# CS143, Spring 2019
# Project2CS143 Report

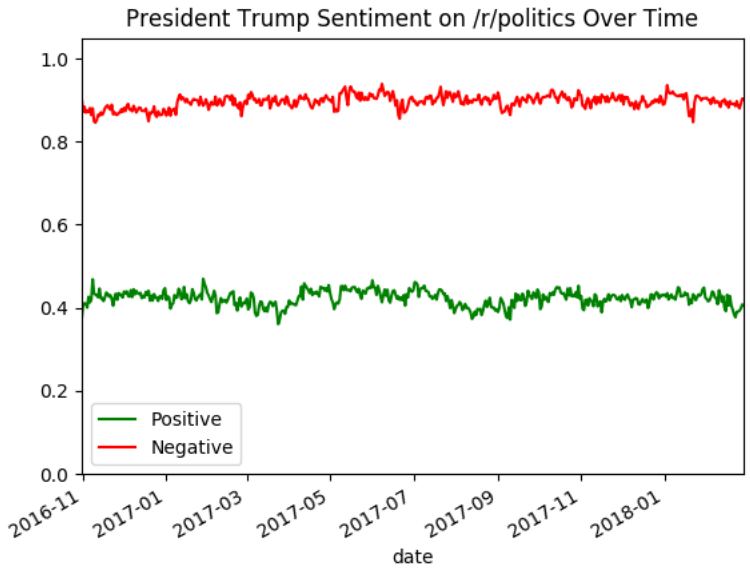
Justin Chang, Quentin Truong

06/1/2019

## 1 Graphs and Explanations

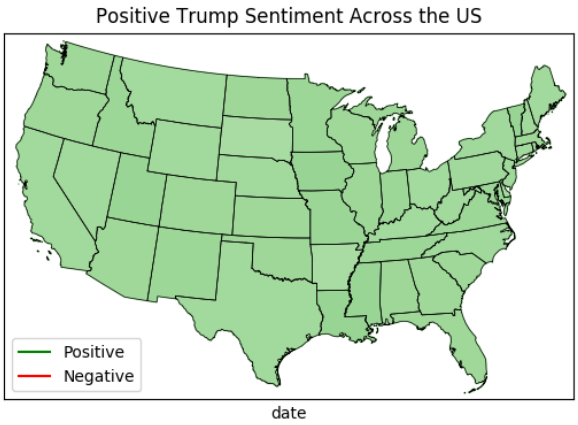
(a) Graph 1. Create a time series plot (by day) of positive and negative sentiment. This plot should contain two lines, one for positive and one for negative. It must have data as an X axis and the percentage of comments classified as each sentiment on the Y axis.

**Solution:** Below is the graph requested




(b) Graph 2 and 3. Create 2 maps of the United States: one for positive sentiment and one for negative sentiment. Color the states by the percentage.

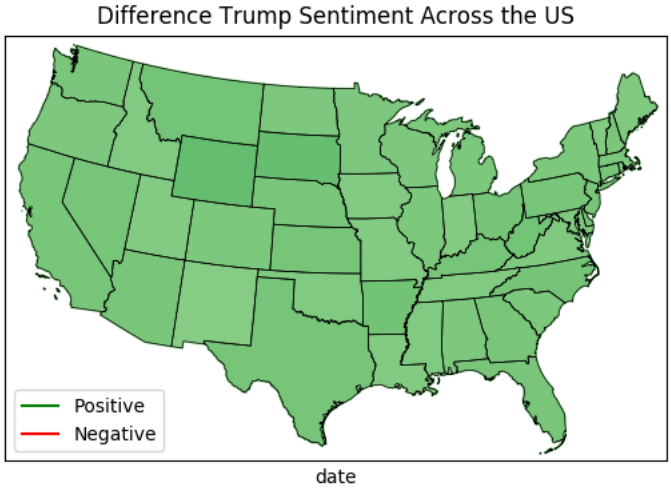
**Solution:** Below are the 2 graphs requested





(c) Graph 4. Create a third map of the United States that computes the difference: Positive - Negative percentage.

**Solution:** Below is the graph requested




(d) Graph 5. Give a list of the top 10 positive stories (have the highest percentage of positive comments) and the top 10 negative stories (have the highest percentage of negative comments). This is easier to do in Spark.

