Pranay Rao - 50169207
Sourav Mukherjee - 50168377

1. Summary & Related Work / Novelty of technique:

Understanding the reaction of the people for various events have become a critical part of any organization. At UB, events such as the recent racial incidents, need the university's PR machinery to be always active and respond instantaneously to the incidents and mitigate the damage caused to the brand by negative publicity. Twitter is a good source of people's initial and unbiased reaction to such incidents and analyzing them at the onset would have help contain the spread of the negative buzz. Our project tries to capture the sentiment of the social media users using twitter where they express how they currently feel about the university or events revolving around it. Analyzing tweets of millions of users needs parallel processing with the power of distributed computing system where tweets gets stored in multiple t. Although for the scope of the project implementation, we are using a single node, we are trying to demonstrate how the same can be scaled up to multiple nodes using Hadoop.

2. Design Issues:

1. Understanding the underlying sentiment expressed in 140 characters - eg Sarcasms, multiple sentiments in a single tweet - example "hate the weather at UB but love the campus."
2. Time intervals at which the tweets should be collected. Our design is based on the assumption that there would be sufficient amount of tweets if collected over a span of few hours. Conversely if we run for too long we need to understand at what number of tweets would suffice for analysis to get a more real time view. We wouldn't want the API to run for a short time nor do we want it to run continuously and create a process of cleaning up twitter data that may not be relevant in the first place.

3. Algorithm Description:

Even though we will use a single node installation, the steps mentioned below can be implemented on separate compute systems and then the final data can be collated on a single node where analytical reports can be generated.

- Download and Extract Twitter data: Flume can be used as a data aggregator to move data from various sources to a centralized data store. Once API streaming is done, this data would be moved into Hadoop Distributed File System
- We will use a hive .sql script to convert the raw data into tabular data. Since we are employing a lexicon approach, a data dictionary would be used to assign a sentiment score(positive, negative or neutral) to each tweet based on the number of positive words to the number of negative words contained in that tweet, and then load these sentiment values into a table along with other data.

The above two steps can be scaled up to a cluster implementation where each node performs these tasks to ultimately help work on a large data sets.

- Using excel we will access the refined dataset.
- Once the sentiment data has been imported, Powerview will be used to create analytical reports.

Pranay Rao - 50169207
Sourav Mukherjee - 50168377

4. Data sets and Software:

For now we used the Twitter API to extract data in JSON format. The curl command was used to fetch and store data from the API in the VM terminal into JSON format. We have collected around 300 tweets over one hour of running the API. But we will be using Flume to directly ingest this data into HDFS. We collected data from Twitter API based on the following keywords - University at Buffalo, Suny Buffalo, #ub, ub football, ub bulls. We have set up Apache Flume on Hortonworks Sandbox to process the data.

5. Bibliography:

1. https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
2. http://www.sciencedirect.com/science/article/pii/S2090447914000550
3. http://mike.teczno.com/notes/streaming-data-from-twitter.html

6. Appendix:

Dataset:

Example of a negative sentiment-

{"created_at":"Wed Oct 28 14:27:39 +0000 2015","id":659375982193848320,"id_str":"659375982193848320","text":"Welcome to University at Buffalo, where the hardest part about college is finding a parking spot.","source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003eTwitter for Android\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":2901885076,"id_str":"2901885076","name":"Kimmy Maloney","screen_name":"KimmyMaloney","location":"Buffalo, NY","url":null,"description":"I can pretty much promise you I'm gonna find a way to make it awkward.","protected":false,"verified":false,"followers_count":39,"friends_count":45,"listed_count":0,"favourites_count":319,"statuses_count":518,"created_at":"Tue Dec 02 04:17:04 +0000 2014","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/658799359493017600\/zhx2hwe8_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/658799359493017600\/zhx2hwe8_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/2901885076\/1423801396","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":{"id":"7dabbf75534f6cee","url":"https:\/\/api.twitter.com\/1.1\/geo\/id\/7dabbf75534f6cee.json","place_type":"city","name":"Amherst","full_name":"Amherst, NY","country_code":"US","country":"United States","bounding_box":{"type":"Polygon","coordinates":[[[-78.832497,42.948903],[-78.832497,43.069330],[-78.696766,43.069330],[-78.696766,42.948903]]]},"attributes":{}},"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"en","timestamp_ms":"1446042459425"}

Count of tweets: