Project 2B Report

Yutong Han 705025619
Yufeng Huang 704944399
Yufei Hu 404944367
Tianyi Liu 705035425

June 10, 2018

1 Time Series Plot

The Figure.1 shows the time series plot(by day) of positive and negative sentiment.

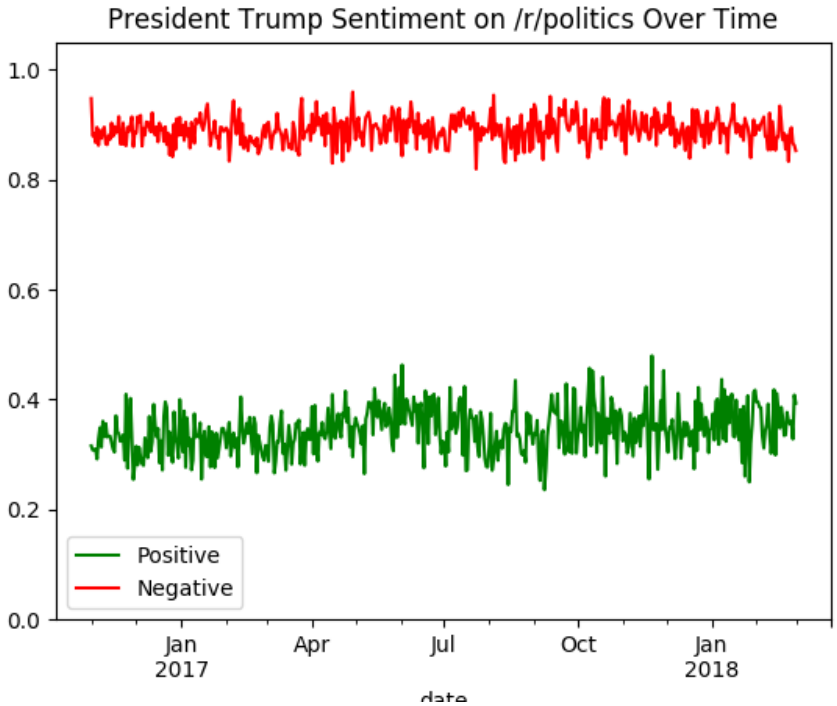


Figure 1: Time Series Plot

2 Positive and Negative Sentiment of US State

The Figure.2 and Figure.3 shows maps of the states of the United States for positive sentiment and one for negative sentiment.

