Random Forest and Decision Tree Model on both the Bag of Words models, based on video title and that on channel title.

⇒ This can be used to categorize videos based on title or channel title, to retrieve statistics of a particular category, and to provide recommendations.

6. Conclusions and Future Work

6.1. Conclusions

⇒ Traffic rank can be predicted with the best model proposed above.

⇒ Privacy prediction requires further data to model classifiers and make predictions.

⇒ Clusters could not provide any useful insights

⇒ Video category can be predicted based on video title or channel title with the best model proposed above.

6.2. Limitations

⇒ Data was not sufficient to devise efficient Privacy prediction model. More information on factors that impact privacy is required.

⇒ Requires better understanding of the various model performance evaluation metrics and other classification models.

6.3. Potential Improvements or Future Work

- Other Dimensionality Reduction techniques can be used, esp. for Bag of Words model to enhance prediction accuracy and faster performance
- Recommendation for videos in each video categories
- Data mining can be carried out for other countries video data set, using which we can identify popular channels.
- Model popular website data with other ensemble classifiers
- Model performance evaluation by other measures, such as Precision, recall, etc.