CS 143 Project 2 Report

/r/politics Sentiment Analysis

Question Answers

1. The data in labeled_data.csv has the functional dependency:
   Input_id => { labeldem, labelgop, labeldjt }
2. No, the comments schema is definitely not normalized by any means. For example, it contains the property author_flair_text, which, in a normalized database, should only show up in relation to the author, and not the author's various comments. Another example is that both the subreddit_id and subreddit are present in the schema, which is obviously redundant. To decompose the schema, we would create a separate relations for each "entity" other than comments that exists in the schema, but include a foreign key for each of these new relations, e.g. make new relations such as:
   { subreddit_id, subreddit }
   { author, author_flair_text, author_flair_css, author_cakeday, can_gild }
   …
   but maintain references like { subreddit_id, author } within the comments schema.

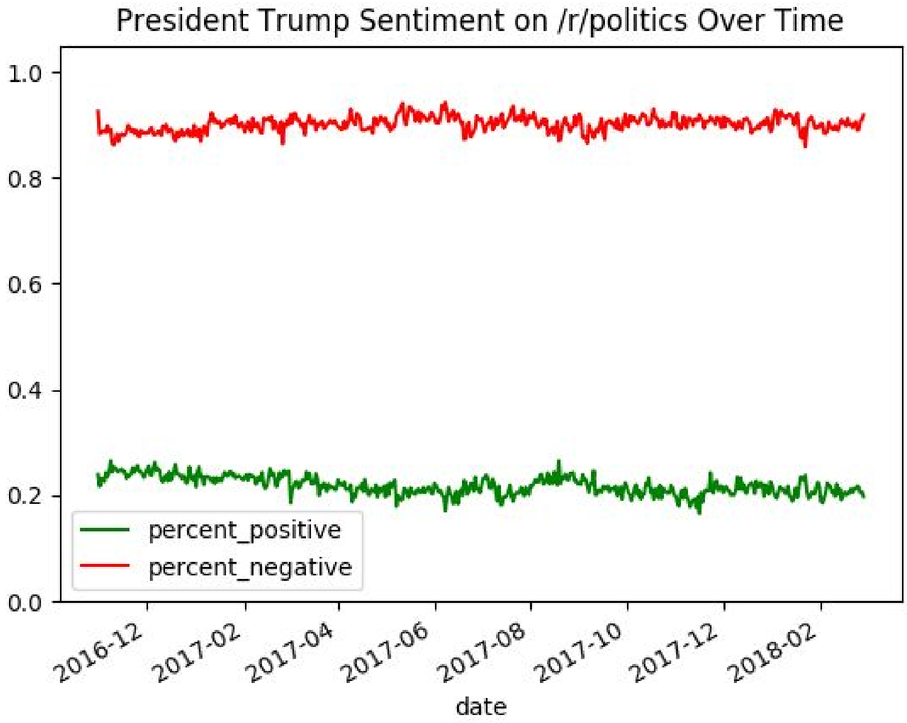
   It's likely that the data collector chose to create a schema that isn't normalized because it allows us to have a lot of potentially relevant data in a central location without needing to perform joins. Normalized data is important in a long-term database but isn't as important when doing data analytics like in our situation.
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 that we ran the .explain() on is labeled_data.join(comments, labeled_data['Input_id'] == comments['id'], 'inner').explain(). Spark seems to be using a Hash join. The hash join build the left table and broadcasts it in order to compare it to the other table. It then projects some of attributes of the join, where one of the attributes can't be Null. The data is read from labeleddata.csv and in the process of reading it creates an index for the data that is read.
Now the right hand side, which is read from the comments.parquet file, has a in memory index. It selects 2 attributes from the RHS table and checks if the id isn't null. It checks that it isn't null since it is doing a left inner join on the comments[id] and the input_id.