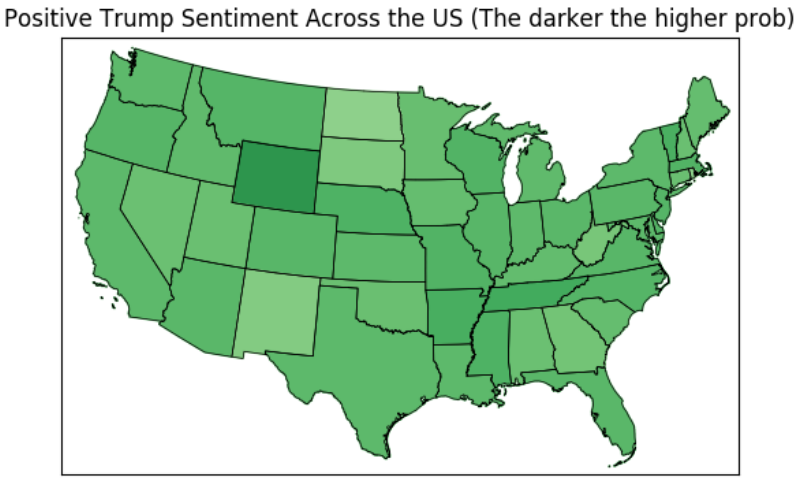


Figure 2: Positive Sentiment over US States

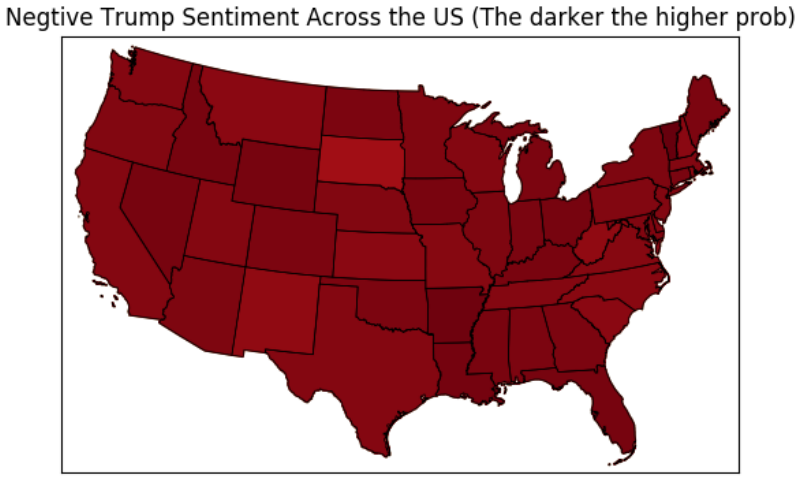


Figure 3: Negative Sentiment over US States

3 Sentiment Difference of US State

The Figure.4 shows the different of positive and negative sentiment across the US States.

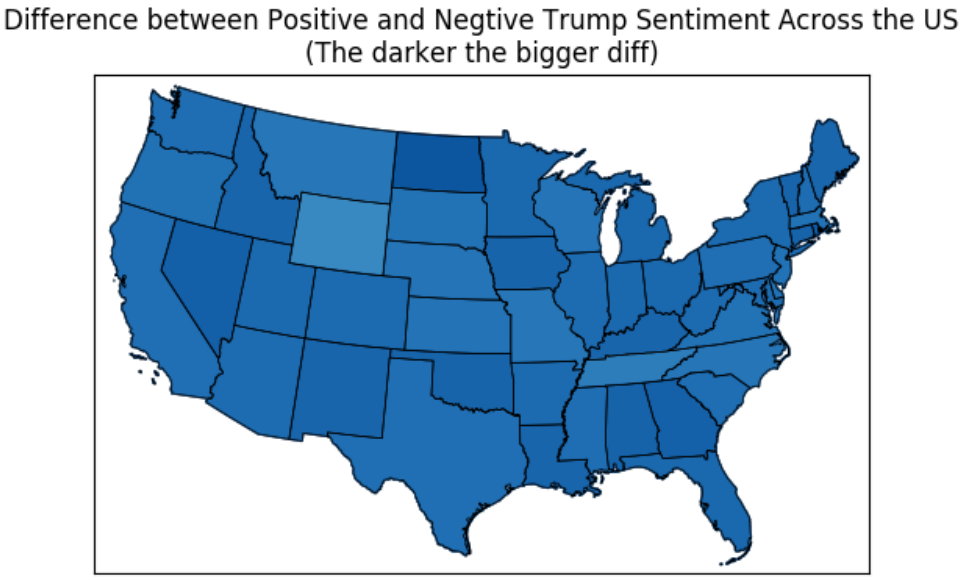


Figure 4: Difference between the positive and negative sentiment across US States.

4 Top 10 positive stories

The following two tables, table 1 and table 2 shows the list of top 10 positive stories and top 10 negative stories.

Table 1: Top 10 positive stories

ID	Title
dnzno4z	Diplomat regrets new turn with U.S
dfs0vzl	Federal judge blocks Indiana abortion ultrasound mandate
dop6avh	President Trump Promotes Book by Wonderful Pastor Who Says Satan Founded the Catholic Church
dqsb8v2	Prosecutors say longtime Manafort colleague has ties to Russian intelligence, the first such allegation by the special counsel
dhsmsbd	Anderson Cooper to Trump surrogate: 'If he took a dump on his desk, you'd defend it'
dht0d5u	WH Official: NYT, WaPo Reports Are 'Coordinated Attack' on Trump
dryjzw2	Russian tankers reportedly smuggling oil to North Korea: Trump has been silent, though one day earlier, he blasted China for the same thing.
dkxzvnm	Ben Shapiro: 'Views Should Never be Banned'
dlbi7zb	Feinstein: We Can't Increase Immigration Enforcement Because No One Will Pick Our Fruit
dmtjwsg	Mr. President, They're Never Going To Like You

