ML 10-805 Project: Topics Authority Detection and Sentiment Analysis on Top Influencers

Manuel Diaz-Granados	mdiazgra@andrew.cmu.edu	10 - 805
Shubham Anandani	sanandan@andrew.cmu.edu	10 - 805
Chaitanya Modak	cmodak@andrew.cmu.edu	10 - 805

- **Dataset**: Initially our project will be focused on Twitter. Twitter is a platform that allows fetching the tweets using the Twitter API. We found a Python wrapper of this API which makes getting a lot of data fast and easy. The Python wrapper is located in GitHub at [1] Also, we found a Twitter data set from Arizona State University [2], this data set contains two files as described in [2]:
  1.**Nodes.csv**: it's the file of all the users. This file works as a dictionary of all the users in this data set. It's useful for fast reference. It contains all the node ids used in the dataset.
  2.**Edges.csv**: this is the friendship/followership network among the users. The friends/followers are represented using edges. Edges are directed.
  Nodes.cvs contains 11316811 of instances and Edges.cvs contains 85331846 number of instances.

- Project idea:
  - In this project, we will build a robust and highly accurate authority detection platform, mainly using the idea at [3] in which they investigated the dynamics of user influence across topics and time of Twitter users. The project doesn't end when we found the main Twitter authorities or influencers in the topics of interest. We also want to determine the view of this "leaders" and if they are influencing other people in a positive or a negative way with respect to the topics previously mentioned. To this end, we want to determine the authorities with the most influence and also identify their viewpoint (sentiment analysis).

- Softwares and toolkits
  - There are a number of existing language modelling toolkits that we can use for analyzing the text of the tweets. We plan to use python for most of our implementation. Scikit Learn has a number of useful clustering modules that we can use for topic analysis. Some other platforms that we intend to use include Word2Vec, which provides shallow neural networks that map words to vectors in semantic space. We can implement concepts such as Latent Semantic Analysis, Latent Dirichlet Allocation, Brown clustering etc. both using existing libraries as well as building our own. Finally, we have a number of sentiment analysis tools for Twitter that we can use to analyze emotional outcomes such as optimism and pessimism that result from tweets from particular figures.

- Teammates and work division (We welcome students from 605 in our project).
  - We expect projects done in a group to be more substantial than projects done individually. Between the members of our group in the 805 sections, the work will be distributed equally, each of us taking care of an independent task. One task is to set up our cloud environment to run our ML algorithms. Second is to prepare the data set in a way that the workers in the cloud can fetch it and run independent tasks so we can gather all the results at the end. Third is to develop the main algorithms to conduct training and testing on the collected datasets. In conclusion, we will conduct useful analysis of the results obtained from the algorithms.

- Baseline techniques:
  - The baseline technique is the one used in [3] in which they calculate the in-degrees of a twitter user and his out-degree, given this metrics and his retweets or likes a rank is built. Then a baseline of sentiment analysis is the used of a simple bag of words corresponding to the 'positive' or 'negative' sentiment of with respect of a certain topic, using naive bayes for example.
- After mid report baseline
  - On top of the authority detection, explore sentiment analysis of the found "authorities" and if time permits, extend the work to other social platforms like Facebook.

References:

[1] https://github.com/bear/python-twitter
[2] http://socialcomputing.asu.edu/datasets/Twitter
[3] Measuring User Influence in Twitter: The Million Follower Fallacy Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, Krishna P. Gummadi, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media