CS 143 Project 2B Report

1.

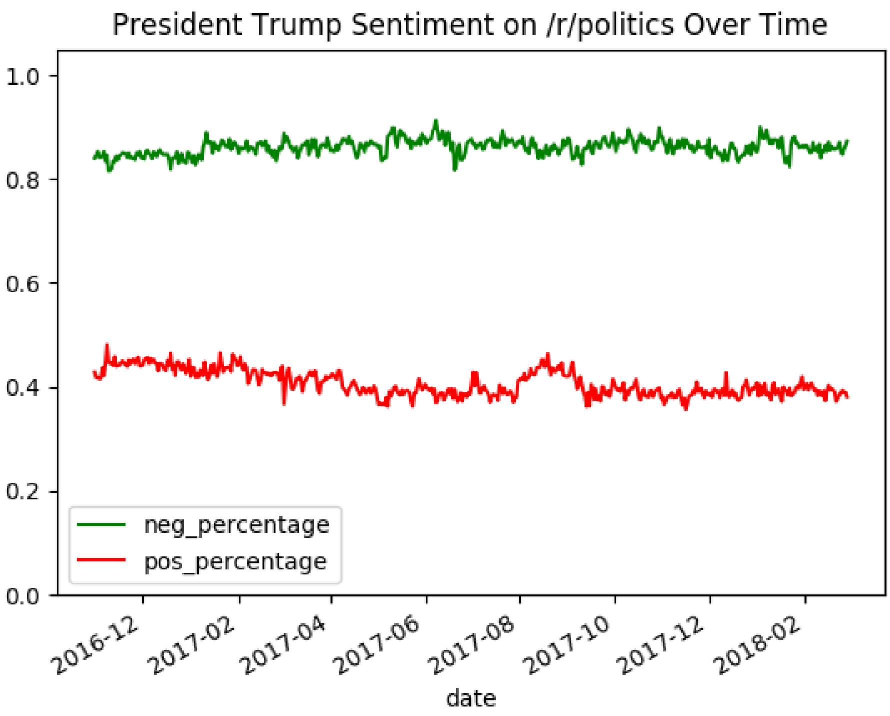


Figure 1

2.

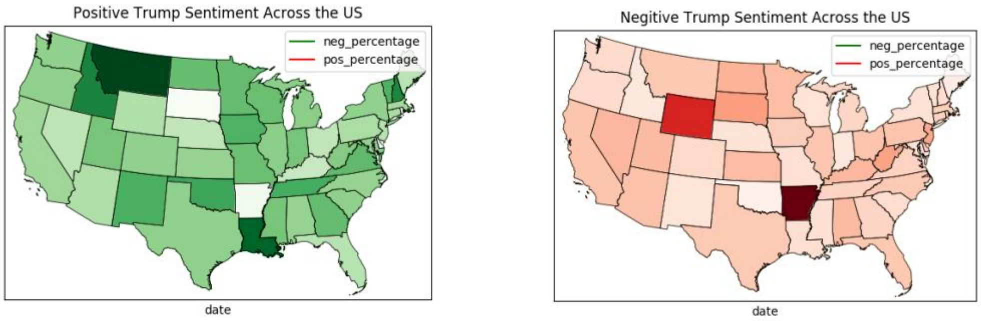

Positive Trump Sentiment Across the US
neg_percentage
pos_percentage
date
Negitive Trump Sentiment Across the US
neg_percentage
pos_percentage
date

Figure 2

3.

