PROJECT 2B REPORT

Yining Wang and Zhicheng Ren

QUESTION 1: Take a look at `labeled_data.csv`. Write the functional dependencies **implied** by the data.

Input_id->labldem
Input_id->labelgop
Input_id->labeldjt

QUESTION 2: Take a look at the schema for comments. 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?

It's obviously not normalized.

There are lots of redundancies. For example, subreddit -> subredditid should be moved out because for a subreddit there is only one subredditid. We can have a seperate table of (subreddit, subredditid) and deletes the subredditid attribute from the comment table. Also, author->can_glided is also redundant because of the same reason. We can remove can_glided from the comment table and make a seperate table of (author, can_glided). Also, Author, subreddit ->author_flair_text is also redundant for the same reason. Just like the previous two, we can just delete author_flair_text from the comment table and build a seperate table of (author, subreddit, author_flair_text).

The collector stored it in this way probably because it makes it easier to look at everything about a certain comment because he can just looks directly at one table- the comment table. Had he makes it normalized he would need to do a join if he wants to look t every attribute of a comment and that would cost a lot of time considering the scale of the data.

QUESTION 3: Pick one of the joins that you executed for this project. Rerun 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?

The join I used explain on is:

```
"""select
        Input_id, labeldem, labelgop, labeldjt, body
        from labeled_data
        join comments
        on Input_id = id"""
```

in test 2 and 3

the output is:

```
+-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------+
|plan
|
+-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------+
|== 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, true]))
    :   +- *(1) Project [Input_id#170, labeldem#171, labelgop#172, labeldjt#173]
    :       +- *(1) Filter isnotnull(Input_id#170)
    :                                                   +- *(1) FileScan parquet
[Input_id#170,labeldem#171,labelgop#172,labeldjt#173] Batched: true, Format: Parquet,
Location: InMemoryFileIndex[file:/media/sf_vm-shared/labeled_data.parquet],
PartitionFilters: [], PushedFilters: [IsNotNull(Input_id)], ReadSchema:
struct<Input_id:string,labeldem:string,labelgop:string,labeldjt:string>
    +- *(2) Project [body#4, id#14]
        +- *(2) Filter isnotnull(id#14)
            +- *(2) FileScan parquet [body#4,id#14] Batched: true, Format: Parquet,
Location: InMemoryFileIndex[file:/media/sf_vm-shared/comments.parquet],
PartitionFilters: [], PushedFilters: [IsNotNull(id)], ReadSchema:
struct<body:string,id:string>|
+-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------+
```

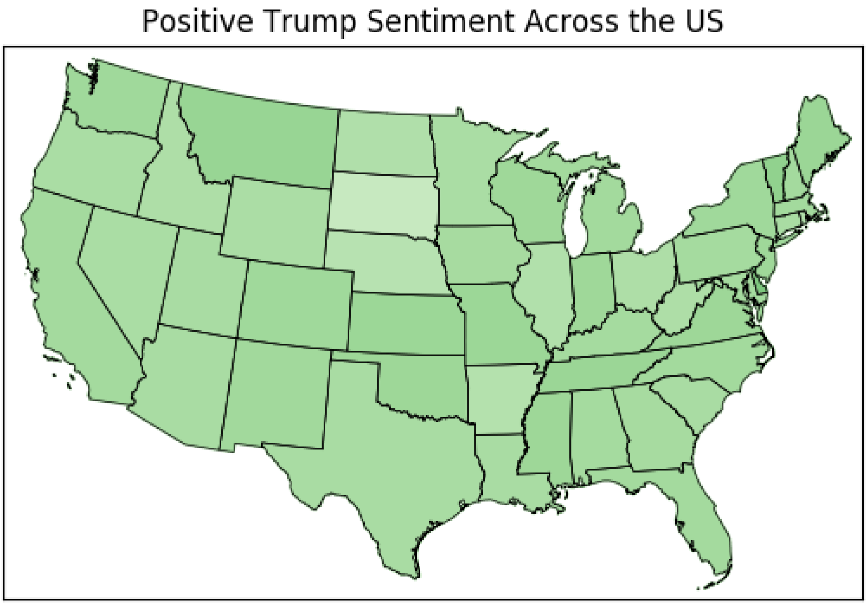
I notice that spark is using broadcast hashjoin to join the tables

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.