Table 2: Top 10 negative stories

ID	Title
dtl0kfs	FBI director prepared to issue rebuttal if Nunes memo released
den7nsu	Adding a Dislike Button to Twitter Could Neutralize Trump's Social Media Presence
di419p4	'This is off the map': Former intelligence officials say the reported Kushner-Russia plan is unlike anything they've ever seen
dhnhxbt	Graham invites Comey to testify before Senate panel
diw3ehp	Rep. DeSantis: Man Asked Whether Republicans or Dems Were on Field Before Scalise Shooting
dozvoc1	Former President Obama called for jury duty in Chicago
dg76w4g	Trump Says He May Freeze Subsidies to the Poor Until Democrats Repeal Obamacare
dhc6yws	Republican Senator Graham to examine Trump's business deals: CNN
di8wivz	GOP taps anti-Clinton strategy to damage Elizabeth Warren early
dje8p0t	Russian Agents Used High Tech Alien Technology To Collude With Trump While Hiding Under James Comey's Floorboards, Sources Say

5 Scatterplots of Submission Score and Comment Score

The Figure.5 and Figure.6 are the 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. The two different colors for positive and negative.

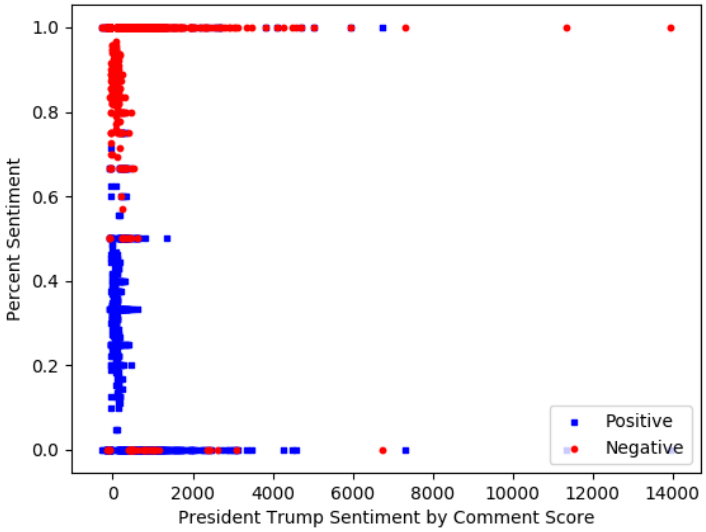


Figure 5: Sentiment by comment score

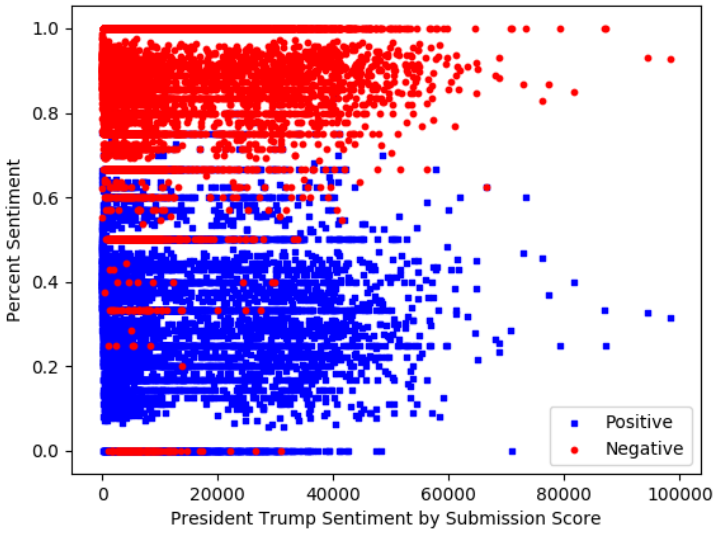


Figure 6: Sentiment by submission score

6 Producing the ROC curves

The Figure.7 plot the ROC for the classifier.