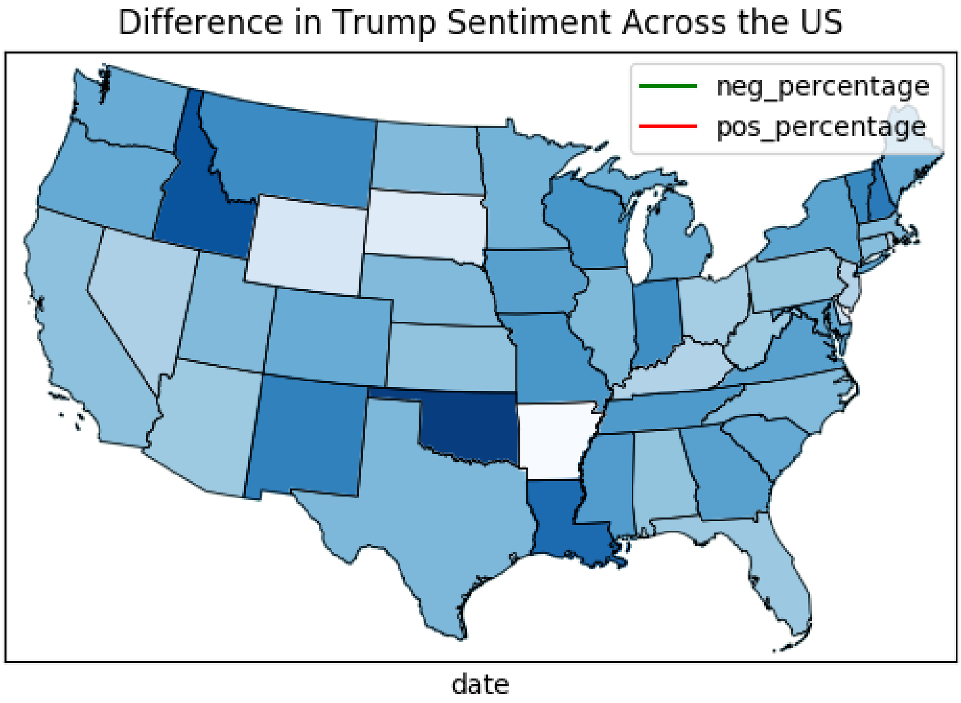


Figure 3

4.

Top 10 Positive Stories	Top 10 Negative Stories
Alt-right leader expelled from CPAC after organizer denounces ‘left-wing fascist anti-semtic group’	"\"You Don't Do Evidence Well I Do\" Trey Gowdy Demands Answers On Trump Russia Collusion"
Beyonce To Join Hillary Clinton on Campaign Trail in Ohio	'Merry Christmas' for Trump is more than a wish
American firms hoard trillions offshore led by Microsoft	"'Come Here, Katy': How Donald Trump Turned Me Into a Target"
"\"Intel and law enforcement officials agree that none of the investigations have found any conclusive or direct link between Trump and the Russian government period,\" the senior official said."	'I will be proven right': Trump tells golfing buddy he is confident his claims about Obama ordering a wiretap will turn out to be accurate

Andrew Wakefield calls Trump “on our side” over vaccines after meeting	"'Dear God, America what have you done?': How the world and its media reacted as Donald Trump poised to become US president"
"\"Suddenly Trump's losses become Putin's loss...\""	"\"Intel and law enforcement officials agree that none of the investigations have found any conclusive or direct link between Trump and the Russian government period,\" the senior official said."
AntiFaâ□□s Moral Superiority and the Potential for Left-Wing Unity	"\"Railway Walo Aaj Raat Neend Nahi Aayegi\" Indian Prime Minister Now Move on to Indian Railway"
"3 Doors Down has risen from the dead to play at the Trump inauguration, and Twitter is having a field day."	"\"Suddenly Trump's losses become Putin's loss...\""
Are pads and tampons taxed but Viagra and Rogaine not?	"\"TRUMP is a DISGUSTING, SHAMEFUL MANBABY!\" CNN's Ana Navarro ANGRY Reaction To Trump's TWEETS"
"A comparison between /The_D and this subreddit about the \"Hundreds of thousands rally in Iran against Trump, chant 'Death to America': TV\" article. There couldn't have two more opposing opinions about this."	'F*ck Donald Trump! F*ck White People!’: 4 in Custody for Torturing Young Man Live on Facebook

5.

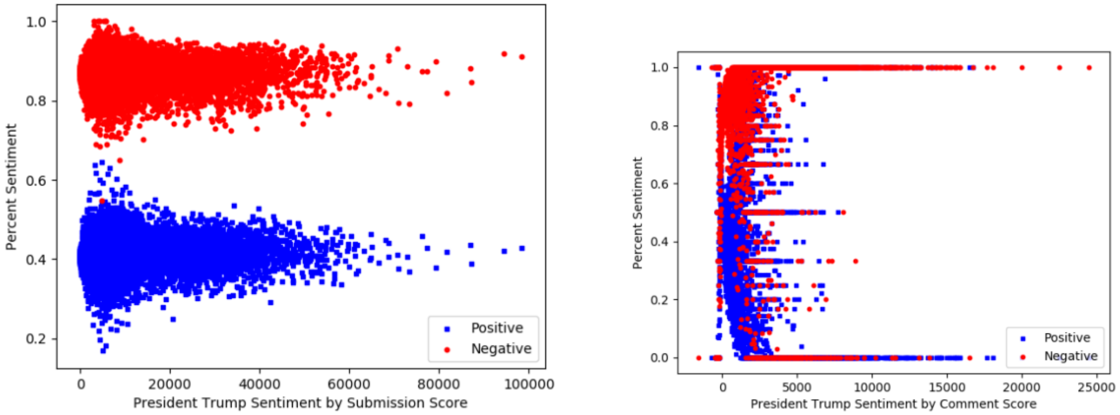


Figure 4

Findings

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

We could see from Figure 1 that there are more negative sentiments against President Trump on /r/politics than positive sentiments, in general not fluctuating greatly over time. From Figure 2 and Figure 3 we could see that the sentiments vary by states. A common trend is that the states that have deeper color (a higher probability of intense sentiment) on one side will have lighter color on the other. As we could observe from Figure 3, the deeper the color is, the greater the difference between the sentiment probabilities for /r/politics opinion on the President for that state. From Figure 1 we could see that the sentiments vary by dates with small fluctuations, but the trend does not change much. From Figure 4 we could conclude that submission score should be a better classifier than comment score because the latter has more overlaps between the two categories of sentiments.

