Utkarsha Devkar

ANALYZING TWITTER-USER DATA

PROJECT REPORT

Introduction

Online communities generate major economic value and currently form pivotal parts of corporate expertise management, marketing and product support. In order to ensure their growth and popularity, the creation of analysis methods that can help community owners and managers to monitor and understand the dynamics of their communities is required. Of particular importance is understanding the behavior that users exhibit in online communities, since changes in these behaviors could affect the utility of the community. This project carries out user-behavior analysis of Twitter, which aims to help the students and other researchers interested in similar research areas to get better understanding of the different user-characteristics associated with different activity levels. This can further help the company develop recommendation engines specific to each group of users.

Methodology

Data Extraction

The user data was extracted from the Tweet object of Twitter users. Using the Twitter API, the user details of 41691 users were extracted. This keywords used for extracting the tweets included a list of the most frequently used words used in tweets as published by Techland.Time.com (*http://techland.time.com/2009/06/08/the-500-most-frequently-used-words-on-twitter/*) where their frequencies are as compiled by the lexicographers of Oxford University Press, based on a sampling of 1.5 million tweets. The data extraction script was developed using Python.

Preliminary Data Cleaning and Analysis

In case of multiple tweet entries by a particular user, the latest tweet corresponding to the particular users were selected while removing the previous tweets. Similarly the data-format of timestamps with respect to user-account creation and tweet-creation were modified to facilitate feature engineering. In addition to the user characteristics (such as friends count, followers count, listed count). A new feature was computed using these timestamps and the total number of tweets issued by a user. This feature is named ‘Tweet-Frequency’ and was calculated as:

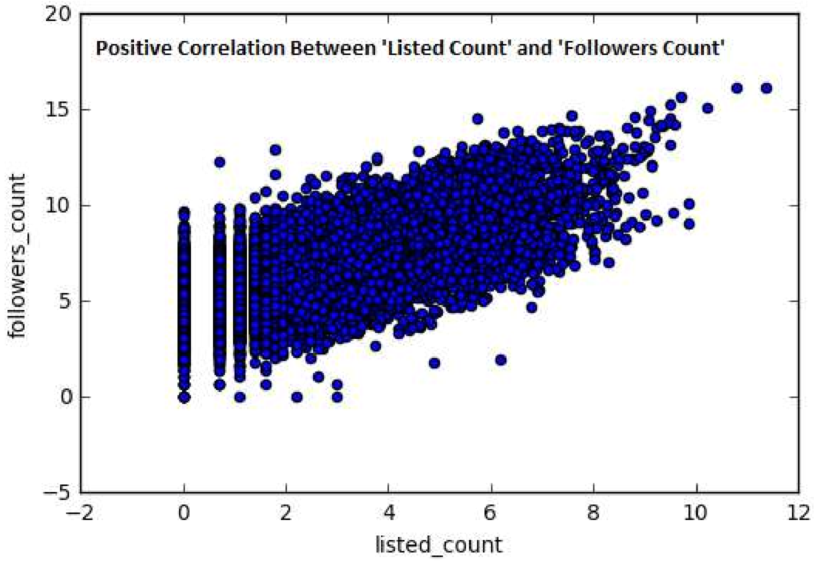
(Total Number of Tweets) / (Number of Days User has been on Twitter)

Here:

Number of days User has been on twitter = (Date of the latest tweet extracted) – (Date on which user created their account). Thus, a final dataset consisting of 41691 users with 9 variables was created.

Statistical Analysis

In order to evaluate correlations among various user characteristics, hypothesis testing was carried out to determine if any correlation exists followed by the correlation coefficient in case the null hypothesis (stating that there is no association between the two tested variables) in rejected and a correlation exists. The Pearson's R Correlation Test (also called the Pearson product-moment correlation coefficient) was implemented which tells us how strong the linear correlation is for paired numeric data. The results indicated that a strong positive correlation was found for the user-characteristic pair of ‘listed_count’ and ‘followers_count’ with a correlation coefficient as high as 0.818 (where listed-count is the number of public lists that the user is a member of and followers count is the number of followers the user-account currently has). This means as the ‘listed_count’ of user increases, the ‘followers_count’ of user increases linearly along with that. Additionally, a positive correlation was found between ‘statuses count’ and ‘favorites count’ of a user account where the null hypothesis was rejected and the correlation coefficient was 0.26. Here ‘statuses count’ is the total number of tweets (including retweets) issued by the user and ‘favourites count’ is the number of tweets the user has liked in the account’s lifetime.

