JiHwan(Evan) Kim
504596254

CS143 Project 2B Report

Final Deliverable

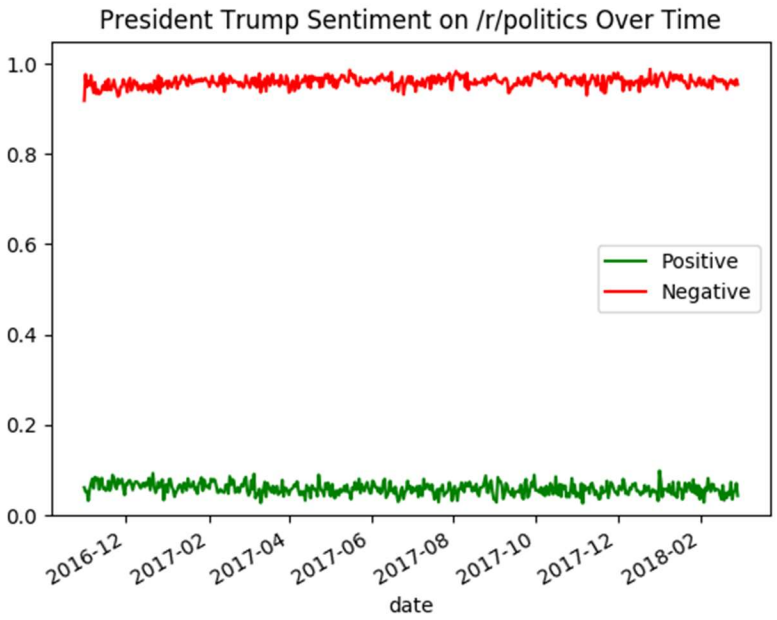


Figure 1 time series plot by day of positive and negative sentiment percentage

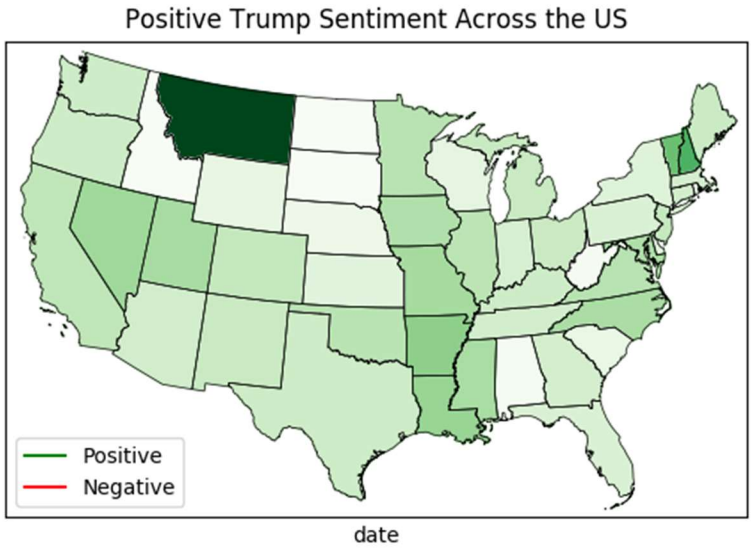


Figure 2 Percentage of positive sentiments across the US

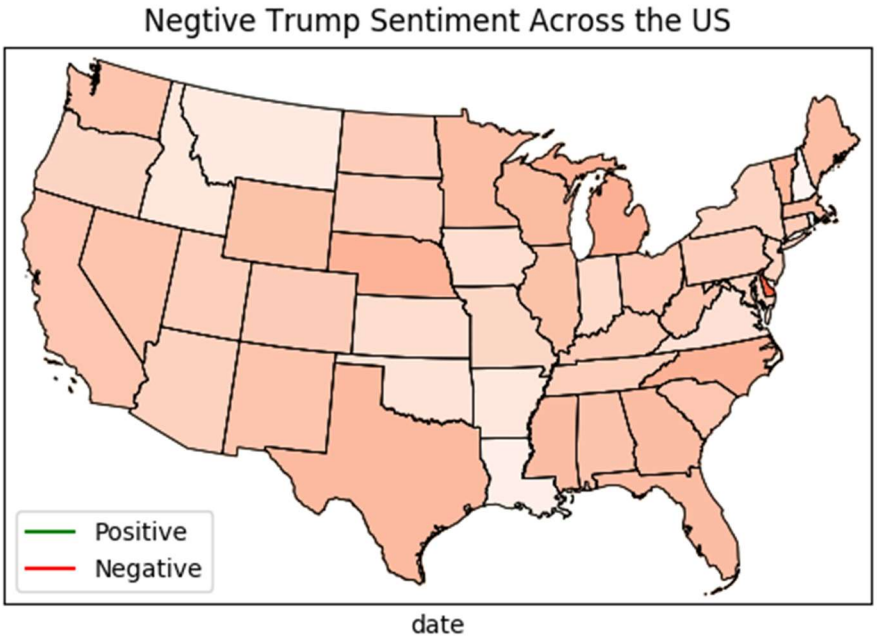


Figure 3 Percentage of negative sentiments across the US

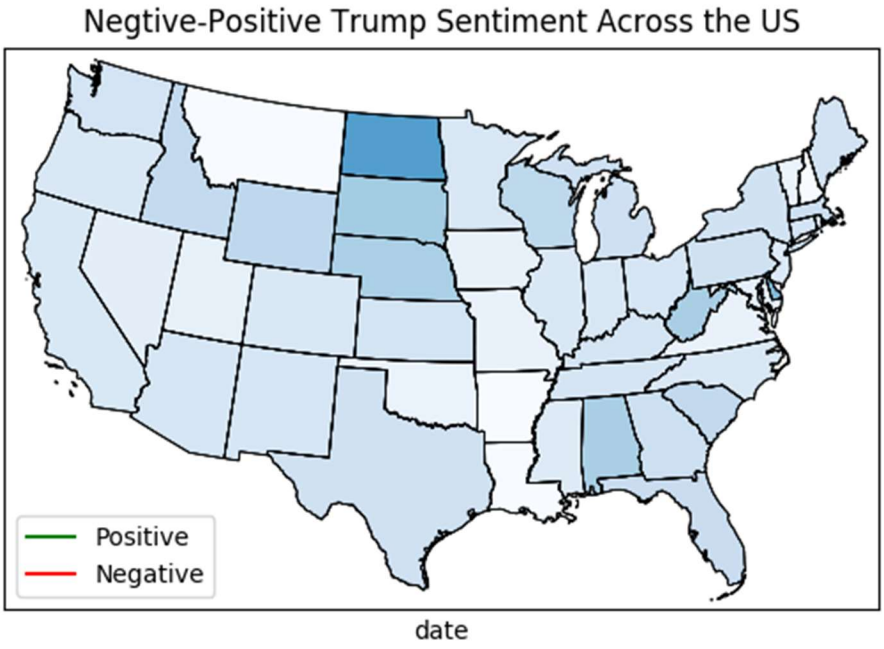


Figure 4 Percentage difference of Positive and Negative sentiments across the US

title	Percentage
Trump: I’m President, So No More Protest or Media Accountability	1
The states taking the opioid epidemic seriously (and not), in one map	1
Republicans are sending in Ted Cruz to salvage a Kansas congressional race that should have been a sure thing	1
Trump aides staged a Twitter 'intervention'	1
While Honoring Pearl Harbor Veterans, Trump Says He'll Serve 7 More Years	1
Hillary Clinton finds power in vulnerability in moving first speech since the election	1
Trump Election Commissioners Are Resisting Efforts to Protect Elections From Hacking	1
Former US spy says Vladimir Putin is trolling Donald Trump over 2016 election meddling	1
Intercepts suggest Sessions discussed Trump campaign matters with Russia envoy: report	1
House passes NRA-backed gun bill	1

Table 1 Top 10 positive sentiments

title	Percentage
Trump: I never said 'Israel' in meeting with Russians	1
Americas Largest Churches Are All Anti-LGBT And Led By Mostly White Men	1
The GOP Can't Stop Blaming Each Other Over Their Health Care Failure	1
Dem group releases new ad showing Nick Saban as Alabama write-in candidate	1
What's behind rich people pretending to be self-made?	1
Hillary Clinton, free to speak her mind	1
Jeff Sessions Just Confessed His Negligence on Russia	1
The stupidity of Trumpcare: Government will spend $34 billion more a year to cover 8.9 million fewer Americans	1
GOP Votes To Roll Back The Americans With Disabilities Act	1
Trudeau offers condolences after Kentucky school shooting	1

