Project 2B Report

Yutong Han 705025619

June 10, 2018

1 Time Series Plot

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

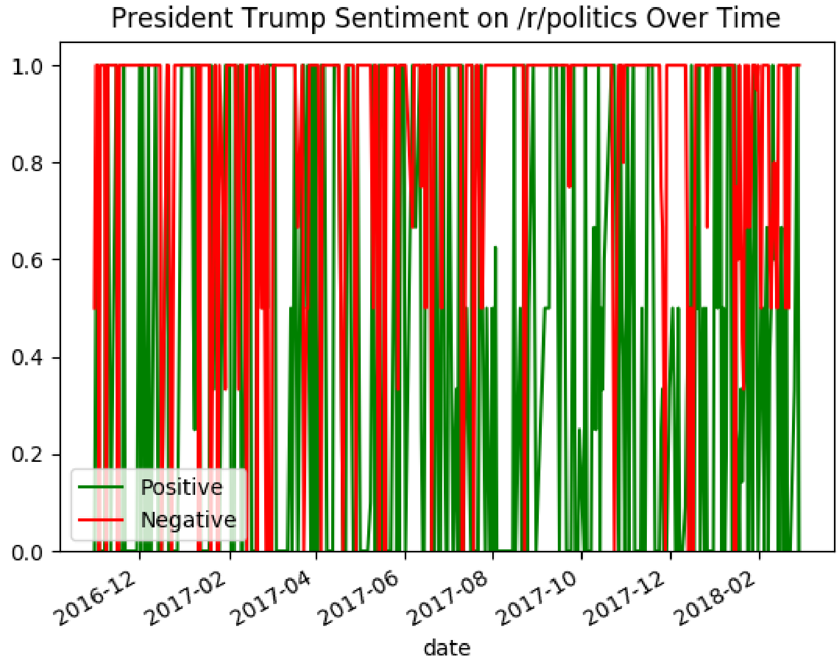


Figure 1: Time Series Plot

2 Positive and Negative Sentiment of US State

The Figure.2 and Figure.3 shows the time series plot(by day) of positive and negative sentiment.

