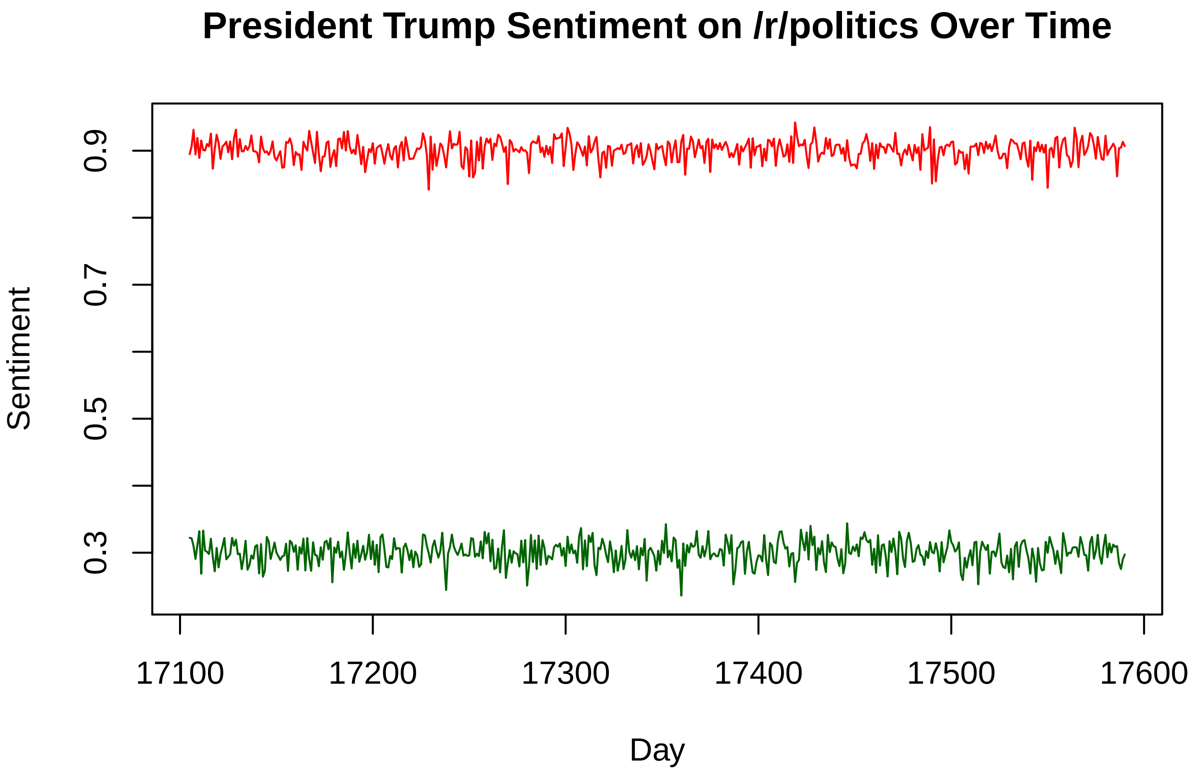
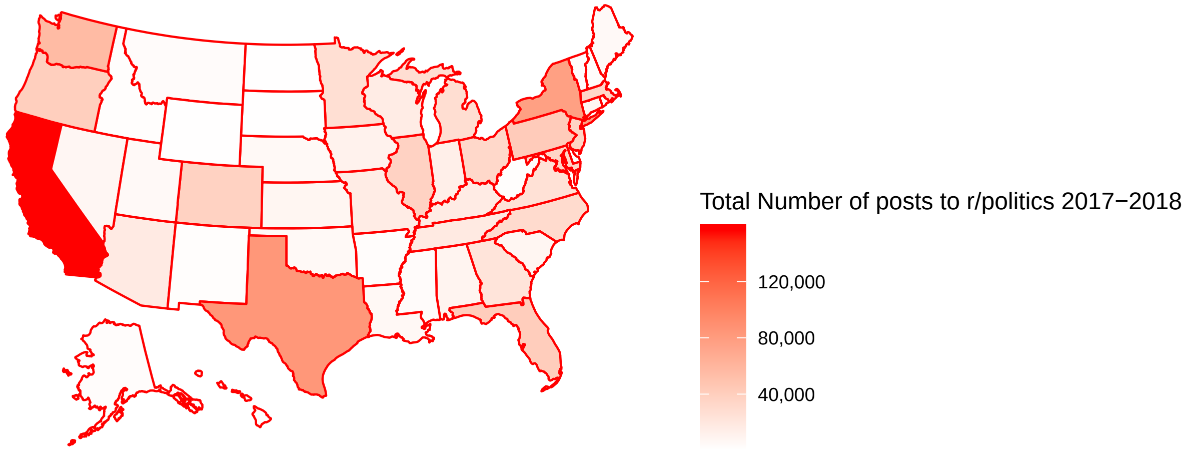
Cs 143 Project 2B Report