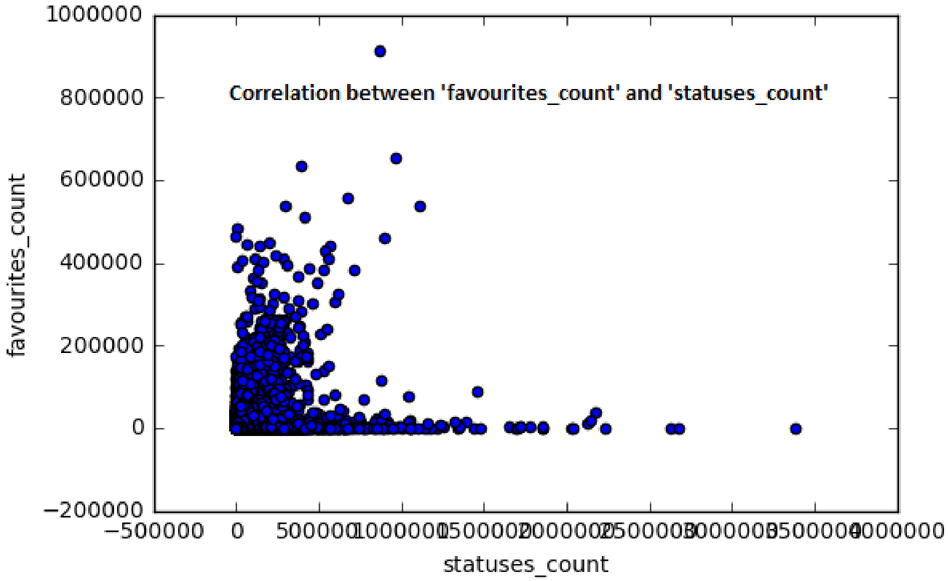


The correlation results:

P value <0.005. Hence we reject the null hypothesis

T-statistic = 291.24

Correlation co-efficient = 0.8188



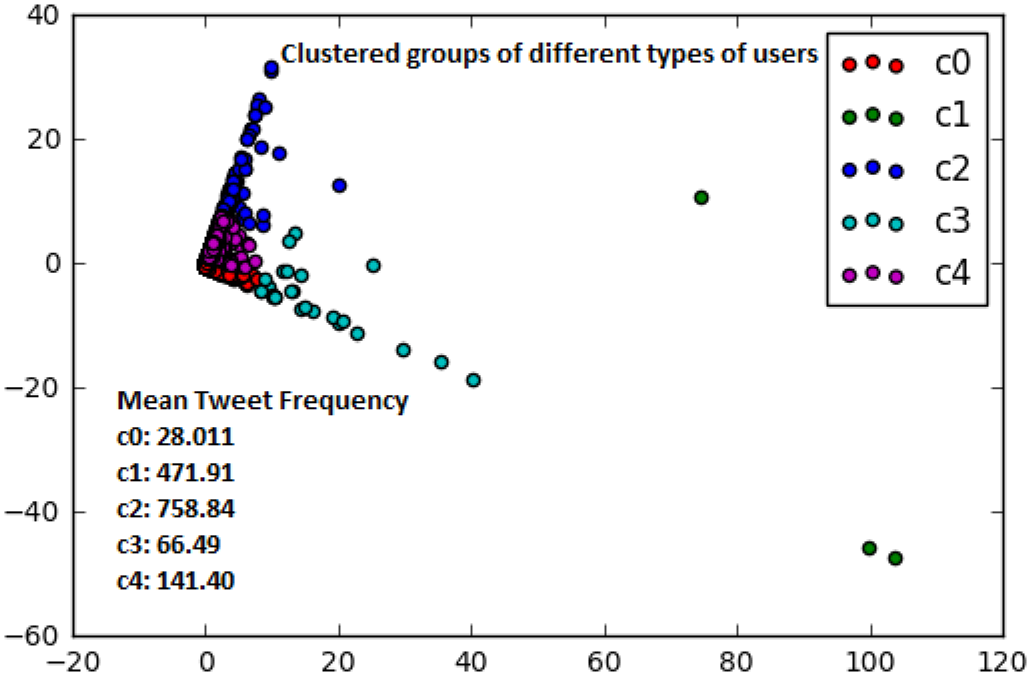
P-value <0.005. Hence we reject the null hypothesis

T-statistic = 56.686

Correlation co-efficient = 0.2675097

Machine Learning:

After the statistical analysis, an unsupervised k–mean clustering algorithm was implemented in Python. K-means is one of the unsupervised learning algorithms that solves the well-known clustering problems. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The K-means clustering algorithm attempts to show which group each person belongs to. However, prior to implementing the clustering algorithm, a Principal Component Analysis (PCA) procedure was applied to the data. A weakness, which is common to clustering in general, concerns the visualization of the obtained clusters. A possible solution is to preprocess the data using PCA. Hence, first, the PCA procedure was applied to the data. Using the principal components the data is mapped into the new feature space. Then, the k-means algorithm is applied to the data in the feature space. The final objective is to be better able to distinguish the different clusters.Thus, PCA was used for dimensionality reduction as a feature extractor, and to visualize the clusters. The finalized cluster is as shown below.



