Impactweets: New Techniques for Filtering Interesting Tweets

Finding the big picture of tweets for a particular user. We will try to find new ways of filtering interesting tweets for individual users using the characteristics like RT, Mentions, Topics, URLs and categorization of tweets.

Team members and roles:

- Seema Puthyapurayil :: Visionary, Network Analyst
- Ignacio Pérez :: Developer, Classifiers and Content Analyst
- Corey Hyllested :: Developer, Data Analyst
- Yao Yue :: Twitter Project Mentor

Project strategy:

We will find:

i Interesting conversations based on replies in the user's network
ii Interesting tweets based on links, hashtags and retweets
iii We will try to give a hint of the possible topic of the tweets using Naive Bayes Classifiers and Support Vectors.

We have divided these tasks as they are modular and assigned one person work on conversations and retweets, one person to work on hashtags and links and one person to work on the classification of tweets.

Project timeline: (Green : Completed, Yellow: Half done, Red: Not started)

Date: Oct 27th

Tasks due:

- Finalizing data formats, programming language of choice
- Identifying potential risks and planning mitigation measures:
    - Issues recognized were each of us working on different sets of data, whether categorization based on Naive Bayes would be achievable, training set for Naive Bayes and rate limiting as we are dealing with REST API. We have found solutions to each of these problems. We would run one data collection program and include ways to dodge the rate limiting in it. The problem for the training set for the Naive Bayes was solved by filtering tweets related to a topic by tracking keywords related to the topic.

Date: Nov 2nd

Tasks due:

- Identifying a list of users whose networks we will analyze.

    - CoreyHyllested, Coldspire, aditi_deshmukh, myy_precious, imedha, akthiel
- Code for extraction of data in the format decided.
    - We are using a mix of python and java, depending of the availability of libraries related with the different aspects of the data analysis required.
- Visualize the format of the analysis output.
    - We already decided that we will generate a text report for our users showing the information that we found inside their network:
        - Interesting tweets based on conversations, hashtags, links, retweets and a classification for each of these tweets.
- Logical framework of the code to perform the analysis should be ready by now
    - We have defined several tools to analyze the FoF network for a user
        - We are using our own Naive Bayes Classifier to classify tweets, implemented in a combination of PIG and Java code.
        - We are using NetworkX in Python to analyze the network and find relations between users.

Date: Nov 9th
Tasks due:

- Run the analysis code on the first set of data.
    - Finding tweets by order of retweets is working
    - Finding conversations for one user's network is working, need to scale it to work for every user's network and counting @mentions in each.
    - Aggregation of most popular links and hashtags is working.
    - Classification is working based on a dataset of 60,000 classified tweets.
- Identify problems in execution and possibly correct them before the iteration
    - We need to improve the data gathering process. the restrictions in the Twitter API are delaying our planned timeline. Hence we plan to run the data collection overnight for different users. We have decided to test out all the analysis with one user's data and test it, find issues

Dates: Nov 9th to 25th

- Testing recommendations with users. (In Progess activity)
- Iteratively refine analysis code and data based on user-input.(In Progess activity)

Work Percentage:

Tasks	Corey	Seema	Ignacio
Data Collection	60	20	20
Literature Review	30	30	40
Analysis (Network, Topics & Data Gathering)	33	33	33
Meetings	33	33	33
Coding	33	33	33
Project Report	20	40	40

Literature Review:

- **A plan for SPAM**
  http://www.paulgraham.com/spam.html
  This article explains how to decide if one particular email is SPAM or not, based on Bayesian Filtering. Using a set of previously classified mails, it is possible to analyze the frequency of words and finally to take that decision. We will show in this work that the same technique is applicable in this project to decide if a tweet is about certain topic or not, using topics like politics, sports, technology, business, etc.
- **Better Bayesian Filtering**
  http://www.paulgraham.com/better.html
  This is a second article about bayesian filtering and explains some new ideas aiming to improve the bayesian classifiers. In the document, it is mentioned that the structure of an email is important to decide if that email is SPAM or not. In the context of our work, we believe that it is possible to improve the bayesian classifier based in the structure of a tweet, using features like places, user, links, hashtags and others. We will try to analyze the structure of a tweet to improve the bayesian classifier.
- **Twitter Sentiment Classification using Distant Supervision**
  http://cs.wmich.edu/~tllake/fileshare/TwitterDistantSupervision09.pdf
  This article extends the idea of using Naive Classifiers and Support Vector Machines to classify tweets. In particular, instead of topics, the article aims to classify tweets based on sentiment: positive or negative sentiment about a subject. We will try to use support vectors in this project to improve the

results of the bayesian classifier and to improve the results in the search of impact tweets.

