From Canvas:

For the final project, you are expected to:

Pick an interesting Social Web problem; For example, you may want to help people affected by a given socially-related issue; or you may want to help a given category of workers; or, again, you may want to understand how many users discuss a given topic, how, why, and what does this imply;

An EXTENDED analysis and report of Social Web data

Pitch your idea and present the final result. Describe your (individual) contribution to the assignment (individual chapter of the report). Collect the results of your work in an ANALYSIS REPORT (to be included in the report):

Step 1: Formulate a research question to answer with your data Step 2: Describe the data used and the way it was collected Step 3: Describe what pre-processing you performed on the data, if any; explain why it was necessary Step 4: Visualise the non-preprocessed data, if any, to show the importance of data cleaning Step 5: Carry out the necessary analyses and visualise your results (e.g., graphs, barplots etc.). Show how you answer your research question with these analyses Step 6: Describe the limitations of your data and analysis or app Step 7: Share your code via GitHub or alternative services. Preferably you should construct interactive visualisations

IS ABSTRACT REQUIRED?

General Idea

The US elections in 2020 is a trending topic on Social Network Services (SNS) such as Twitter and Facebook. SNS have proven themselves to be a great influence on society sparking controversy in for example the 2016 US elections. This gives rise to the following question: To what degree did significant events related to the US elections of 2016 have an influence on the actual prediction outcomes as presented by data available on social network services? The idea is to mine data from Twitter and Facebook related to the USA elections in 2016 surrounding the republican candidate "Donald Trump" and democratic candidate "Hillary Clinton". The data collected can be compared to data of the real US elections outcomes of 2016. The gathered information and analysis of the data can present a great insight into actual SNS influence on world politics.

Data Collection

Twitter data collection

After analysis on twitter's structure, we found 5 following useful twitter attributes to collect (For privacy reasons we didn't collect user's id):

- We collected each twitter's id to identify which specific twitter it is, so that we can traceback when we have data analysis problem on specific twitter.
- We collected each twitter's permalink, it is the link to specific twitter.

- We collected each twitter's text and we analysis them to do prediction of 2016 US election.
- We collected each twitter's date so that we can see how are people's attitudes change towards 2016 US election candidates through 100 days.
- We collected each twitter's hashtags, it is kind of human defined category, we can use it to find out core topics during specific periods.

Furthermore, we found that twitter system has two different kinds of twitter, one is top twitter, which are selected the most focused twitters base on twitter's algorithm, the other is normal twitters.

In data gather stage, we collected ALL Trumps' and Hillary's related top twitters, at the same time we random collect 10K Trumps' and Hillary's normal twitters, both 2 kinds of data contains 5 attributes mentioned above and their data date is around 100 days from August 01, 2016 to November 09, 2016.

In code implementation part, we didn't use official twitter python package and its corresponding API. This is because of official API does not allow free developer account to search twitter before a week, and the API access time limitation is 6000-times per week, which means we can not get enough amount of data for this topic. After research we used a python package called GetOldTweets3, it is a kind of python twitter crawler. This kind of package should not be used under commercial condition due to twitter's policy, and only can be used in education condition.

2016 US president important dates collection

From the paper [Mentioned on CLASS] we know that political issues discussed on social web is event driven. So we collected 2016 US president important events and dates as follows, they are collected and selected manually.

Description	**Event date**
US Presidential Election 2016	November 8, 2016
First presidential debate; Hempstead, New York	26 September
Donald Trump Access Hollywood tape	October 7, 2016
U.S. Says Russia Directed Hacks to Influence Elections The Guardian: US officially accuses Russia of hacking DNC and interfering with election	October 7, 2016
The Guardian: Newly discovered emails relating to Hillary Clinton case under review by FBI The Guardian: How we got here: a complete timeline of 2016's historic US election	October 28, 2016

Data pre-process

Data pre-process is happened in data collecting period, it implementation file is called `data-clean.py`.
In this part we transfer date time to standard date-only format, we lower hash tags text and we clean tweet text by removing links, special characters, which is the most important operation in this part. We transferred date time to date only because 2016 US president prediction output result is generated in daily way. Lower hash tags is very useful for future hash tag statistic to avoid case related problem, for example “#TRUMP”, “#Trump”, “#trump” should have the same meaning and regard as the same hashtag. Besides, tweet text cleaning is target at removing no meaning symbols and links, so that we can get precise sentiment analysis and token/NER analysis in the next stage.