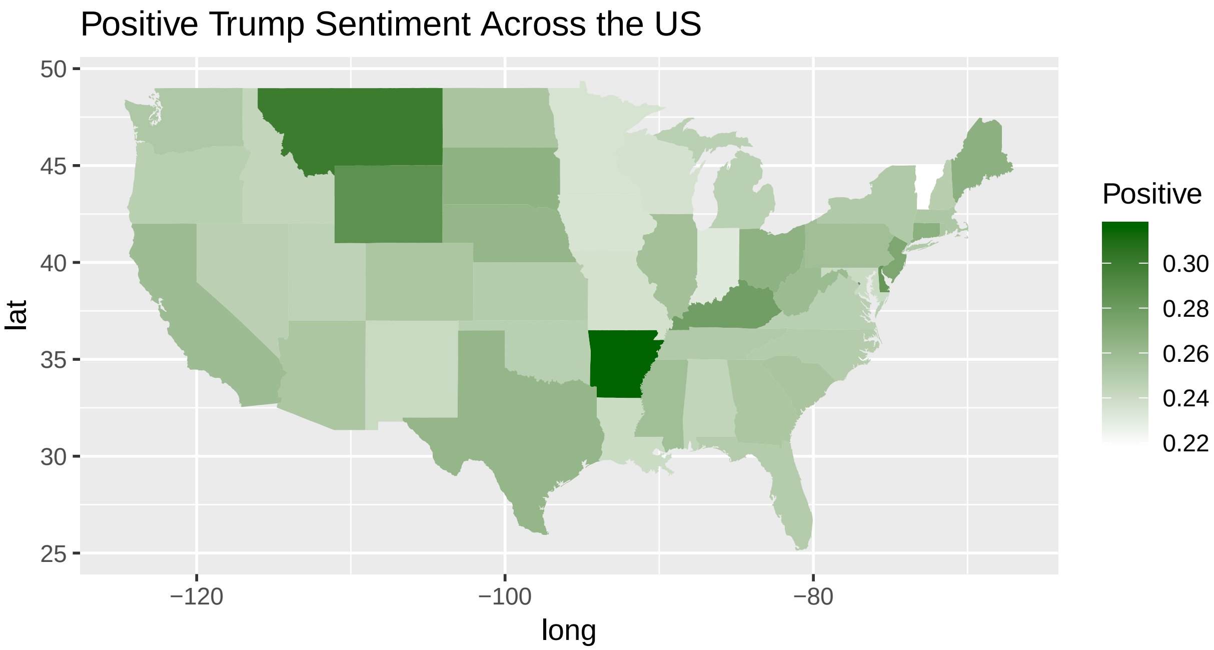
Vincent Chi 304576879

Overall it seems r/politics is overwhelmingly negative towards Trump. From our time series plot, we see that the negative sentiment percentage for trump has not declined at all over the past couple years, and the positive sentiment has not risen. The negative sentiment flucuates slightly around .9 while the positive sits at around .3. These natural fluctuations likely arise from random while perhaps the more significant fluctuations results from current events. With regards to our maps, we see that for negative sentiment, the darker states include Maine, oklahoma, Idaho, and Indiana. This is contrary to my expectations as I would have expected New York or California to be the most negative, rather than these flyover states. For the states with the most postive sentiments, we see that these include Arkansas, Montana, and Wyoming which makes sense as those states are traditionally more right leaning. With regeards to higheset difference by state, we see that it also belongs to the same states as those with the most positive. This is likely due to how trump is so controversial and so negatively viewed upon on reddit that having any positive sentiment for trump is enough to stand out and cause this difference. From the submission score vs percentage, we see that the higher the submission score, the slightly more neutral the