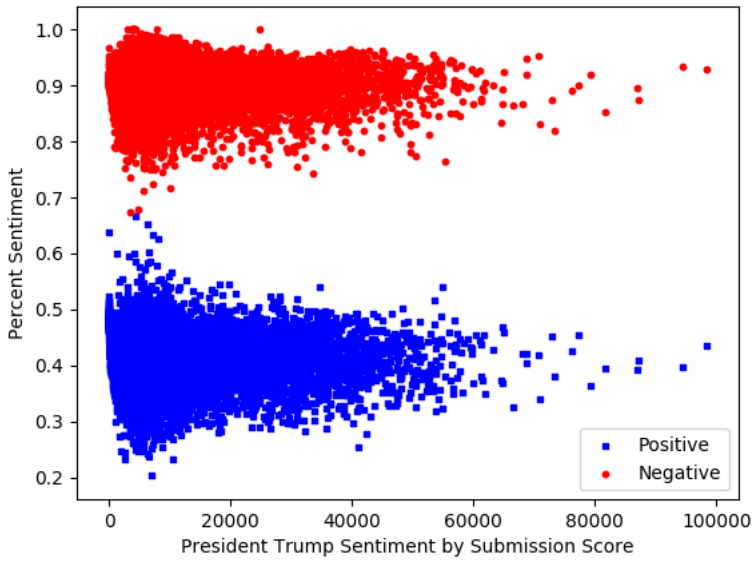
**Solution:** Below are the 2 csv lists requested. These were generated in spark. The first csv list is top positive and second csv list is top negative

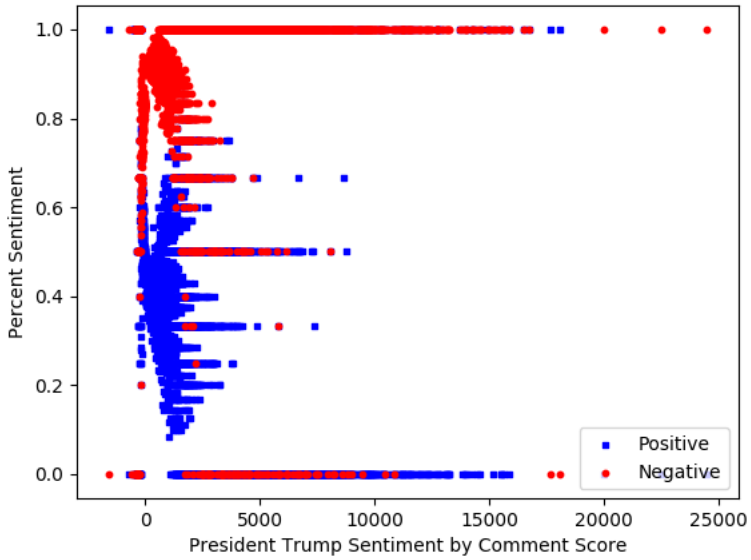
```
link_id,title,Positive,Negative
5bg1au,U.S. Supreme Court Allows Arizona To Enforce Ban On Ballot-Collecting,1.0,1.0
5bhydj,F.B.I. Says It Hasn't Changed Its Conclusions on Hillary Clinton Email Case,1.0,1.0
5an1ff,Republican activists to monitor Election Day polls,1.0,1.0
5bio8r,Top Democrats say Clinton took a real hit from Comey. But they’re cautiously optimistic.,1.0,1.0
5aj4wf,Report: Trump used dubious tax avoidance scheme in 1990s,1.0,1.0
5bld35,Clinton versus the Turnip,1.0,0.5
5b1263,Historic Mississippi black church burned and vandalized with ‘Vote Trump’ graffiti,1.0,1.0
5blnih,"In Colorado, gap between Democratic and Republican strategies is clear",1.0,1.0
5bnzs5,"Janet Reno, First Woman to Serve as U.S. Attorney General, Dies at 78",1.0,1.0
5bnnum,Clinton has the edge one day before election,1.0,0.0
```

```
link_id,title,Positive,Negative
5appvm,Meet the unopposed Assembly candidate who says climate change is a good thing that hurts 'enemies on the equator',0.0,1.0
5ao8j6,Hillary Canceled Her Last Public Event Because The Crowd Yelled “Lock Her Up”,0.5,1.0
5ahvea,"Investigating Donald Trump, F.B.I. Sees No Clear Link to Russia",1.0,1.0
5ahyqk,Just told my mom im not voting for Hillary Clinton and this happened,1.0,1.0
5aq5q6,Bill Weld on Rachel Maddow: ‘I’m Here Vouching for Mrs. Clinton’,0.3333333333333333,1.0
5al4fo,Election maps are telling you big lies about small things,1.0,1.0
5amict,Hillary Clinton's challenge: Shift focus back to Trump,0.0,1.0
5an1ff,Republican activists to monitor Election Day polls,1.0,1.0
5ao7m9,U.S. House Speaker Ryan renews call to suspend classified briefings for Clinton,0.0,1.0
5au5ax,James Comey's casting of innuendo reminiscent of J. Edgar Hoover,0.0,1.0
```

(e) Graph 6. Create TWO scatterplots where the X axis is the submission score, and a second where the X axis is the comment score, and the Y access is the percentage positive and negative. Use two different colors for positive and negative. This allows us to determine if submission score, or comment score can be used as a feature.

