**Group Data Analysis Project**

**CS 4319 Data Mining and Warehouses**

**Dr. Oner Celepcikay**

James Castillo, Liejo Fernandus,

Muazam Jafry, Imran Hussain

**Abstract**

The team was assigned a project to access Twitter Application Interface (API) using python to implement a program that would help stream and count the number of supporters each Presidential Candidate (Donald Trump and Hillary Clinton) have, based on the hashtags “#HillaryClinton” or “#realDonaldTrump”. This project was split into two phases; the first phase involved the collection of voters data through the use of twitter API. The contents of this phase have already been submitted to Dr. Celepcikay. This report talks more in depth about Phase 2 of the project where the team implemented its own twitter API to create own polls using tweet analysis.

**Introduction**

At the end of phase 1, the team met to decide on what course to take to complete the second phase of the project. Among the multiple methods suggested, James came up with a design to code in python to analyse tweets and try to predict if a specific tweet was a vote for Hillary Clinton or Donald Trump. The team went about this project by first doing research on twitter analysis API’s and took some create courses on Python, some of which was provided by Dr. Celepcikay.

**Design**

The team used Python 2.7, Matplotlib and Pandas plugins to create the python program which in turn helps plot how many tweets that insinuate votes are for Clinton or Trump. The program was setup to compare tweets between each candidate directly as well as to record tweets depending on location or origin of the tweet. The team was not able to design the program where it would show Tweet analysis info from each state separately but was able to grab data from a collective group of states and use this to make the vote prediction.

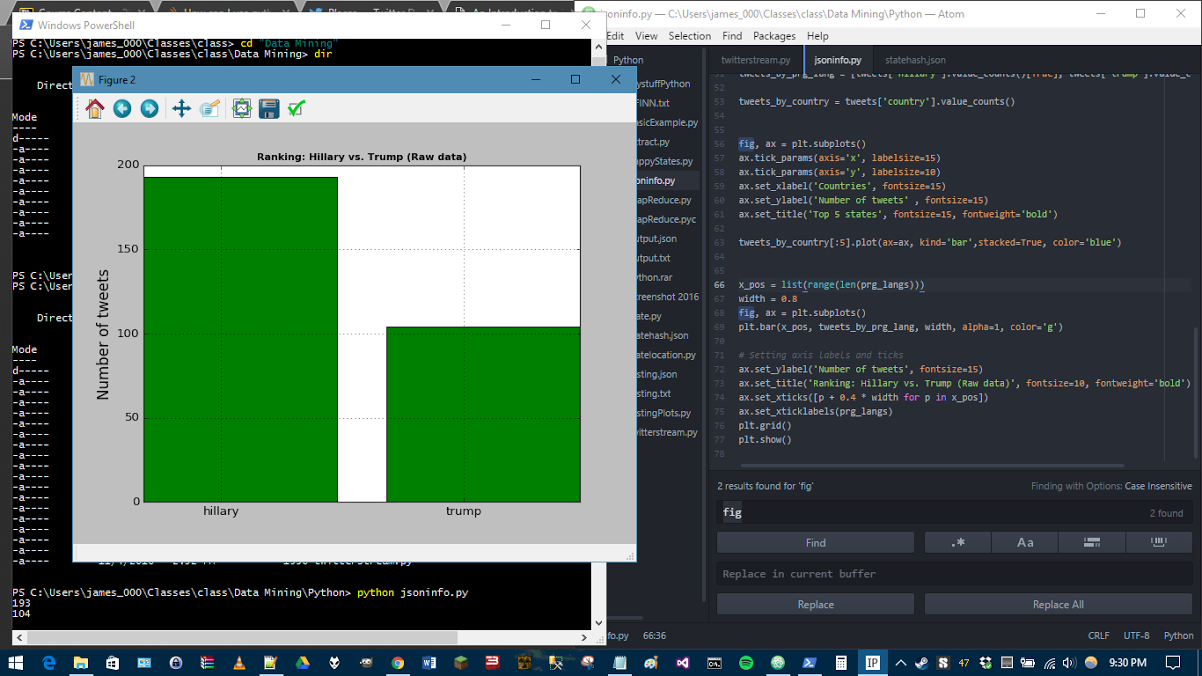


Figure 1. Screenshot of tweet analysis of votes for Presidential Candidates

The direct comparison of tweet analysis which predicts votes for the Presidential candidates through the use of the python program the team created is shown in Figure 1. The results obtained from running the python program the team created is very much similar to the data that was collected for Phase 1 of this project and turned in to Dr. Celepcikay.