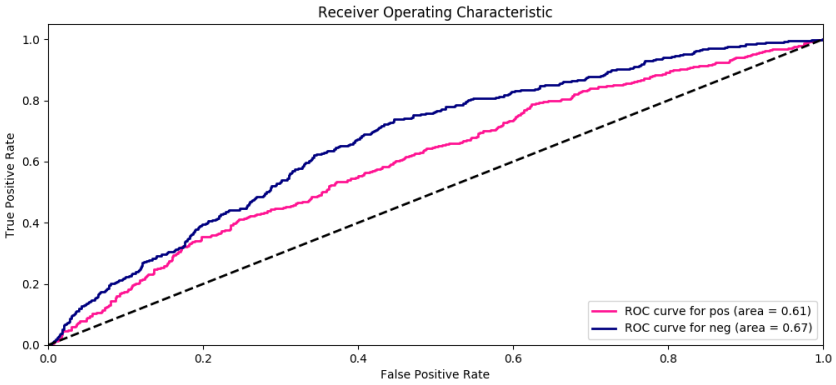


Figure 7: Receiver Operating Characteristic

7 Summarizing finding

Through the change of the time curve, we can clearly see that Trump's positive comment has steadily and slightly increased with the increase of Trump's ruling time. However, on the whole, the fact that the majority of the comments about President Trump is negative did not change at all.

What's more, surprisingly, the distribution of positive and negative sentiment towards Trump across the US States is not based on the states that we envisioned to support Republicans or Democrats. In other words, there is no significant difference between the states.

What we do expect is that the comment's sentiment differs in different stories. People inclined to express their negative comment for the Trump Russia scandal, the performance of Trump in social media and some crazy and some his irresponsible speech. But when it comes to immigration and religion, people have more ratio of positive sentiment.

As far as I am concerned, all the mentioned phenomenon can be explained by an extremely important and essential question that who are the voters of Trump, which is also the reason why Trump won the election the president of America. According to plenty of sociology and political literature, the claim that white working-class voters were a crucial block of support for Trump in the 2016 presidential election has been the consensus in the academic. It is the silent majority that makes Trump become president and at the same time those people's opinion and sentiment toward Trump is very difficult to detect by political blog statistics or public opinion poll.

8 Question 1

Take a look at *labeled_data.csv* Write the functional dependencies implied by the data.

Answer: The *Input_id* is the primary key of the scheme. So the $Input\_id \rightarrow labeldem$, $Input\_id \rightarrow labelgop$, $Input\_id \rightarrow labeldjt$ and their combinations can be implied.

9 Question 2

Take a look at the schema for the comments dataframe. Forget BCNF and 3NF. Does the data frame look normalized? In other words, is the data frame free of redundancies that might affect insert/update integrity? If not, how would we decompose it? Why do you believe the collector of the data stored it in this way?

Answer: It does not look normalized. In the *subreddit_id* and *subreddit* part, it seems that the $subreddit\_id \rightarrow subreddit$. So the *subreddit* may store repeatedly. So it can be decomposed into two scheme with the *subreddit* be the key of other scheme. Since this database may contains about the posts about the politics, the subreddit is just the */r/politics*