Since clustering is basically used for data discovery rather than prediction, a 'Random Forest Classification' model was then developed to predict the cluster a new user would belong to.

Random forests is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

Final Results

Clustering Algorithm:

The user tweeting-characteristics pertaining to each cluster-group is as summarized in the table below:

Cluster Group	Measure	Favourites count	Followers count	Friends count	Listed count	Statuses count	Tweet frequency
c0	Mean	9052.117668	2.778261e+03	1.063621e+03	37.305	1.750946e+04	28.011
	std	17816.65	1.921853e+04	4.682493e+03	207.748	2.308764e+04	248.585
	Min	0	0	0	0	1	0.0003
	Max	481907.00	7.783810e+05	3.145420e+05	9775.00	1.531560e+05	30628.571
c1	Mean	344.00	9.078444e+06	2.717667e+03	50586.00	1.241307e+06	471.912
	std	294.068	2.341209e+06	3.938101e+03	35370.17	1.854470e+06	715.773
	Min	10	6.385896e+06	1.530000e+02	16299.00	1.666300e+05	56.517109
	Max	564.00	1.063424e+07	7.252000e+03	86948.00	3.382659e+06	1298.412406
c2	Mean	52623.098	7.709320e+04	2.482237e+04	1155.46	1.182828e+06	758.847
	std	159439.15	1.942521e+05	5.612750e+04	2126.26	4.814594e+05	494.194
	Min	0.00	5.900000e+01	0.00	0.00	6.718780e+05	253.830
	Max	912954.00	1.435443e+06	3.930980e+05	13438.00	2.677612e+06	2465.329
c3	Mean	3723.230	1.630586e+06	8.035446e+04	6262.34	1.805733e+05	66.495
	std	9257.125	8.624104e+05	3.245466e+05	6302.699	2.642369e+05	94.922

	Min	0.00	8.617110e+05	0.00	308.00	6.600000e+02	0.486
	Max	35470.00	4.134115e+06	1.663139e+06	27193.00	8.998360e+05	315.157
c4	Mean	44203.291	1.333832e+04	5.718534e+03	320.475	2.090667e+05	141.404
	std	69911.740	4.142851e+04	2.120245e+04	947.585	1.104751e+05	159.115
	Min	0.00	5.000000e+00	0.00	0.00	9.655000e+04	29.571465
	Max	632481.00	6.461480e+05	5.245380e+05	19293.00	6.906680e+05	1809.413

(From above table, std = standard deviation)

Thus, the clustering algorithm is used to cluster users based on their similarity in their tweeting characteristics which is followed by a classification model which is trained by the results of the clustering algorithm in order to predict the cluster group a new user would belong to. Once the cluster group is determined, we are then able to predict the tweet frequency along with the measure of other tweeting characteristics of this user.

The evaluation metric for the random-forest classification model developed are as below:

Accuracy: 0.917836152093

Precision: 0.894565209282

Recall score: 0.917836152093

f1 score: 0.904423923376

The scripts written for data extraction and implementation of algorithms are made available on GitHub to facilitate other researchers for reference.

Future Scope:

Based of differentiating users into different cluster groups, customized recommendation engines can be developed, so as to enhance their activity level. At the same time, Twitter would have statistics of each user group which will facilitate them in better understanding their users.

Conclusion:

This project aimed at understanding the best practices to extract Twitter data followed by implementing unsupervised machine learning algorithm so as to differentiate users on basis of their tweeting characteristics. On training the data based on the clusters formed, the classification algorithm is thus able to predict the cluster a new user would belong thereby helping the company predict their user activity level and characteristics.

Glossary:

1. favourites_count: The number of Tweets this user has liked in the account’s lifetime.
2. followers_count: The number of followers this account currently has.
3. friends_count: The number of users this account is following (AKA their “followings”).
4. listed_count: The number of public lists that this user is a member of.
5. statuses_count: The number of Tweets (including retweets) issued by the user.