- **Don't follow me: Spam detection in Twitter**
  http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5741690&tag=1
  This article explains how to use automatic classifiers to filter spam in Twitter, but it also suggests an interesting idea about how to use the network of friends and their relationships to manage and process SPAM. This network analysis could be useful in our project to suggest new conversations and topics to the users based on the information in the Friends of Friends (FoF) network.
- **Why Do People Retweet? Anti-Homophily Wins the Day! Sofus A. Macskassy and Matthew Michelson**
  http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2790/3291
  Here authors are trying to analyse the information diffusion through retweets. They come up with four models: 'General'(retweet anything randomly seen, higher probability of what occurs first on their timeline), 'Recent Communication'(retweet a tweet from user who they have communicated with in the past 24 hours), 'Topic' (related to topics of their interest) and 'Profile'(retweet tweets by user whose profile matches theirs). In our project we are trying to enhance the users twitter experience by bringing impactful tweets to the front, based on recent communications (conversations), topics of interest to the user and thus helping users to not rely on the 'General' model of having high probability of retweeting what occurs first. If time permits we will also think about classifying users. The authors here have used a very interesting model based on Wikipedia entries to classify users but due to time limitations we will be using our NB classification which is a pretty strong starting point.
- **Unsupervised Modeling of Twitter Conversations: Ritter, Alan; Cherry, Colin; Dolan, Bill** http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=16885300
  The researchers in this paper are looking to find acts that can incite conversations called 'dialogue acts' within twitter data. They have used an unsupervised approach to find the words and tokens that are dialogue acts. Tweets are classified as 'Status' which is the traditional update about what the user is doing, 'Question to Followers' represents a user asking a question to their followers, 'Reference Broadcast' which contains URLs meant to be broadcast to the general public on Twitter, 'Question' which is a question to the general public, and 'Reaction', 'Response' and 'Comment' which arise as response to any one of the type of questions, statuses or broadcasts. They

then generated a very interesting list most frequent words and tokens that come up in each of these categories. In terms of our project, in the future we could use this corpus of words to find what type of tweets generates most number of responses. Another interesting find from this paper is that the majority of conversations on Twitter are very short; those of length 2 (one status post and a reply) accounted for 69% of the data that they collected. For our project this means that if a tweet contains more than 1 reply it lies in the higher 31% of tweets which one or more user has expressed interest in. Hence tweets with more replies will be significantly of more interest to a user. Also they find that users are very likely to respond to URLs that interest them much more than the regular status messages the other users post. This validates our decision to find links that people talk about and measure their importance.

- **Beyond Microblogging: Conversation and Collaboration via Twitter: Courtenay Honeycutt, Susan C. Herring**
  http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4755499&tag=1
  Twitter was initially designed for people to answer the question "What are you doing?", but as is evident from the recent avatar of Twitter, tweets are being used in many more ways than its traditionally imagined usage. This paper explores the usage of the '@' symbol in tweets. They discuss the high addressivity of the symbol, how it asks for and initiates a response and thus promotes information diffusion. Many research questions are discussed but the most pertinent question for us is about how long, and how coherent, are interactive exchanges, and to what extent do they make use of the @ sign. The paper also discusses the purpose and content of tweets with and without the @ sign. The paper particularly concludes that interface views that allow all conversations to be viewed together on a user□s homepage, would increase the usability of Twitter for conversational purposes. This helped us think about links with the @ symbol, and the .@ symbol, which is used when a user intentionally puts a dot before the @ symbol to make increase its audience.
- **Networks, Crowds, and Markets: Reasoning About a Highly Connected World: David Eakley and John Kleinberg**
  http://www.cs.cornell.edu/home/kleinber/networks-book/
  We are using this book and the lecture by Stan Nikolov (http://blogs.ischool.berkeley.edu/i290-abdt-s12/2012/10/26/video-lecture-information-diffusion-on-twitter-by-snikolov/) to study the diffusion of tweets. The concepts of homophily, and the interplay of selection and social influence, explained in the book are especially important to us, to find

the most impactful tweets from a user's timeline. Stan also used a very novel way to display diffusion using Gephi and such infographics could really help a user visualize his network. Although this is out of scope for us now, we would like to accommodate this if we have the time.

- **Social networks that matter: Twitter under the microscope. - Huberman, Romero, and Wu**
  Scholars, advertisers and political activists see massive online social networks as a representation of social interactions that can be used to study the propagation of ideas, social bond dynamics and viral marketing, among others. But the linked structures of social networks do not reveal actual interactions among people. Scarcity of attention and the daily rhythms of life and work makes people default to interacting with those few that matter and that reciprocate their attention. A study of social interactions within Twitter reveals that the driver of usage is a sparse and hidden network of connections underlying the "declared" set of friends and followers. In the context of our project, we are interested in understand how those small interactions could have an interesting impact in the process of filtering of important tweets to the users.

Further readings (read but not discussed here):

- Twitter Power: Tweets as Electronic Word of Mouth - Jansen and Zhang
- Everyone's an Influencer: Quantifying Influence on Twitter Bakshy, Hofman, Mason and DJ Watts.
- I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience - Marwick and Boyd.

Accomplishments to date on the project:

Further refinement of project goals:

After discussion with our mentor, we have refined our collection of links, hashtags, most retweeted to be from the friends network only. Only for conversations, as a user's friend can have conversations with his friends which many not be in the users network, we will need to use the friends-of-friends network.

We have also decided to explore the area of images posted on twitter, to find the interesting images posted in a users network.