percentages actually lie, meaning the positive sentiment percentage actually raises up a little bit while the negative percentage lowers on average. However, this difference is very slight and r/politics still has an overwhelmingly negatiev view towards trump. With comment score, this is more evident as those comments with the highest score have the most negative sentiment towards trump. With regards to the stories, the most positive ones seem to almost be sarcastically/ironically positive. Example: how compared to hillary clinton, trump's sexual assault isn't so bad or how a psychologist releases a terrifying diagnosis of him, or his most controversial moments such as his remarks towards Haiti which some people found funny and true and his war on "Fake News" which a lot of peoplee on both the right and left can get behind. His most negative stories are highly critical of trump and his actions as expected. from our extra plot of number of posts by state we see that a large proportion of posts come from California as well, which is very left leaning.

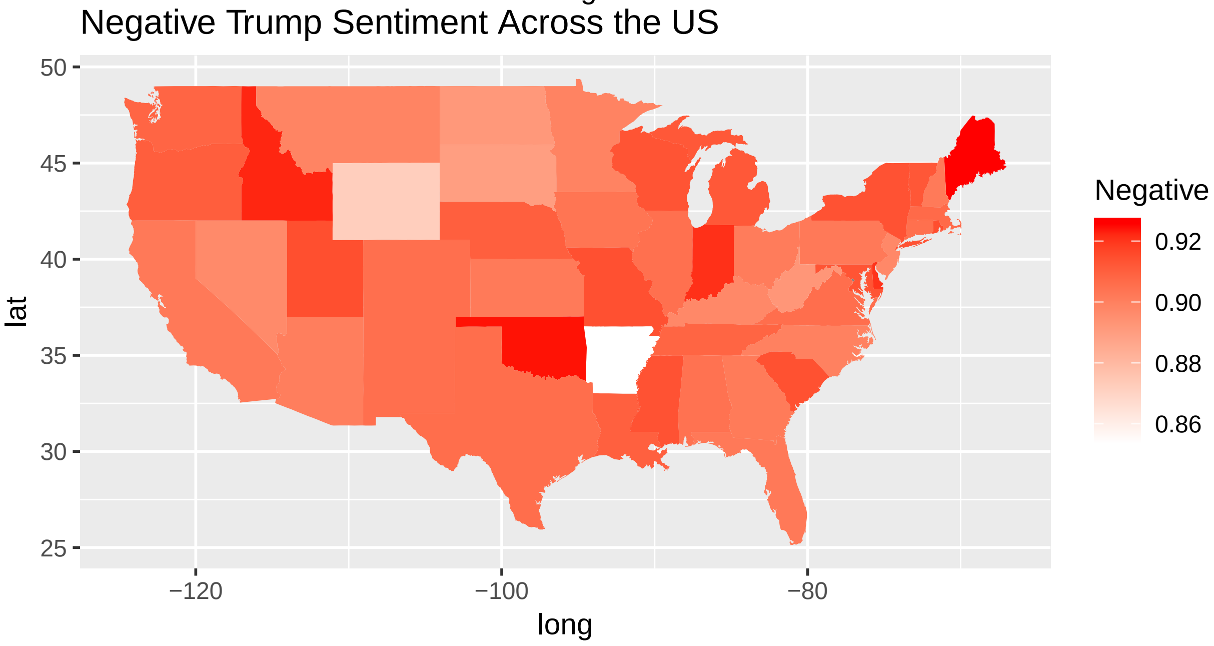
However, we need to realize that the data came only from reddit, which has its own demographic of users and can't be used to generalize to the population that's also filled with aging flyover-state boomers who probably love trump but don't go on reddit, which is likely filled with Bernie Fanatics.