Table 2 Top 10 negative sentiments

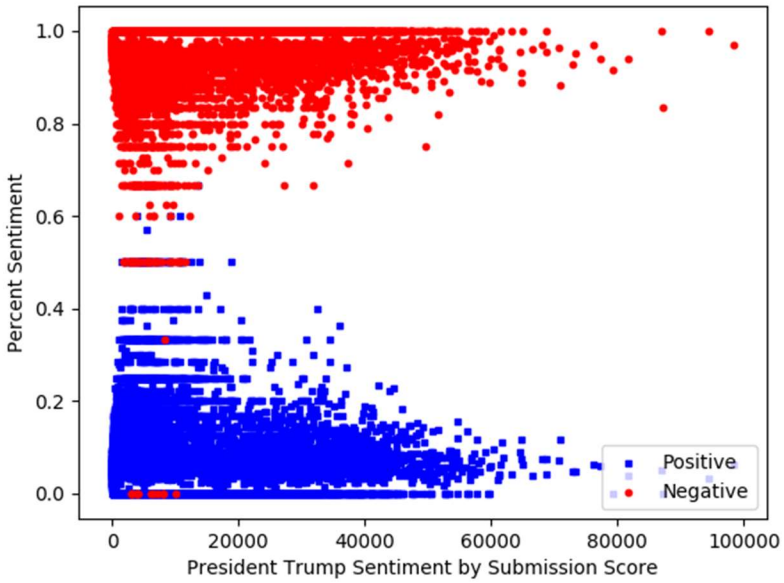


Figure 5 Submission Score scatterplot

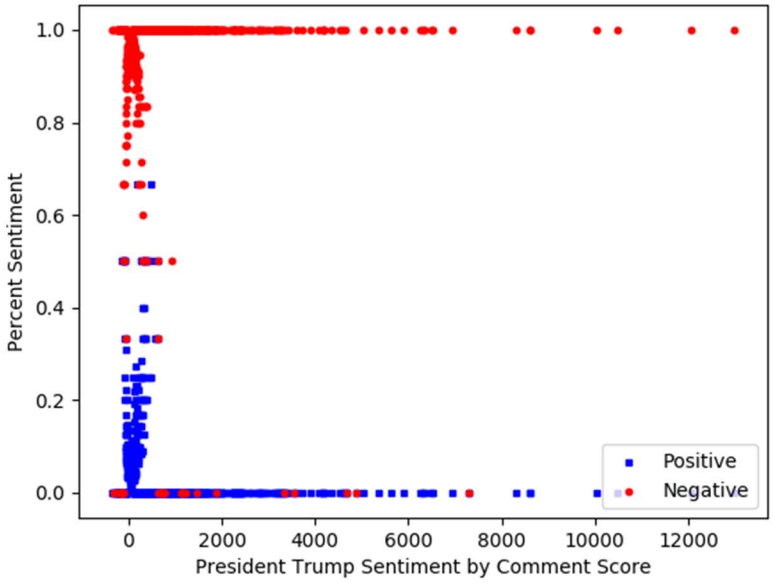


Figure 6 Comment Score scatterplot

In figure 1, it shows that there are significantly more negative sentiments than positive sentiments. It also shows that sentiments do not vary over time as sentiments are stable their percentage. From figure 2 and figure 3, it is shown that sentiments do vary across states and the difference of negative and positive sentiments is show in figure 4 where lighter the color is, the difference in negative and positive sentiments are low and where the deeper the color is, the difference of sentiments is bigger meaning there are more negative sentiments than positive sentiments. Finally from Figure 5 and 6, I can conclude that sentiments vary by submission and comment scores

Questions

Question 1.