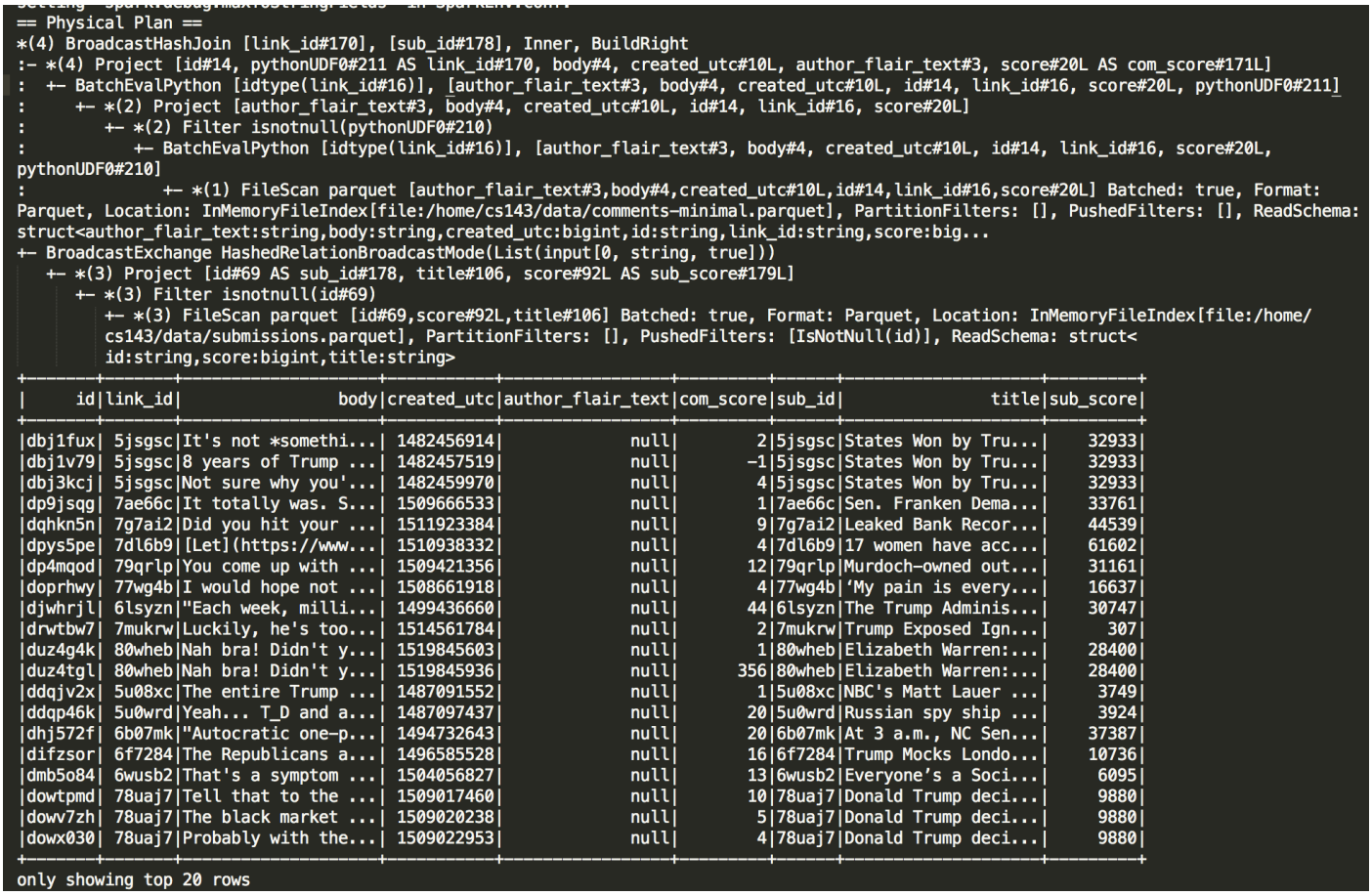
10 Question 3

Pick one of the joins that you executed for this project. Return the join with .explain() attached to it. Include the output. What do you notice? Explain what Spark SQL is doing during the join. Which join algorithm does Spark seem to be using?

Answer: The Figure.8 shows the output of the .explain on the join operation

```
df_full = df_com_full
  .join(df_sub_full,
  df_com_full.link_id == df_sub_full.sub_id,
  'inner')
```