President Trump Sentiment on /r/politics Over Time
Sentiment
0.3
0.5
0.7
0.9
17100
17200
17300
17400
17500
17600
Day

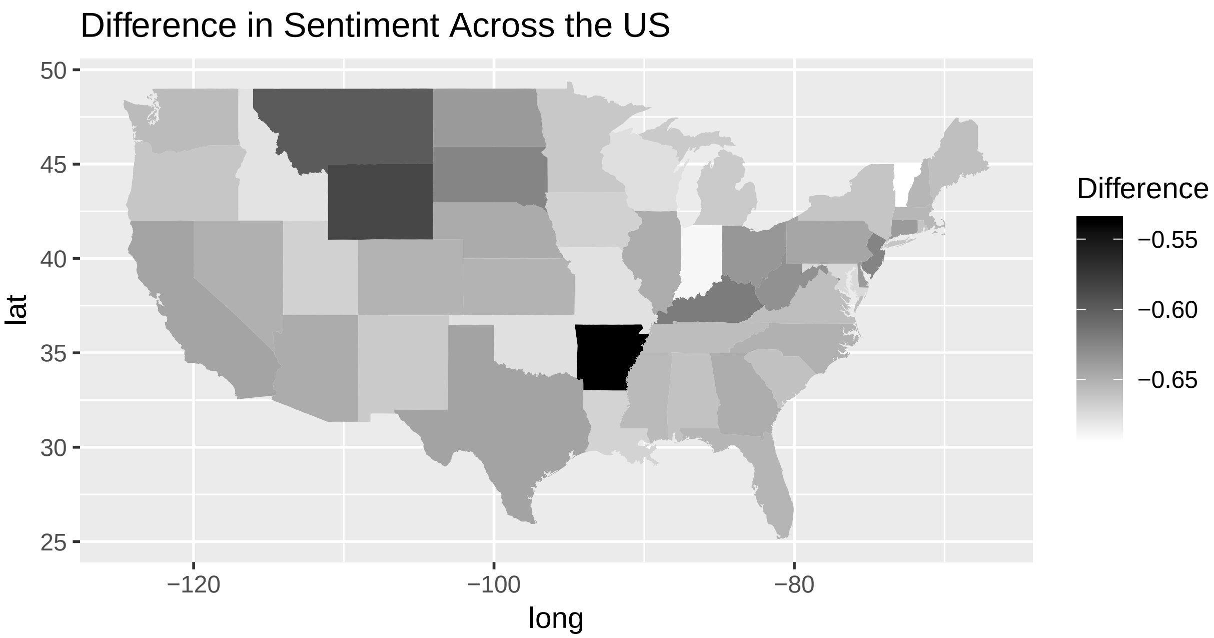