Visualize the non-preprocessed data

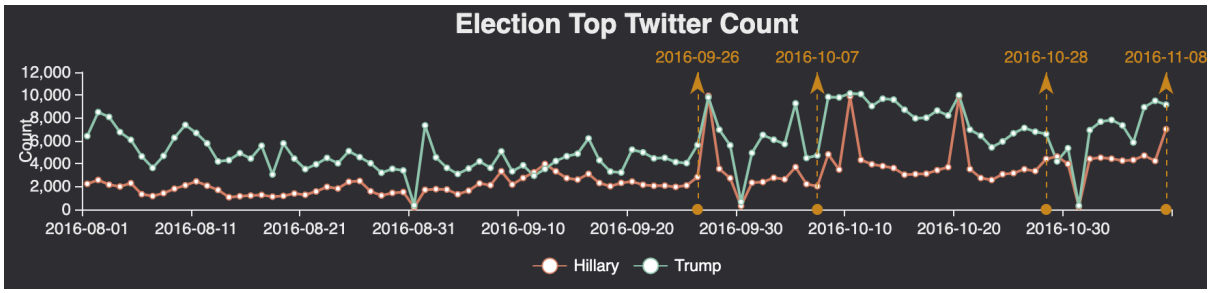
Raw text data

id	text
787080512502325249	The Mic, the Teleprompter. His greatest enemy is technology. Maybe Trump is a hero sent from the future to fight the Rise of the Machines?
787080462703263744	Trump insults one accuser with "she wouldn't be my first choice." Guess who's first choice you're not, Donald? The American voters.
787080429568454656	.@realDonaldTrump @HillaryClinton which one of you gonna bring the aux cord back to iPhone?
787080411927175172	why did Obama have to ether Trump like this
787080394130595840	#PodestaEmails7 @megynkelly #kellyfile dropping like a rock @FoxNews. BOYCOTTS WORK @realDonaldTrump supporters.
787080367756967936	yes it's always someone's else's fault it's called king baby syndrome .. he needs a lot of help..DUMP trump now
787080347167186944	I donated today to get my Official Trump Card. Did you? Stop Tweeting and Donate NOW to support .@realDonaldTrump !!!!
787080331954311168	The greeting.....for me?? (Probably not) @realDonaldTrump rally Charlotte, NC #Campaign2016 #NorthCarolina
787080298722762756	@Fahrenthold @timfunk @realDonaldTrump Religion is big business. Follow the money.
787080235514785792	"We are going to make our country so much STRONGER and so much SAFER." - @realDonaldTrump
787080114353926144	.@realdonaldtrump in CLT: African-American community "living in the hell" of these inner cities. Says, "What the hell do you have to lose?"
787080110360760320	We LOVE you, Mike Pence! #PENCE2020
787080109698134016	I am with @realDonaldTrump ALL THE WAY TO THE WHITE HOUSE!! #MAGA
787080104270766080	.@realDonaldTrump TelePrompTer has failed. Hang on to your hats
787080077653704704	Well, at least it begins to explain Franklin Graham's exceptional devotion to Trump...
787080070301188097	"Murder rates are at the highest rate they've been in 45 years. We are going to make our country so much safer." - @realDonaldTrump
787080046435528704	Donald Trump: I "wasn't impressed" when Hillary Clinton walked in front of me at debate http://cnn.it/2egOBkA
787080023861784577	One America is possible. Not a black or white America, but one America. @realDonaldTrump #TrumpPence16
787079884652711937	Obama: Let's get everyone insured. "YOU'RE DESTROYING AMERICA!!!" Trump: I hate minorities & force finger women. "Hmm let's hear him out."
787079829770436608	@realDonaldTrump We all know these women are lying so do not talk about them. Do not give them their 5 minutes of fame. Stay on policy! <3 U

As we can see from this picture there are a lot of texts, from non-processed data we can not show prediction directly sentiments analysis is required. Moreover, large data set means it not available for us to do sentiment analysis one by one manually. So we need use program to do sentiment analysis. Besides, what are voter concentrated on during different period? It need NLP to extract information from twitter text then do statistics.

Top twitter numbers and important dates

We got all top twitters for two candidates, so we can got sum number of two different candidates, the sum numbers of top twitters can reflect popularity of different candidates. We also have important dates so we combined two data into one chart to show if big issues can influence candidates popularity



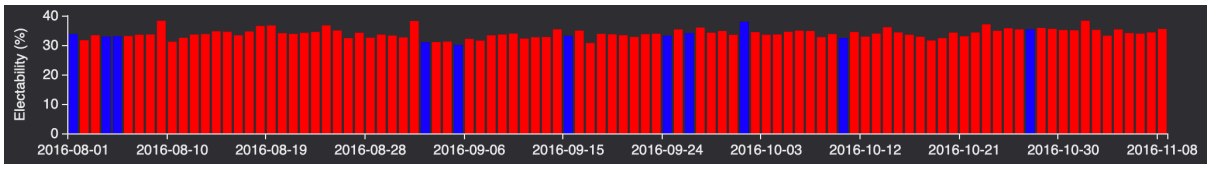
Analyses and Visualization

Analyses part and visualization part contains 5 main parts:

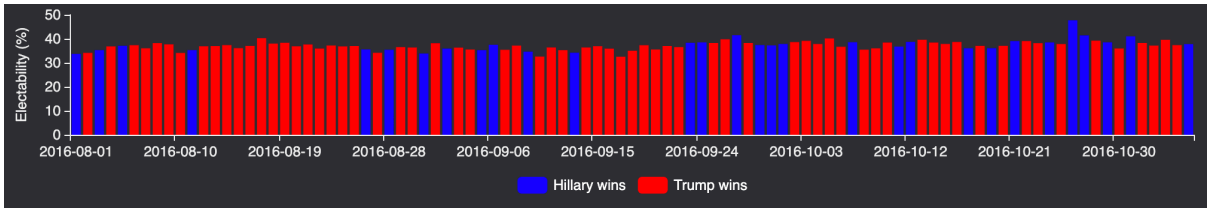
- Twitter sentiment analysis and US president prediction
- Twitter NER process (Named-entity recognition) and Twitter token statistics.
- Twitter hashtag statistic
- Top twitter's & Normal twitter's result similarity check
- Findings

Twitter sentiment analysis and US president prediction

The 2016 US president prediction is based on the sentiment analysis on everyday's twitter texts. The citizens attitudes towards two different candidates have three basic status: "positive", "neutral", "negative", and the ratios of three attitudes towards two candidates are changed from day to days.First, we get people everyday's twitter sentiment analysis result. Then compare Hillary's Positive ratio with Trump's Positive ratio and take higher positive ratio candidate as daily winner. Finally we put it in a bar chart to see who got more support through 100 days. For this part's implementation we used python textblob module to do sentiment analysis and used multi-threads spark to boost the analysis speed.



Normal twitter sentiment analysis



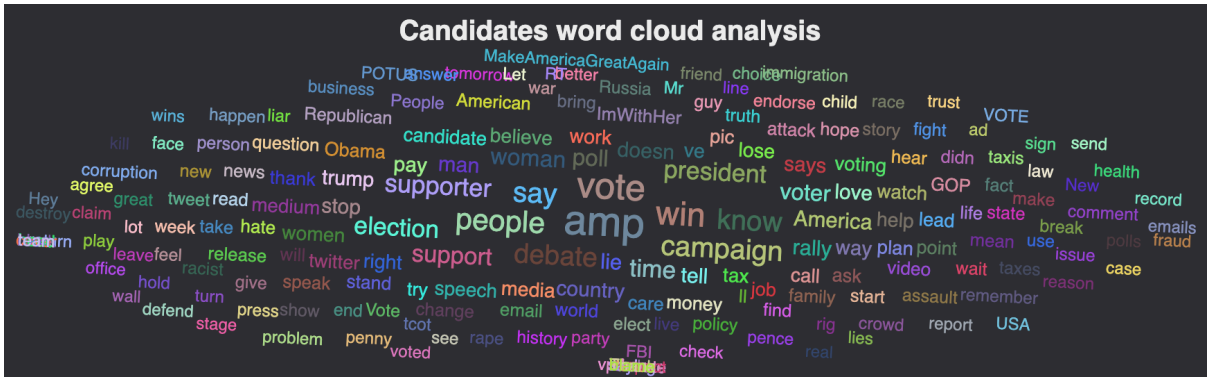
Top twitter sentiment analysis

Twitter NER process and Twitter token statistics

Twitter NER (Named-entity recognition) process targets at extracting important named-entity related to president like "Barack Obama", "Russia", "Republican", "immigration policy" and so on from twitter texts to see what the highly concentrated entities during president election period.

Twitter token statistic is target at getting people's operations and attitude words and most mentioned entities. We split twitter sentences into multiple single words as tokens. We filtered out number words, symbols and conjunction, subordinating or preposition words because they have no useful meaning in twitter text. Furthermore, we converted nones or verbs into dictionary based words, for instance we converted "went","gone","goes","go" to word "go", so that all tokens have a standard form in statistic.

In code implementation we used python spacy to do NER recognition and token procession at same time we used python spark to boost this part's data process. Finally, we used e-chart JS's word cloud library to do visualization.



Twitter hashtag statistic

Hashtag is also very important, because they are human created which tell us directly what twitter users are focusing on in their tweets. We need to split tweets hashtags and do statistic, and finally use bar chart to show what are the most mentioned tags.