Here we can notice that the Spark use the Hash Join algorithm (BroadcastHashJoin) and build the index on the right relation. Spark first loads each relation and filter the not null key value we want to join on. Then project the values we want to select. And finally do the hash join on two relations using the Hash Join alogorithm.



```
== Physical Plan ==
*(4) BroadcastHashJoin [link_id#170], [sub_id#178], Inner, BuildRight
:- *(4) Project [id#14, pythonUDF0#211 AS link_id#170, body#4, created_utc#10L, author_flair_text#3, score#20L AS com_score#171L]
:  +- BatchEvalPython [idtype(link_id#16)], [author_flair_text#3, body#4, created_utc#10L, id#14, link_id#16, score#20L, pythonUDF0#211]
:     +- *(2) Project [author_flair_text#3, body#4, created_utc#10L, id#14, link_id#16, score#20L]
:        +- *(2) Filter isnotnull(pythonUDF0#210)
:           +- BatchEvalPython [idtype(link_id#16)], [author_flair_text#3, body#4, created_utc#10L, id#14, link_id#16, score#20L,
pythonUDF0#210]
:              +- *(1) FileScan parquet [author_flair_text#3,body#4,created_utc#10L,id#14,link_id#16,score#20L] Batched: true, Format:
Parquet, Location: InMemoryFileIndex[file:/home/cs143/data/comments-minimal.parquet], PartitionFilters: [], PushedFilters: [], ReadSchema:
struct<author_flair_text:string,body:string,created_utc:bigint,id:string,link_id:string,score:big...
+- BroadcastExchange HashedRelationBroadcastMode(List(input[0, string, true]))
   +- *(3) Project [id#69 AS sub_id#178, title#106, score#92L AS sub_score#179L]
      +- *(3) Filter isnotnull(id#69)
         +- *(3) FileScan parquet [id#69,score#92L,title#106] Batched: true, Format: Parquet, Location: InMemoryFileIndex[file:/home/
         cs143/data/submissions.parquet], PartitionFilters: [], PushedFilters: [IsNotNull(id)], ReadSchema: struct<
         id:string,score:bigint,title:string>
+-------+-------+--------------------+-----------+----------------+---------+------+--------------------+---------+
|     id|link_id|                body|created_utc|author_flair_text|com_score|sub_id|               title|sub_score|
+-------+-------+--------------------+-----------+----------------+---------+------+--------------------+---------+
|dbj1fux| 5jsgsc|It's not *somethi...| 1482456914|            null|        2|5jsgsc|States Won by Tru...|    32933|
|dbj1v79| 5jsgsc|8 years of Trump ...| 1482457519|            null|       -1|5jsgsc|States Won by Tru...|    32933|
|dbj3kcj| 5jsgsc|Not sure why you'...| 1482459970|            null|        4|5jsgsc|States Won by Tru...|    32933|
|dp9jsqg| 7ae66c|It totally was. S...| 1509666533|            null|        1|7ae66c|Sen. Franken Dema...|    33761|
|dqhkn5n| 7g7ai2|Did you hit your ...| 1511923384|            null|        9|7g7ai2|Leaked Bank Recor...|    44539|
|dpys5pe| 7dl6b9|[Let](https://www...| 1510938332|            null|        4|7dl6b9|17 women have acc...|    61602|
|dp4mqod| 79qrlp|You come up with ...| 1509421356|            null|       12|79qrlp|Murdoch-owned out...|    31161|
|doprhwy| 77wg4b|I would hope not ...| 1508661918|            null|        4|77wg4b|‘My pain is every...|    16637|
|djwhrjl| 6lsyzn|"Each week, milli...| 1499436660|            null|       44|6lsyzn|The Trump Adminis...|    30747|
|drwtbw7| 7mukrw|Luckily, he's too...| 1514561784|            null|        2|7mukrw|Trump Exposed Ign...|      307|
|duz4g4k| 80wheb|Nah bra! Didn't y...| 1519845603|            null|        1|80wheb|Elizabeth Warren:...|    28400|
|duz4tgl| 80wheb|Nah bra! Didn't y...| 1519845936|            null|      356|80wheb|Elizabeth Warren:...|    28400|
|ddqjv2x| 5u08xc|The entire Trump ...| 1487091552|            null|        1|5u08xc|NBC's Matt Lauer ...|     3749|
|ddqp46k| 5u0wrd|Yeah... T_D and a...| 1487097437|            null|       20|5u0wrd|Russian spy ship ...|     3924|
|dhj572f| 6b07mk|"Autocratic one-p...| 1494732643|            null|       20|6b07mk|At 3 a.m., NC Sen...|    37387|
|difzsor| 6f7284|The Republicans a...| 1496585528|            null|       16|6f7284|Trump Mocks Londo...|    10736|
|dmb5o84| 6wusb2|That's a symptom ...| 1504056827|            null|       13|6wusb2|Everyone’s a Soci...|     6095|
|dowtpmd| 78uaj7|Tell that to the ...| 1509017460|            null|       10|78uaj7|Donald Trump deci...|     9880|
|dowv7zh| 78uaj7|The black market ...| 1509020238|            null|        5|78uaj7|Donald Trump deci...|     9880|
|dowx030| 78uaj7|Probably with the...| 1509022953|            null|        4|78uaj7|Donald Trump deci...|     9880|
+-------+-------+--------------------+-----------+----------------+---------+------+--------------------+---------+
only showing top 20 rows
```

Figure 8: Explain on Join Operation