Total Number of posts to r/politics 2017−2018
120,000
80,000
40,000



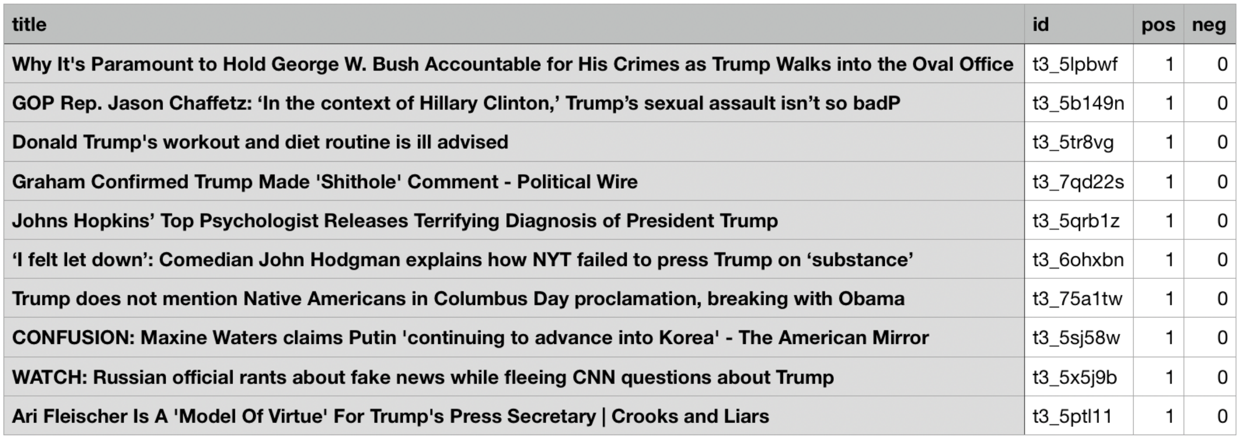
Positive Trump Sentiment Across the US