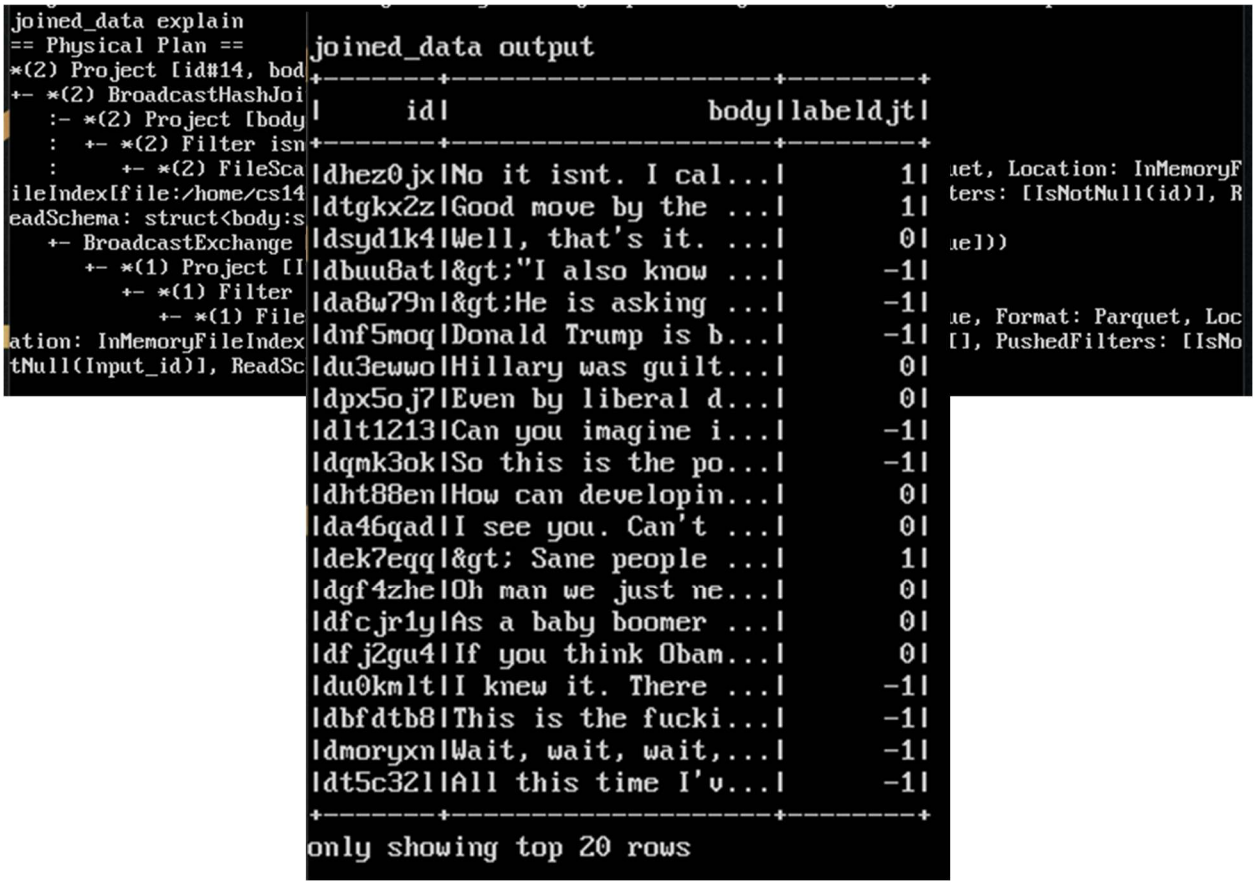
The Functional Dependency implied by labeled_data is Input_id -> {labeldem, labelgop, labeljt}

Question 2.

The comments dataframe does not look normalized since it contains redundant data. In the comments dataframe, there are subreddit and subreddit_id comment. This means that I can decompose the comments dataframe to one relation containing subreddit_id and its corresponding subreddit and another relation containing all of the columns of comments dataframe without subreddit column.

However, decomposing the comments dataframe to two relations like above could cause insert/update integrity. Thus the comments dataframe is stored in its current manner

Question 3.



The above figures are the explain paragraph and output of joined_data in my code. The explain() shows that the sparkSQL is doing a hash-join on the keys. Firstly, the tables are loaded and filter rows with null values on specified key. Then the sparkSQL selects the same attributes in both table and finally joins them with hash-join procedure.