We are not able to draw it because we don’t know how take care about the day in a year thing, but we did get a csv file of it

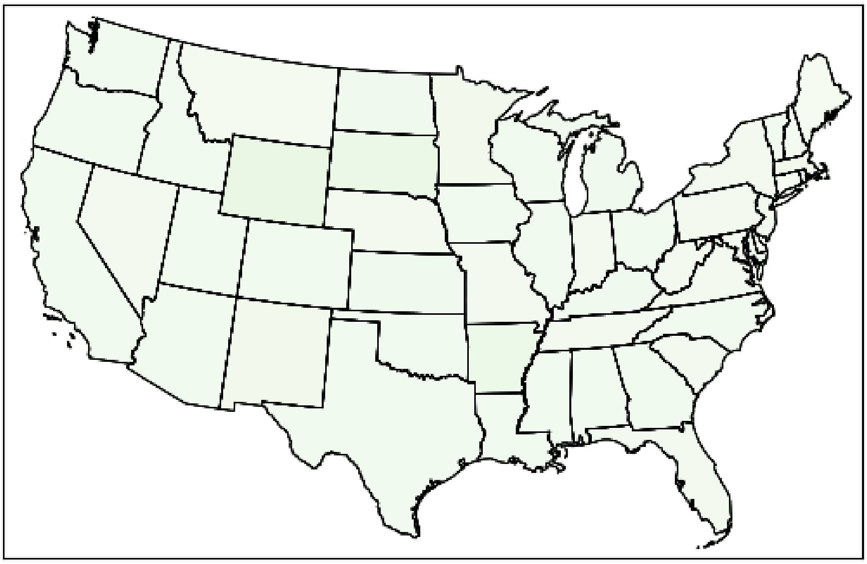
2. Create 2 maps of the United States: one for positive sentiment and one for negative sentiment. Color the states by the percentage.

Positive:

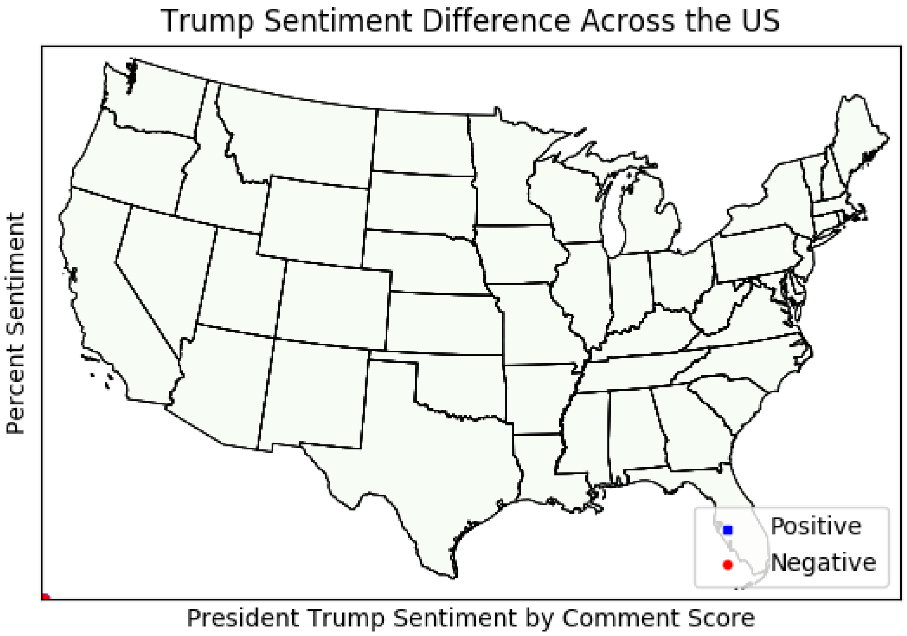


Negative:

Negative Trump Sentiment Across the US



3. Create a third map of the United States that computes the *difference:* %Positive - %Negative.


Trump Sentiment Difference Across the US
Percent Sentiment
Positive
Negative
President Trump Sentiment by Comment Score

4.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.