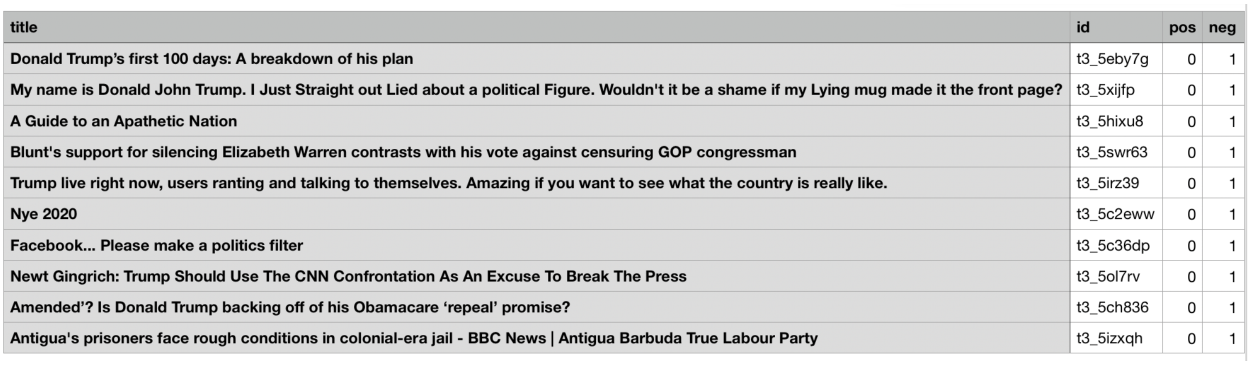
Negative Trump Sentiment Across the US



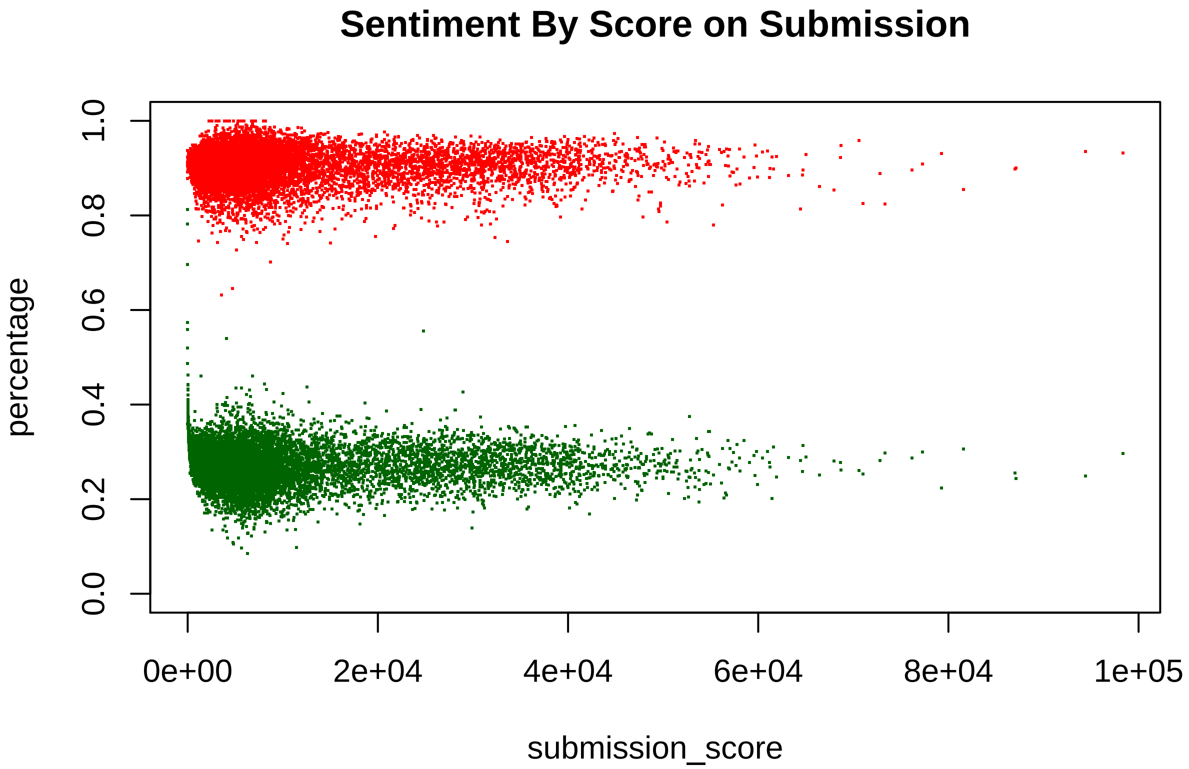
Top 10 Positive and Negative stories by Percentage