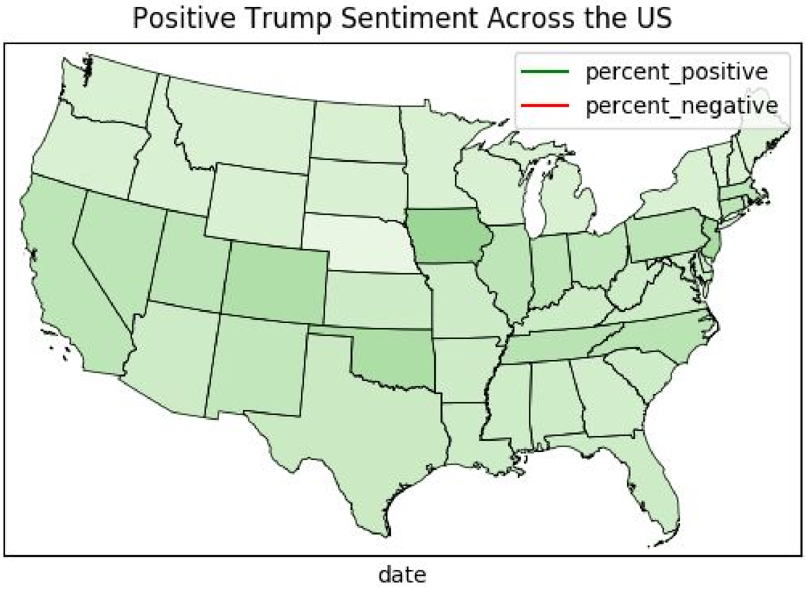
Here is the output of the .explain() command:

```
== Physical Plan ==
*(2) BroadcastHashJoin [Input_id#180], [id#14], Inner, BuildLeft
:- BroadcastExchange HashedRelationBroadcastMode(List(input[0, string, true]))
:  +- *(1) Project [Input_id#180, labeldem#181, labelgop#182, labeldjt#183]
:     +- *(1) Filter isnotnull(Input_id#180)
:        +- *(1) FileScan csv [Input_id#180,labeldem#181,labelgop#182,labeldjt#183] Batched: false,
Format: CSV, Location: InMemoryFileIndex[file:/media/sf_vm-shared/2B/labeled_data.csv],
PartitionFilters: [], PushedFilters: [IsNotNull(Input_id)], ReadSchema:
struct<Input_id:string,labeldem:int,labelgop:int,labeldjt:int>
+- *(2) Project [id#14, body#4]
   +- *(2) Filter isnotnull(id#14)
      +- *(2) FileScan parquet [body#4,id#14] Batched: true, Format: Parquet, Location:
InMemoryFileIndex[file:/media/sf_vm-shared/2B/comments.parquet], PartitionFilters: [],
PushedFilters: [IsNotNull(id)], ReadSchema: struct<body:string,id:string>
```