Top twitter's & Normal twitter 's result similarity check

Do top twitters are controlled by several org or people that cannot reflect real president selection? We will use python scipy stat module to do paired t-test on each candidate's normal & top twitter sentiment analysis result and overall election prediction output based on normal & top twitter to reflect if top twitter and normal random twitter got the same result. From over all prediction, we got :

```
Prediction t = 3.7199509224307223
Prediction p = 0.0002594805556168014 < 0.05
```

From each candidate's analysis sentiment analysis, we got:

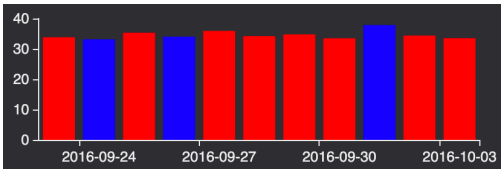
```
Hillary sentiment analysis t = -11.778465339325374
Hillary sentiment analysis p = 1.017535140109869e-24 < 0.05
Trump sentiment analysis t = -12.510471871318002
Trump sentiment analysis p = 5.775976722535195e-27 < 0.05
```

As we see no matter under what condition p-value is lower than 0.05 which means top twitter's and normal twitter's output results are different!

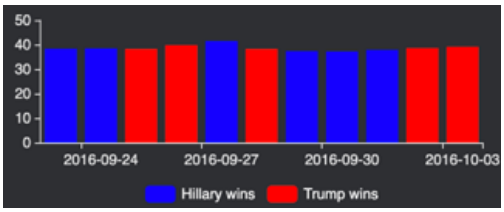
Findings

From "Twitter sentiment analysis and US president prediction" and "Top twitter's & Normal twitter 's result similarity check" we can see that from Top twitter prediction and Normal Twitter's overall prediction "Trump" is more likely to win the president election, but from daily prediction their outcome are not the same! This outcome means there are some orgs are trying to influence public opinions. Let's have a look what happened near important dates:

First Debate

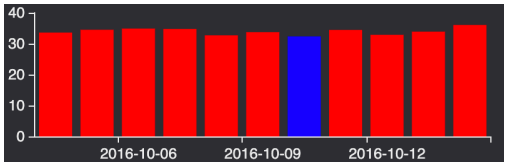


Normal Twitters

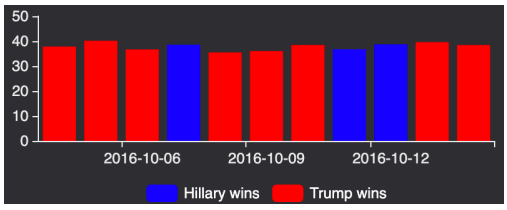


Top Twitters

Russian Hack

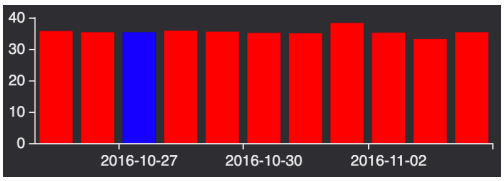


Normal Twitters

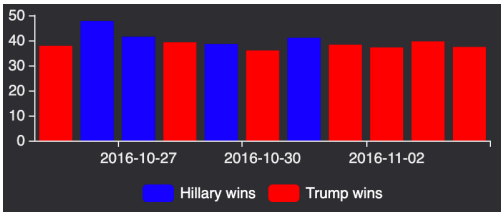


Top Twitters

Hillary Mail Issue

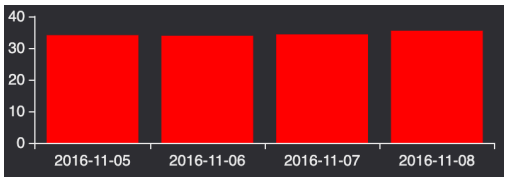


Normal Twitters

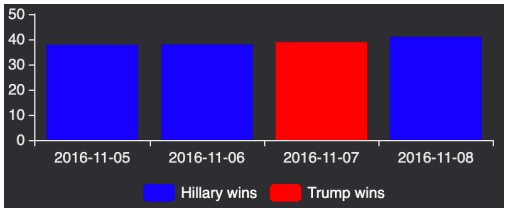


Top Twitters

Election Day



Normal Twitters



Top Twitters

From 4 groups of charts we can see that no matter under what condition, Hillary top twitter's support is higher than Hillary normal twitter's support. Furthermore, except for Russian Hack event, Hillary got higher support than Trump under top twitter condition while normal twitter condition vice versa. These means Hillary related group was trying to influence public opinion however their media has a very weak influence on public opinion.

From the word cloud and hashtags, we found that Russia, Muslin, Immigration, Tax, Women Right Related issues are citizen's mainly concentrated issues during 2016 US president election. Moreover, the citizens have strong patriotic feelings, so they really want to make American again!

Limitations of your data and analysis

For sarcastic analysis is not easy to implement in NLP, we did not check sarcastic in this project. In this circumstance the result is not that precise, if a day two candidate have similar positive ratio this may influence the prediction of that day.
As we mentioned that political issues in social web are event driven, but how does an event will influence on different candidates we cannot figure it out.