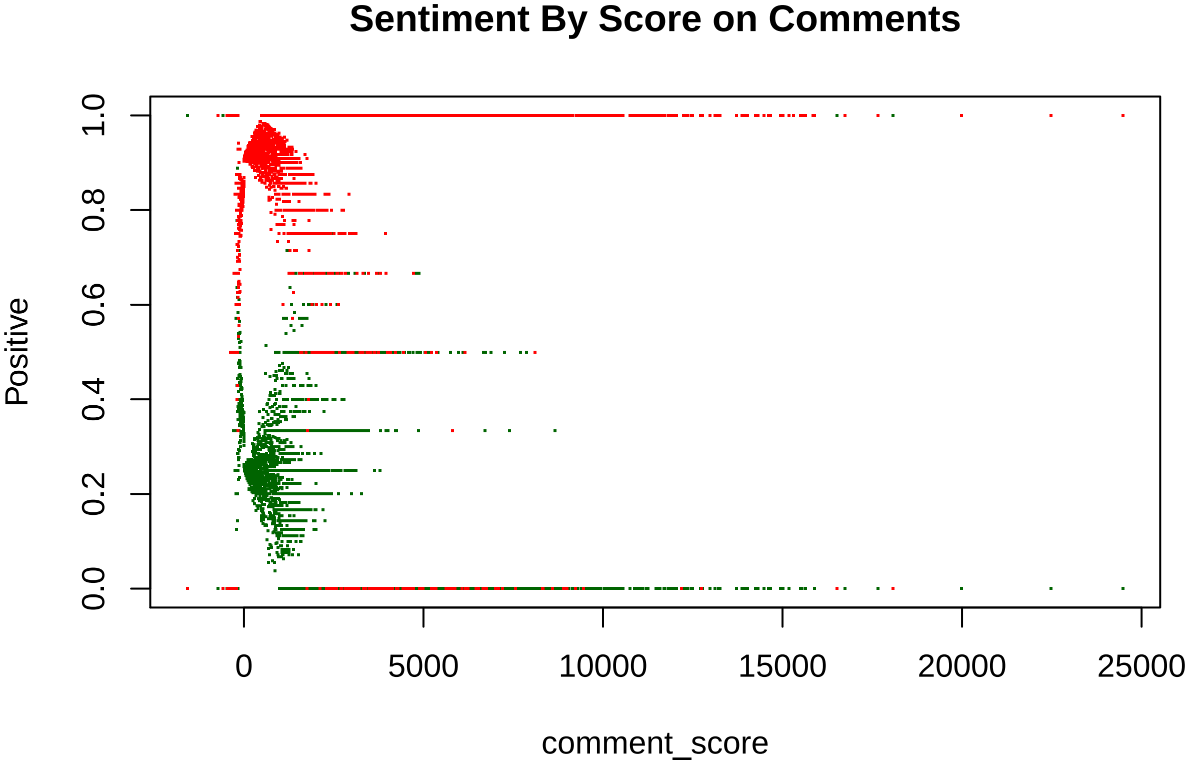


title	id	pos	neg
Why It's Paramount to Hold George W. Bush Accountable for His Crimes as Trump Walks into the Oval Office	t3_5lpbwf	1	0
GOP Rep. Jason Chaffetz: ‘In the context of Hillary Clinton,’ Trump’s sexual assault isn’t so badP	t3_5b149n	1	0
Donald Trump's workout and diet routine is ill advised	t3_5tr8vg	1	0
Graham Confirmed Trump Made 'Shithole' Comment - Political Wire	t3_7qd22s	1	0
Johns Hopkins’ Top Psychologist Releases Terrifying Diagnosis of President Trump	t3_5qrb1z	1	0
‘I felt let down’: Comedian John Hodgman explains how NYT failed to press Trump on ‘substance’	t3_6ohxbn	1	0
Trump does not mention Native Americans in Columbus Day proclamation, breaking with Obama	t3_75a1tw	1	0
CONFUSION: Maxine Waters claims Putin 'continuing to advance into Korea' - The American Mirror	t3_5sj58w	1	0
WATCH: Russian official rants about fake news while fleeing CNN questions about Trump	t3_5x5j9b	1	0
Ari Fleischer Is A 'Model Of Virtue' For Trump's Press Secretary | Crooks and Liars	t3_5ptl11	1	0



title	id	pos	neg
Donald Trump’s first 100 days: A breakdown of his plan	t3_5eby7g	0	1
My name is Donald John Trump. I Just Straight out Lied about a political Figure. Wouldn't it be a shame if my Lying mug made it the front page?	t3_5xijfp	0	1
A Guide to an Apathetic Nation	t3_5hixu8	0	1
Blunt's support for silencing Elizabeth Warren contrasts with his vote against censuring GOP congressman	t3_5swr63	0	1
Trump live right now, users ranting and talking to themselves. Amazing if you want to see what the country is really like.	t3_5irz39	0	1
Nye 2020	t3_5c2eww	0	1
Facebook... Please make a politics filter	t3_5c36dp	0	1
Newt Gingrich: Trump Should Use The CNN Confrontation As An Excuse To Break The Press	t3_5ol7rv	0	1
Amended’? Is Donald Trump backing off of his Obamacare ‘repeal’ promise?	t3_5ch836	0	1
Antigua's prisoners face rough conditions in colonial-era jail - BBC News | Antigua Barbuda True Labour Party	t3_5izxqh	0	1