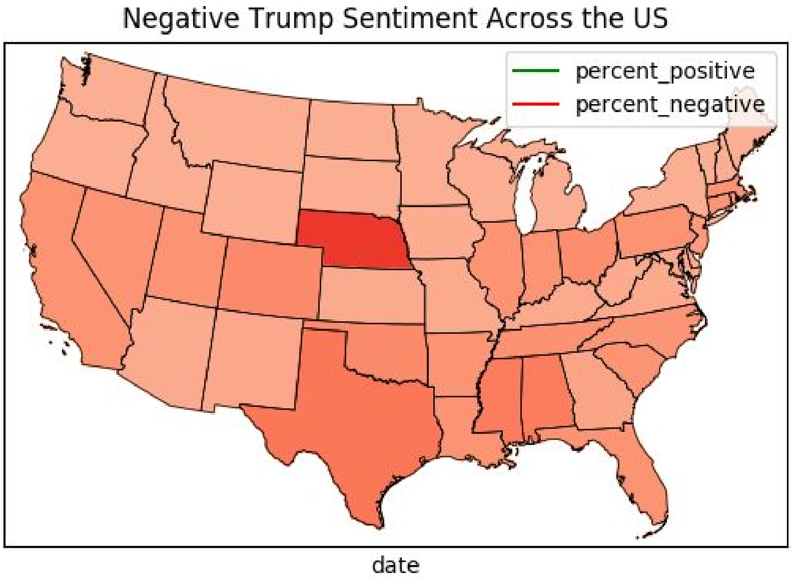
Plots



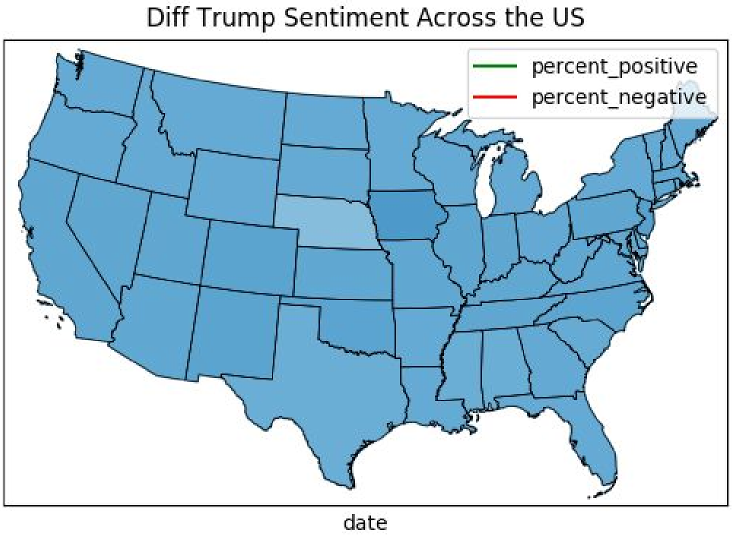
Plot 1: Positive and negative sentiment over time



Plot 2: Positive sentiment per state



Plot 3: Negative sentiment per state



Plot 4: Difference in sentiment across state

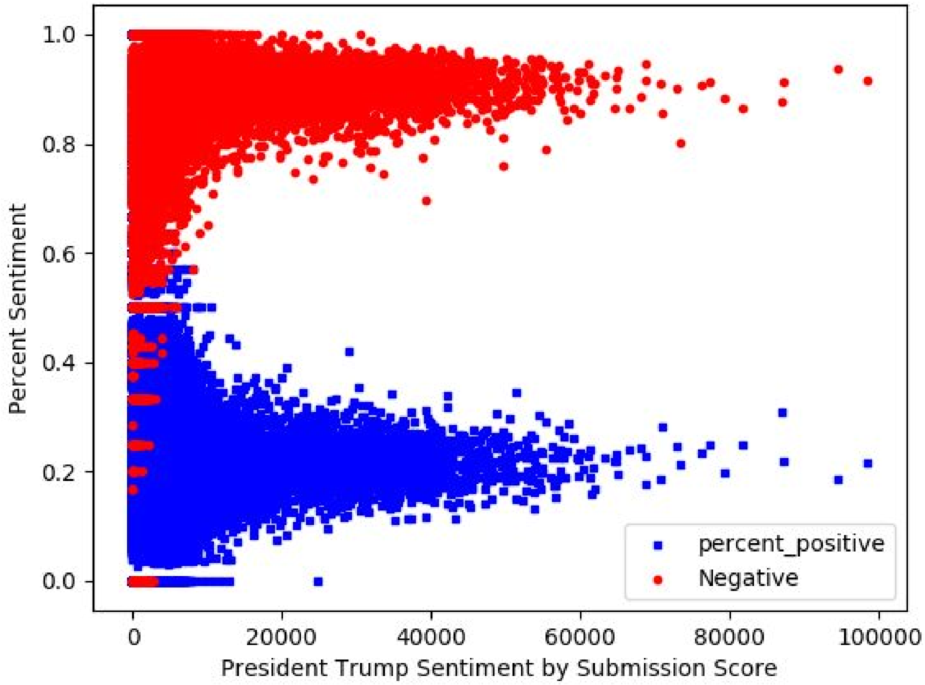
title	percent_negative
Donald J. Trump elected president of the United States	1
A Clinton win is 70 to 75% likely, according to a Pimco analyst	1
British newspapers react to judges' Brexit ruling: 'Enemies of the people' | Politics	1
U.S. House Speaker Ryan renews call to suspend classified briefings for Clinton	1
There was no gun!	1
Bill Weld on Rachel Maddow: ‘I’m Here Vouching for Mrs. Clinton’	1
If you're in Philly, there are free rides via Uber and Lyft to the polls on Election Day	1
Historic Mississippi black church burned and vandalized with ‘Vote Trump’ graffiti	1
#nonpartisan #thissucks #wawa	1
Investigating Donald Trump, F.B.I. Sees No Clear Link to Russia	1

