

CS 7930 Social Media Mining

Homework 2

Gopal Menon

March 26, 2016

1. **Task 1** These are the five features I decided to use in order to distinguish between legitimate users and spammer, and the reason why:
 - a) **Number Of Followings:** I reasoned that a legitimate user would follow the users he was interested in, while a spam user would follow as many people as possible. I have seen from personal experience, that my Twitter account is followed by some people who do not seem to be legitimate users.
 - b) **Number Of Followers:** I thought that a legitimate user would have some followers, while a spam user would not have many. It was possible that spam users would follow each other, but I thought I would include this feature all the same.
 - c) **Number Of Tweets:** I assumed that a spam user would have a large number of tweets, while a legitimate user would only have a reasonable number.
 - d) **Number Of Url Tweets:** Since spam users try to sell services or products, I decided to have a url count in tweets as a feature in order to identify spammers.
 - e) **Change in Following:** Following churn would be a good indicator of a spammer and so I decided to use this as a feature.

In order to extract the features from the training and testing data sets, I used Microsoft Excel to import the data as tab separated text. In the case of the file with the following count, I used both tabs and commas as separators while doing the import. I used the Excel Search function to look for a url within the tweet text. Then I used a pivot table to summarize the data so that I got the total count of tweets containing a url for a user. For finding the number of followers, I used the absolute value of the difference between the first and last following counts. Once all data files were imported into excel and the url containing tweets were summarized, I pasted the data side by side onto a single worksheet. I used an Excel formula to identify places where the Twitter User Id did not match on the same row as some data was missing. For these cases, I aligned the data by shifting it down, till the Excel formula showed a match.

Once the data were available, I saved the Excel worksheets for training and testing data as tab separated text files that I could import using Python. The extracted data contained the columns listed above and in addition, also contained columns for length of

screen name, length of user profile description and number of tweets in the tweets file per user.

2. **Task 2**

3. **Task 3**