Negative by title

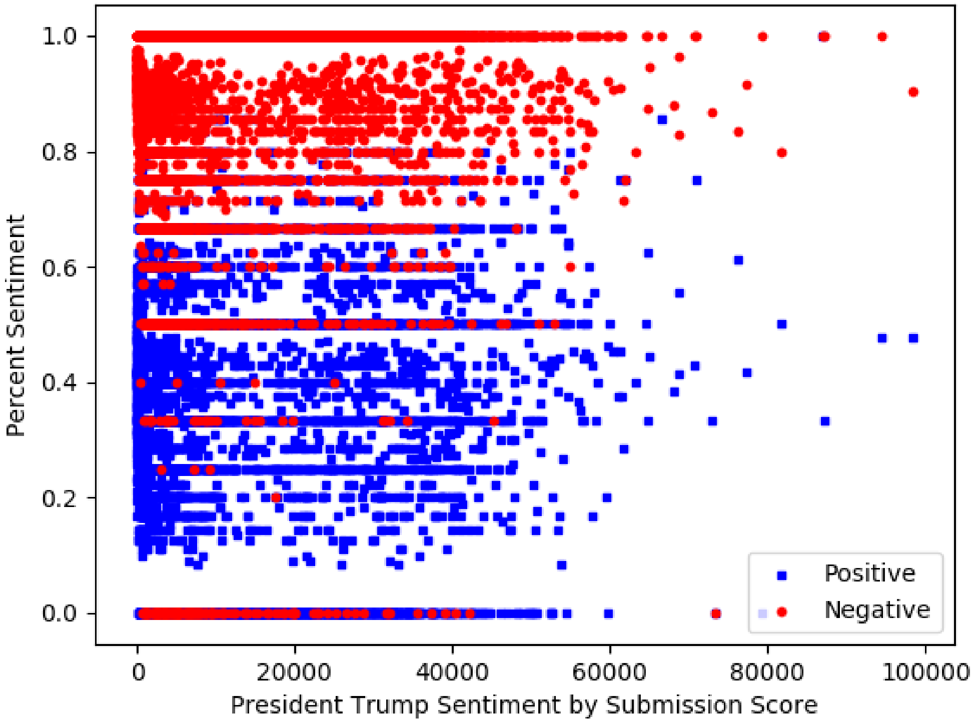
1	Sen. Franken Demands Attorney General Jeff Sessions Explain Himself Amid New Trump-Russia Revelations										
1	Leaked Bank Records Tie Russian Money To Kushner Startup He Didnâ€™t Disclose										
1	Murdoch-owned outlets bash Mueller, seemingly in unison										
1	â€˜My pain is everydayâ€™: After Weinsteinâ€™s fall, Trump accusers wonder: Why not him?										
1	Trump Exposed Ignorance During Trade Talk With Merkel, Leaving White House Aides Humiliated: Report										
1	Elizabeth Warren: \Equifax may actually make money off this breach\""										
1	NBC's Matt Lauer to Conway: Your defense of the WH over Flynn 'makes no sense'										