Figure 1: Top 10 Most Negative Submissions

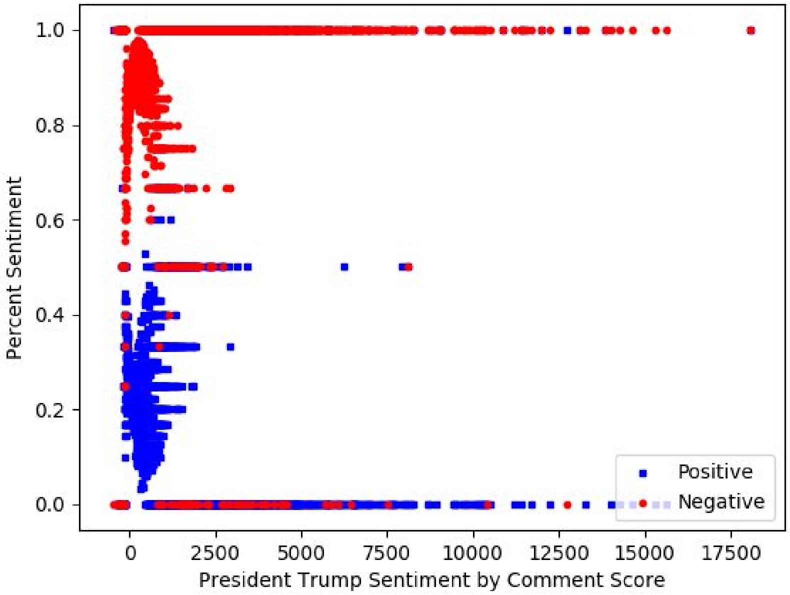
title	percent_positive
Another Clinton Murder? Woman Investigating Child Trafficking In Haiti Dies Under Suspicious Circumstances	1

UNC Professor Allowed to Harass LGBTQ Students	1
Michigan Fights To Avoid Delivering Water To Flint Residents	1
Corbyn is 'The Maddest Person In The Room' - Says Bill Clinton	1
Politics Trump Foundation admits to violating ban on ‘self-dealing,’ new filing to IRS shows	1
How Trump Will Make America Great Again? Infographic	1
Trump taps S.C. Gov. Nikki Haley as U.N. ambassador	1
Google fixes NYC Trump Tower name after changed to ‘Dump Tower’ on Google Maps	1
Trump's Win Upends Climate Fight	1
Lies in the Guise of News in the Trump Era	1

Figure 2: Top 10 Most Positive Submissions



Plot 5a: Sentiment and Submission Score



Plot 5b: Sentiment and submission score

Findings

Looking at our findings it seems that /r/politics tends to have negative feelings towards Donald Trump. Plot 1 shows that the difference between positive and negative sentiment has stayed rather consistent over time. Positive sentiment seems to stay at a little about 20% and negative sentiment at around 90%. There does seem to be some difference per states looking at plots 2 and 3. Nebraska seems to have the most negative sentiment and the least positive sentiment. States like California, Illinois, Florida and Texas also seem to have slightly more negative sentiment. Positive sentiment seems rather uniform across states, with slightly higher positive sentiment in some states like Ohio. The difference in sentiment graph does not seem to show a huge difference in sentiment in states, meaning states have rather consistent view of Donald Trump. It is important to note that this state data might not be super accurate because we sampled the data to only 20% and relied on commenters marking their state of origin in the author flair, which many did not. Looking at comment score it seems that some highly negative comments have high scores. Submission score seems to show that highly negative submissions also can have high scores. However, there are some positive comments that also have high scores. It is possible these highly rated submissions are the same submissions and have some positive sentiment and negative sentiments. Overall, it seems /r/politics has more negative sentiment towards Donald Trump.