We have found interesting information during the classification of tweets. With a small amount of tweets classified to create a base model, (60000 tweets) the classifier is already discarding topics that are not related with a tweet with a 90% of

precision, using an example dataset of 200 tweets of one of our users. We are not filtering the tweets, nor removing stopwords.
We are planning to use external libraries to classify and find the topics of a Tweet. The benefit of using those libraries is that we will use a preexisting set of categories to classify tweets.

Data Gathering

Currently, we have used a very basic approach for collecting data. For each user, we're downloading the tweets from their friend-of-friends network. We're tracking the protected accounts and the spam accounts. To ensure that we're not wasting effort, we reduce the space usage by enforcing uniqueness. For a single user with ~100 followers, the FoF network encompasses 100K. We collect the last 200 tweets per user.
In the context of the data gathering for classification purposes, our initial approach was to manually try to classify an interesting amount of tweets. However, the resulting base model for the Naive Bayes classifier was not useful to classify the "incoming" tweets. Instead of that, we are using the Twitter's search API to download tweets related with certain topic. For example, under the "politics" topic, we are downloading all the tweets that the search API retrieve with the search words obama, romney, politics, election, vote, government, etc.
With a base model for the Naive Bayes classifier using 60,000 tweets, the results were immediately more interesting. We plan to create a base model using at least one million tweets gathered in different dates. This is an example of the current list of topics and keywords that we are using to build the base model.

Challenges and Next Steps:

1. We realized that favorite count is not yet implemented in the API so we will not be able to rank tweets based on favorite count.
2. We decided to do tweet metric post processing in Pig. However, the API only provides the tweet ID and in reply to tweet ID for every tweet. After a lot of trial and error iterations using all kinds of joins in Pig, there was still no way to fold up the tweets and get the entire conversation as an output. After further research it was found that we could find the related tweets to every tweet ID using predefined API functions in Java. Using that in conjunction with reply
3. Data collection takes a long, long time. We may target fewer users.
4. It is possible to use the topic classifier to build a vector with the interests in each topic for any particular user. Then, we will be able to compare different users in terms of the weights of their interests. We can try to use Support Vector Machines to compare different vectors of users, if time permits.

Algorithm development:

For our project, most of the algorithm development is for the Naive Bayes classification. Other development pertains to finding metrics for a tweet. Currently, we're trying to generate a model of what tweets would be important, interesting, jejune, spam by thinking about the available information and the metrics that could signal such facets.

In the context of classifying tweets, we worked in three phases:

1 Retrieval of tweets:
We are using the Twitter's Search API to download tweets related with certain topics. The format of the resulting types is: (TweetID, topic, TweetText)
Reference for the classified tweets:
https://raw.github.com/iaperez/ABDTProject/master/data/alltweets
Reference for the download application:
https://github.com/iaperez/ABDTProject/blob/master/retrieval/src/UsersMentionDownload.java

2 Build a base model with the probabilities for each possible ngram to indicate a topic. Reference:
https://github.com/iaperez/ABDTProject/blob/master/pigfiles/BaseModel.pig
In the context of building a base model, and considering as input a set of classified tweets, we need to calculate for each ngram its relation with each possible topic. In Pig, we implemented this in the following steps:
    - Load input data
    - Counting all the tweets used as input. (ctTweet)
    - Counting for every topic the amount of tweets related with that topic. (ctTweetTopic)
    - Counting the number of times in which an ngram is associated with a particular Topic. (ctNgramTopic)
    - Counting the number of tweets that contain a particular ngram. (ctNgramTweet)
    - Finally we estimate the probability that a ngram is referencing a particular topic (based on "A plan for SPAM" article)

$$\frac{\frac{ctNgramTopic}{ctTweetTopic}}{\frac{ctNgramTopic}{ctTweetTopic} + \frac{ctNgramTweet - ctNgramTopic}{ctTweet - ctTweetTopic}}$$

3. Analyze new tweets using the base model.

Reference:https://github.com/iaperez/ABDTProject/blob/master/pigfiles/tagsAnalisys.pig

We are using the model proposed in the literature ("A plan for SPAM" article) to calculate the probability for each incoming tweet to be related with a certain topic: every tweet will be analyzed in terms of its Ngrams, and the total probability of that tweet to be related to a particular topic is:

$$Tweet\ Probability\ per\ Topic = \frac{\prod ProbNgramTopic}{\prod ProbNgramTopic\ + \prod(1 - ProbNgramTopic)}$$

With ProbNgramTopic as the probability that the ngram is related with that topic (generated from base model). When an ngram is not found in the base model, we are using a probability of 0.4.

Example of the results of this phase:
https://github.com/iaperez/ABDTProject/blob/master/data/twitterclassificationresults.txt

- **Coding**:
  - https://github.com/iaperez/ABDTProject/
  - https://github.com/CoreyHyllested/impactweets
  - https://github.com/seemahari/impactweets

- **Interface design:**
  We have decided that we will generate a text report for our users showing the information that we found inside their network. The report will contain all the interesting tweets in the users network based on conversations, hashtags, links, retweets and a classification for each of these tweets.