1. How much did this assignment cost you?

$5.25

2. What question are you attempting to answer via your analysis of the data?

We sought to find general sentiment of the "Twitterverse" with regards to UVA President Teresa Sullivan and Rector Helen Dragas during the president's resignation and successive reinstatement in June, 2012. Specifically, we wanted to answer the question "how did support for Sullivan/Dragas change over time?" To this end, we tried to find the number of positive and negative tweets with respect to each of the two parties. The two parties are more broadly defined by treating the Rector and the Board of Visitors as a single entity. These and President Sullivan constitute the key figures in this event. The results can also be used to answer questions such as "how does tweet sentiment reflect things happening in the real world?"

3. What external sources of information/code, if any, did you use? (enumerate the sources, with a brief URL and description of each source and specifically how you used it)

Natural Language Toolkit (NLTK) - http://nltk.org/ - This is a python library that performs a naïve Bayesian Classification, determining for us if a tweet is positive or negative. We should note the classifier was trained on a set of movie reviews provided by the package.

StreamHacker - http://streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naivebayes-classifier/ - We used sample code from here for working with NLTK.

Timeline http://uvamagazine.org/university_digest/article/timeline_teresa_sullivans_resignation_and_reinstatement#.UWtQmqKG0sc - A timeline of the events with which we compared our data.

4. What percentage of the tweets did you discard/ignore? How do you explain/quantify this unused data?

We discarded 28553 (~54.6%) of the tweets in our dataset. Of these discarded tweets, only 2612 contain a reference to either Sullivan or Dragas (an equal number in fact), and we could not determine which was the subject of the tweet. We suspect that the majority of the remainder is unrelated to this case, as in the example of a tweet about Mike Scott being taken by the Atlanta Hawks in the NBA draft. It would appear that the data was curated mostly by looking for hashtags starting with "UVA", etc. There are 44416 tweets of this form alone, meaning there are many things that could slip in that are unrelated to the case of interest. Further, the date range encompassed by the archive extends well beyond the period of focus, but we did not exclude these from our analysis. We chose to acknowledge any tweets which we could associate with one of the relevant parties; all others were discarded.

5. Compare your preliminary version with your final version:

a. What is the (semantic) difference between the code in your preliminary version and your final code?

One of the primary changes to our code includes the addition of a library of terms that we hand-picked to categorize a tweet upon sight, meaning that we are more reliant on instinct than on the NLTK library.

These terms are mainly composed of hashtags identifying pro-Sullivan or anti-Dragas/BOV tweets. Further, we also increased the number of terms used to identify a tweet with an entity before it is passed to NLTK. Our reducer code now also provides us with 4 values: positive and negative for both Sullivan and Dragas/BOV, as opposed to just the net positivity for each of the two parties as was in our preliminary code.

b. How would you quantify the improvement in quality in your analysis?

We originally identified a set of approximately 4000 tweets which had a nonzero but equal number of references to Sullivan and Dragas/BOV. We now process 1400 of these, the rest suspected of being relatively neutral news summaries. Our previous analysis only presented a "net positivity," effectively masking approximately half of our data. Presenting the raw assignments as we do now avoids this masking, allowing us to not only gain that report but also present new ones previously unavailable.

6. What were the pros/cons of using Hadoop to analyze this data – e.g. were there any analyses that you wanted to perform but couldn't due to the structure/nature of Hadoop?

The biggest negative of using Hadoop for this data is that Hadoop does not handle small files very well, as HDFS is based on GFS. This led us to running most of our data through a series of shell scripts instead of on the Hadoop framework. On the plus side, though, the data and analysis that we wanted to perform fit well into the Map/Reduce architecture. We don't feel that we were limited in our analysis by Hadoop.

7. Are your results surprising to you? E.g., did you expect a more conclusive analysis? (did you expect a less conclusive analysis?) Does this make you question the intrinsic value (and/or potential bias) of the data?

We were surprised by several things. For one, there is a surprising dearth of tweets on June 24, when over 2000 people congregated on the Lawn for the "Rally for Honor." This may indicate more problems with the underlying dataset. The positive Sullivan tweets on the 26th coincide with the President's reinstatement, but we are surprised to see so much positivity toward Dragas on the 19th, a day which marked leaked emails between Dragas and Vice-Rector Kington (who resigned on that day) as well as the resignation of University Professor William Wulf in protest of the Board. Carl Zeithaml was also selected as interim president, but we consider this to be a neutral event for this analysis. Perhaps the buzz about the leaked emails, posted originally to Twitter, led to an increased excitement that was interpreted by our system as positivity, and the subject of these tweets is most definitely Dragas/BOV. Overall, we find the number of Dragas-positive tweets to be high, possibly indicating a weakness in our analysis architecture. However, we are more concerned with the intrinsic value of the data. We were surprised by how much irrelevant data existed, and the overall amount of Twitter activity doesn't seem to match with real-world events. We do not suspect the data to be biased, namely because the manner in which it seems to have been collected does not favor one side over the other. Still, the analysis is about as conclusive as we expected; natural language processing is messy and is bound to miscategorize some things (sarcasm especially), and the way people write on Twitter is different from the training data. Some line-blurring was bound to occur, but Dragas clearly received more vitriol, and people seemed to like the reinstatement of Sullivan.