We can view in figure 2. how the top 5 states were predicted to vote based on the tweet analysis that the team ran using the python program we created. Hillary was clearly in the lead in both forms of analysis.

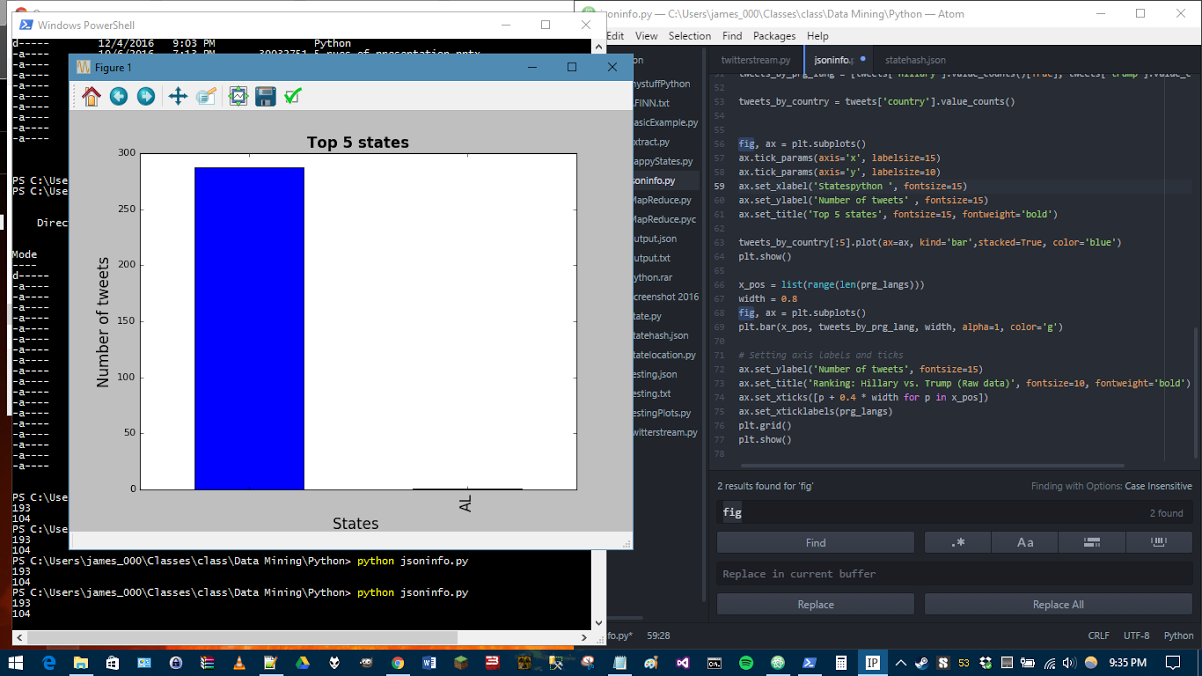


Figure 2. Screenshot of tweet analysis of votes for Presidential candidates based on location

**Conclusion**

Through the previously collected data which was submitted in phase 1 and the newly attained data through the python program created, the team was able to conclude, through tweet analysis, that Hillary Clinton would receive a higher vote count and be named as the next and first female President of the United States. As we found out over time, this was not the case and the data collected was not able to give a completely accurate depiction of the final results of the 2016 US Presidential Election. There are many factors that can be a cause of this, one of which can be that most voters that voted for Donald Trump do not use Twitter and hence the data for these voters was not collected with the program that was created by the team.

Although the program did not give the right prediction to the election, the team had a great experience designing and working on this project which gave us all an opportunity to learn more about Python and Twitter API, as well as expose us to tweet analysis.

**Works Cited**

* "How to Use APIs with Python." *Codecademy*. N.p., n.d. Web. 20 Nov. 2016.
* Http://fivethirtyeight.com/contributors/nate-silver/. "2016 Election Forecast | FiveThirtyEight." *FiveThirtyEight*. N.p., 08 Nov. 2016. Web. 22 Nov. 2016.