**Solution:** Below are the 2 scatter plots requested






**Paragraph Explanation of Observations:**

All of the plots produced above involved using all of the dataset after applying a 0.25 threshold in our model creation in task 7. It is worth noting that the sentiments seem to be more negative than positive when looking at the data in each plot that assumes supporting trump vs against trump. This makes sense since there are usually more liberals and democrats on our reddit politics thread, so sentiment analysis from dataset indicated that there were more support in the politics subreddit that are against trump. We can also see on the maps that there are way more people that are negative against trump than positive across the US in general, and the data from our other graphs indicate that the sentiments remain consistently in that favor. Another thing to take away from the difference graph was that since most of our data indicated that we had more negative trump sentiments, the difference ended up being an absolute difference (defined the difference as essentially, if the value was going to be negative from positive-negative, switch the sign so we only get positive differences). It does not seem to vary too much by state as in most of the sentiments seem to just align towards negative compared to positive. Additionally, over time, there just seems to be around the same (more negative) than positive sentiments. Looking at story/submission, we also see that in the 2 scatter plots that there are just more negative than positive sentiments (both comment and submission scores strongly indicate that) although there may be a few comments that also show support, most are actually indicating negative rather than positive).

## 2 Questions and Answers

(a) Part 1

**Solution:** Solution to Part 1

Functional Dependencies in labeled_data.csv:
Input_id − >labeldem
Input_id − >labelgop
Input_id − >labeldjt

(b) Part 2

**Solution:** Solution to Part 2

No the data is not fully normalized. We could decompose it further by moving all attributes relating to the author into another relation and using an author_id. Also, we could decompose it further by moving all attributes relating to the subreddit into another relation and only using the subreddit_id. I think the data collector stored the data like this to make it easier to perform statistics on the data (A OLAP point of view, this is better to do aggregation functions). It would be easier because we don't need to perform joins. Also, it may be that the data was made available in this format and so it was just collected as is. The data is also thus, going to take up a lot more space.

(c) Part 3

**Solution:** Solution to Part 3

Question 3:

Using explain on this line of code in reddit_model.py:

labeled_comments = labels.join(comments, comments.id == labels.Input_id).select(['Input_id', 'labeldem', 'labelgop', 'labeldjt', 'body']).explain()

Gave the output:

```
Gave the output:
== Physical Plan ==
*(2) Project [Input_id#170, labeldem#171, labelgop#172, labeldjt#173, body#4]
+- *(2) BroadcastHashJoin [Input_id#170], [id#14], Inner, BuildLeft
   :- BroadcastExchange HashedRelationBroadcastMode(List(input[0, string, false]))
   :  +- *(1) Filter isnotnull(Input_id#170)
   :     +- *(1) Sample 0.0, 0.1, false, -4372746341945787401
   :        +- *(1) FileScan parquet
   [Input_id#170,labeldem#171,labelgop#172,labeldjt#173] Batched: true, Format:
   Parquet, Location: InMemoryFileIndex[file:/media/sf_vm-shared/labels.parquet],
   PartitionFilters: [], PushedFilters: [], Rea dSchema:
   struct<Input_id:string,labeldem:int,labelgop:int,labeldjt:int>
   +- *(2) Filter isnotnull(id#14)
      +- *(2) Sample 0.0, 0.1, false, 7407911338482665907
         +- *(2) FileScan parquet [body#4,id#14] Batched: true, Format: Parquet,
         Location: InMemoryFileIndex[file:/media/sf_vm-shared/comments.parquet],
         PartitionFilters: [], PushedFilters: [], ReadSchema:
         struct<body:string,id:string>
-
```

The join algorithms used by spark seem to be a broadcast hash join. Hash joins are generally useful for equality joins on large tables because it computes the join condition based on a hash index. It seems like it is also joining on the hash keys denoted by the number next to the attribute (input_id is 170 and id is 14) and does an inner left join. It is also worth noting that we only used 10% of the data to sampling just to speed things up a bit. The project on the first line after the physical plan shows SQL select being used or in relational algebra terms, the project statement to select the following 5 columns.

Note:
**USAGE FOR PROJECT**:
spark-submit reddit_model.py
Also, each plot probably took around 6.5 hours to save the data for when using the ENTIRE dataset. We could produce similar looking (trendwise) graphs in a matter of minutes as long as we sample at a much smaller data rate (and also get enough data, example is 10 percent of data was already pretty reasonable). Future considerations would be to try to make it so that we don't need to wait that long just to generate data for each plot.
It is also worth noting that our plotting script looks for the dataset.csv by also appending a 1 to the end of each dataset name (before .csv) just to make it easier to manually rename and alias it differently from the .save in redditmodel.py (makes manually renaming the data.csv files easier to find and use for plots). Example would be: time_data.csv's cs file inside the folder, is renamed to time_data1.csv