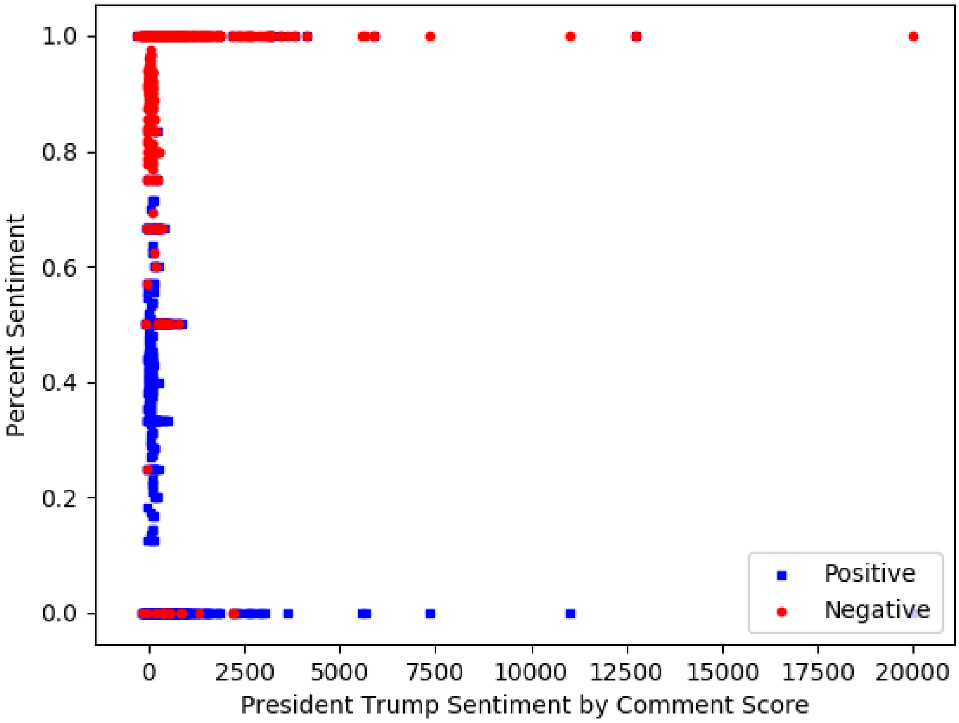
1	Russian spy ship off the east coast of US, officials say						
1	Everyoneâ€™s a Socialist After a Natural Disaster						
1	Donald Trump decided not to deport fugitive accused of rape 'after learning he is a Mar-a-Lago member'						

Positive by title:

1	Trump Exposed Ignorance During Trade Talk With Merkel, Leaving White House Aides Humiliated: Report					
1	Donald Trump's approval rating slips to all-time low in new poll					
1	History wonâ€™t be unkind to Trumpâ€”it will be cruel					
1	Senator Sanders: Our incoming President is a 'Patholigical Liar'					
1	Warren: Doug Jones should be seated without delay					
1	Illinois passes automatic voter registration					
1	Dems introduce act would authorize FBI Director to revoke WH staff security clearance					
1	The Top 10 Racist Dog Whistles Hidden in Trumpâ€™s State of the Union Address					
1	As First Act, New HHS Secretary Imposes More Medicaid Work Requirements					
1	Dems block 20-week abortion ban					

5.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.


1.0
0.8
0.6
0.4
0.2
0.0
Percent Sentiment
0
20000
40000
60000
80000
100000
President Trump Sentiment by Submission Score
■ Positive
● Negative


1.0
0.8
0.6
0.4
0.2
0.0
Percent Sentiment
0
2500
5000
7500
10000
12500
15000
17500
20000
President Trump Sentiment by Comment Score
■ Positive
● Negative

6.Write a paragraph summarizing your findings. What does /r/politics think about President Trump? Does this vary by state? Over time? By story/submission?

I would say that overall /r/politics don't like Trump. Obviously, this vary by state. As we can see from the maps we drew, obviously only several states in the south or middle west has more positive view on Trump, which are exactly the states that has less negative view on Trump. It doesn't really vary by story or submission because we can tell from the titles and the scatter plots that usually the ones getting a lot of positive sentiment would also get a lot of negative sentiment, so all of them are basically controversial. Generally, although, there are more negative sentiment than positive sentiment. Timewise, the negative sentiment remains at the same amount while the positive sentiment towards trump keeps droping over time.

Overall, r/polities doesn't like trump.