Predicting the number of *retweets* and *likes* of a tweet by building a mathematical model using Python

Nashe Mncube

July 12, 2017

Abstract

MUST READ:Before running the programm install the *Tweepy* module. The instructions to do so are as follows:

1. Firstly, for windows 7, open up the command prompt by opening the start button and typing in 'CMD' in the search tool.
2. Open up the app called 'CMD.exe'
3. Type in 'pip install tweepy' in the command line
4. Wait for the module to be installed.

This take approximately a minute to do, and the Tweepy module is required for the function of my app.

1 Introduction

I wanted to know if there was a way to predict how popular a tweet may be by analysing the language within it, and also to build some sort of model which could analyse the language used.

2 How it works?

My program incorporates the twitter RESTAPI as well as other in built Python modules to predict the popularity of a user's tweets.

2.1 The Twitter REST-API

The REST-API is the most important part of this program. This API allows the program to read the tweets of a user as well as access any data which cannot be edited but can be read such as retweet count, favourite count, number of followers e.t.c. The first part of my program concerns the authentication of the program being used. This is required for the user to allow my app to read their tweets. After that is achieved the program uses the data to build a suitable mathematical module. This process is outlined within the program comments.

2.2 Using the data to build a mathematical model

This process is also outlined in comments within the code, but the outline is as follows. The program converts word/text data to numerical data. From working with numerical data and using a statistical model I can look at the distribution of language within a tweet and use that to and relate that back to the number of tweets and favorites and likes within a tweet. From this I build a model of the relationship between the likes an favorites data.

2.3 Predicting the user's tweet popularity

After this model is built, predicting the likes and favorites of a user becomes simple. This part of the program just requires the user to insert what they want to tweet and my program will predict the tweet's possible retweet and favorite count. And from then on the app provides options for the user in terms of trying another tweet or leaving the app.

3 Limitations

The biggest limitation of my app is that it doesn't incorporate the wide range of variables that are involved in determining the tweets popularity. These variables include the users follower count, the number of times they typically tweet etc. This limits my app's functionality as a prediction of retweet and favorite count. It may incorporate these variables in the future but I still my app is still a good predictor for the semi-average twitter user.

Another limitation is the statistical model which is used to for prediction. I found for the test data(my own twitter account) I used, it was accurate. But my own twitter account isn't representative of the all twitter accounts. To overcome this problem in the future I believe that some sort of machine learning method could be used on eve larger data sets. This would allow for the program to refit and improve the model I used without any user input besides their twitter data. However, I don't think I could have reached a large enough scope of people to obtain this data before the project deadline but it is certainly something I believe would've have helped.

4 Conclusion

In conclusion, I believe that although my program does require further improvement to better improve it's text analysis and prediction, I believe I have set decent foundations for the program as it stands to be further improved.

References

[1] https://dev.twitter.com/overview/api/tweets

[2] https://dev.twitter.com/oauth/3-legged

[3] https://dev.twitter.com/oauth/application-only

[4] https://dev.twitter.com/oauth/overview

[5] https://dev.twitter.com/overview/api/twitter-libraries

[6] http://tweepy.readthedocs.org/en/v3.5.0/

[7] http://tweepy.readthedocs.org/en/v3.5.0/api.html

[8] http://cs.brown.edu/courses/csci1951-a/assignments/assignment6/

[9] http://xpo6.com/list-of-english-stop-words/

[10] https://www.youtube.com/watch?v=YeIWV9KOofU

[11] https://www.youtube.com/watch?v=qZgx0pMR-Ps

[12] http://www.toptal.com/machine-learning/machine-learning-theory-an-introductory-primer

[13] https://twittercommunity.com/t/how-to-get-favourite-count-of-tweet/8369

[14] http://docs.scipy.org/doc/scipy/reference/tutorial/interpolate.html

[15] http://cs229.stanford.edu/proj2012/HaponenLindell-PredictingMovieandTVPreferencesfromFacebookProfiles.pdf

[16] http://cs229.stanford.edu/proj2012/BarthelemyGuilloryMandal-UsingTwitterDataToPredictBoxOfficeRevenues.pdf

[17] http://cs229.stanford.edu/proj2012/ChrzanowskiLevick-UsingTwitterToPredictVotingBehavior.pdf

[18] http://cs229.stanford.edu/proj2012/PatersonZhangMwangi-RecommendingMoviesAndTVShowsBasedOnFacebookProfileData.pdf

[19] http://cs229.stanford.edu/proj2012/VenkatesanMai-RecommendTVShowsMoviesBasedOnFacebookData.pdf