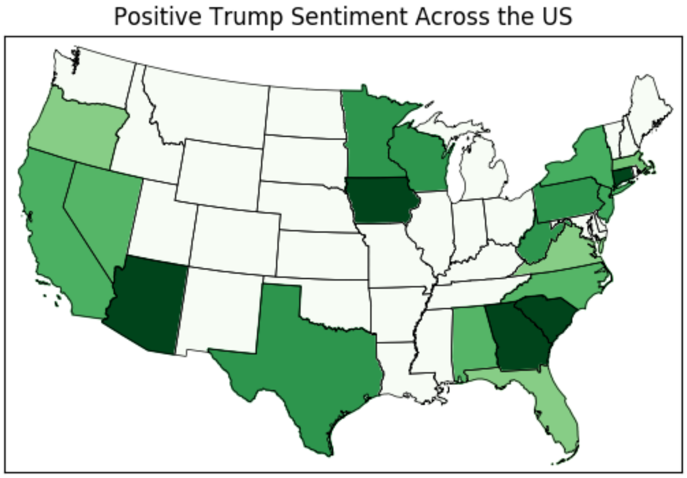


Figure 2: Positive Sentiment over US States

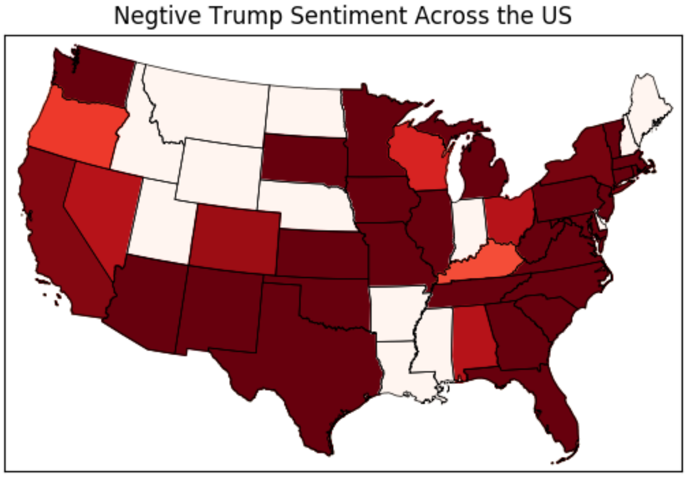


Figure 3: Negative Sentiment over US States

3 Sentiment Difference of US State

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

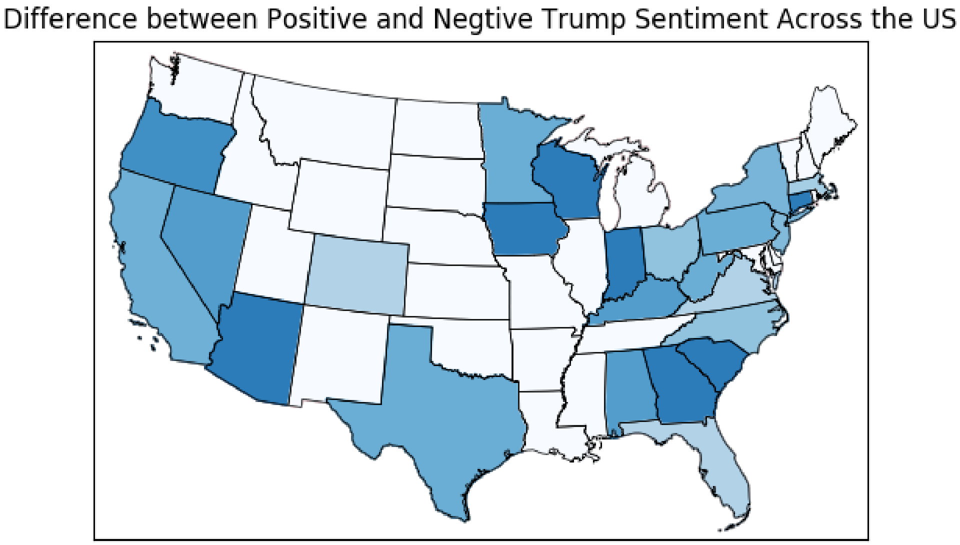

Difference between Positive and Negtive Trump Sentiment Across the US

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

4 Top 10 positive stories

5 Create TWO scatterplots where the X axis is the submission score,

6 Produce the ROC curves

7 summarizing your finding

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* → *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*

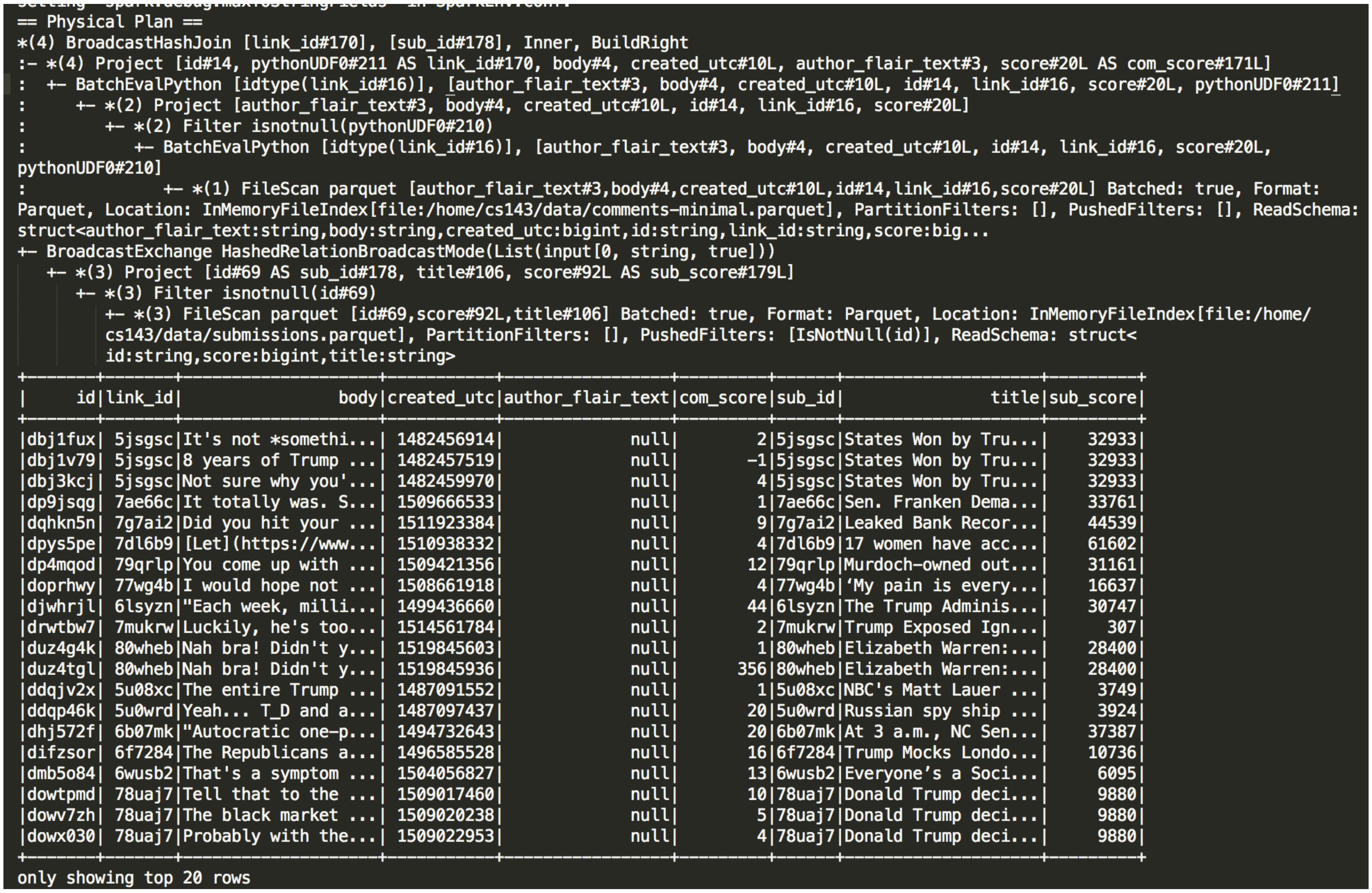
10 Question 3

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.5 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 5: Explain on Join Operation