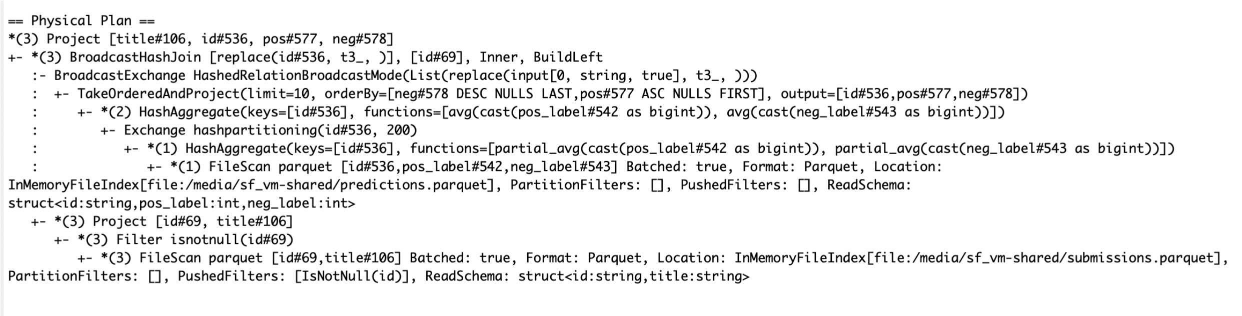




##QUESTIONS:

Question 1

Input_ID -> labeldem Input_ID -> labelgop Input_ID -> labeldjt



```
== Physical Plan ==
*(3) Project [title#106, id#536, pos#577, neg#578]
+- *(3) BroadcastHashJoin [replace(id#536, t3_, )], [id#69], Inner, BuildLeft
   :- BroadcastExchange HashedRelationBroadcastMode(List(replace(input[0, string, true], t3_, )))
   :  +- TakeOrderedAndProject(limit=10, orderBy=[neg#578 DESC NULLS LAST,pos#577 ASC NULLS FIRST], output=[id#536,pos#577,neg#578])
   :     +- *(2) HashAggregate(keys=[id#536], functions=[avg(cast(pos_label#542 as bigint)), avg(cast(neg_label#543 as bigint))])
   :        +- Exchange hashpartitioning(id#536, 200)
   :           +- *(1) HashAggregate(keys=[id#536], functions=[partial_avg(cast(pos_label#542 as bigint)), partial_avg(cast(neg_label#543 as bigint))])
   :              +- *(1) FileScan parquet [id#536,pos_label#542,neg_label#543] Batched: true, Format: Parquet, Location:
InMemoryFileIndex[file:/media/sf_vm-shared/predictions.parquet], PartitionFilters: [], PushedFilters: [], ReadSchema:
struct<id:string,pos_label:int,neg_label:int>
   +- *(3) Project [id#69, title#106]
      +- *(3) Filter isnotnull(id#69)
         +- *(3) FileScan parquet [id#69,title#106] Batched: true, Format: Parquet, Location: InMemoryFileIndex[file:/media/sf_vm-shared/submissions.parquet],
PartitionFilters: [], PushedFilters: [IsNotNull(id)], ReadSchema: struct<id:string,title:string>
```

Figure 1: 'Join Explain Output'

Question 2

The data in is not normalized as comments contains collumns for multiple attributes that are already in the submissions table such as archived, subreddit, subreddit_id, subreddit_type etc. It is possible that the author included these columns to do analysis on the comments independent of submission information, in which case you would not need the submissions file at all, saving some space.

Question 3

From our results in Figure 1, we see that spark is executing a broadcast hash join in order optimize join queries when the size of one side's data is below some threshold, in this case our top negative results, and then broadcasts it onto the larger table which is submissions. We see that spark reads submissions from our stored parquet. Spark also performs our aggregate functions (avg) used in calculating the top negative submissions using aggregate hashing. Alsoo we see project , where we select title, id, positive, and negative percentages.

Running explain on the query to join the top negative submissions with submissions on their id's yields:

context.sql("SELECT submissions.title, top_neg_submission.id, top_neg_submission.pos, top_neg_submission.neg FROM top_neg_submission JOIN submissions on replace(top_neg_submission.id, 't3_',")= submissions.id").explain()