<u>Questions</u>

Question 1:

The functional dependencies are as follows:

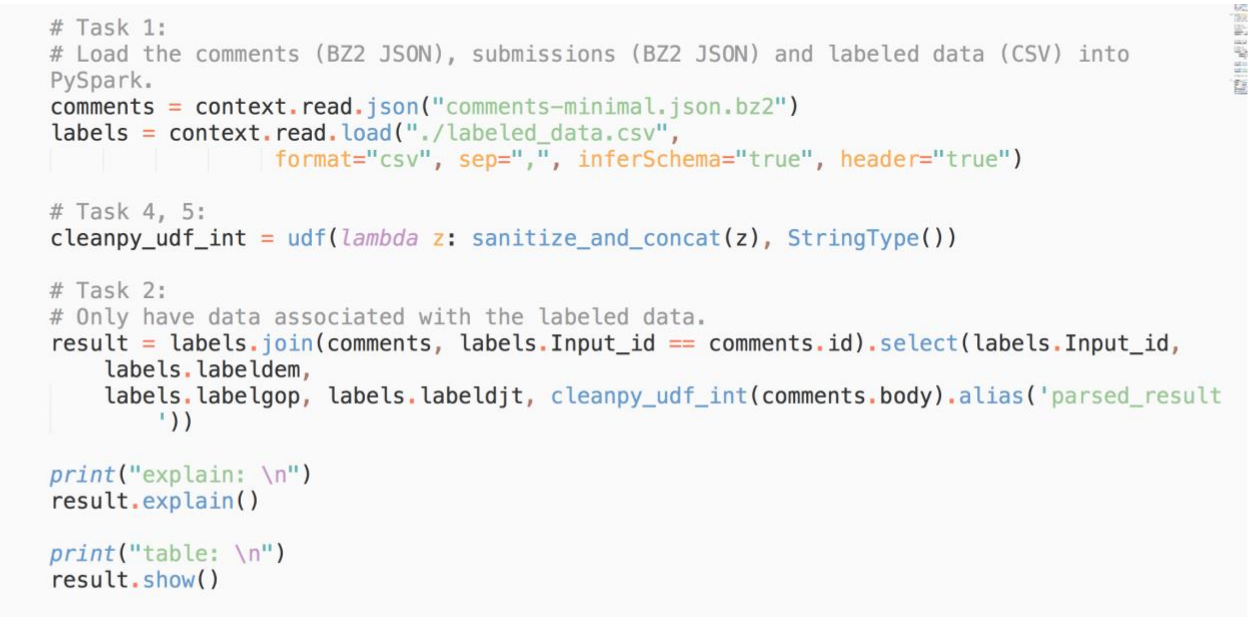
Input_id->labeldem, Input_id->labelgop, Input_id->labeldjt

Question 2:

The schema does not look normalized as there are many fieds that hold the same value (i.e subreddit, gilded, can_gild, just to name a few). The way to resolve this is to split the comments into seperate tables instead of have one giant table. As for why the data was **originally stored this** way, it is likely the output of the webscraper used to get this data and provided a visually simple standard for storing the data as well as making it very simple to limit the comment data set to a certain size.

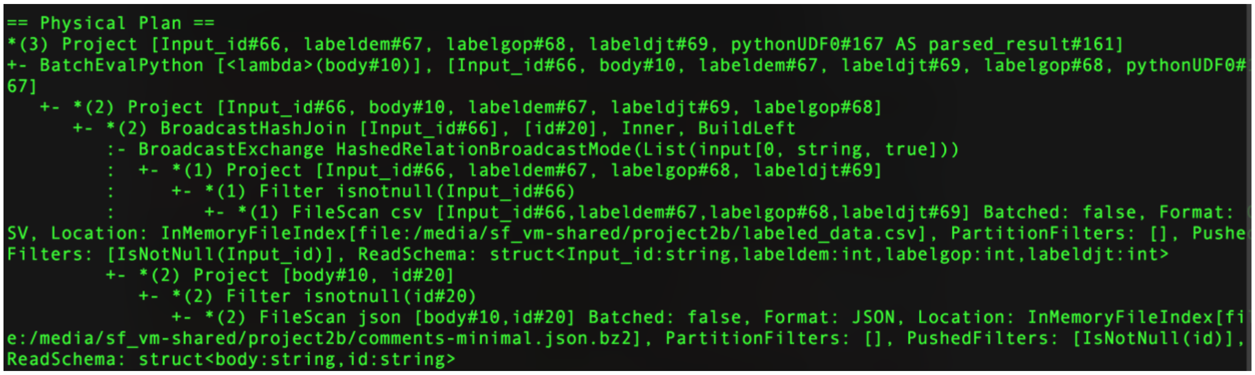
QUESTION 3:

code:



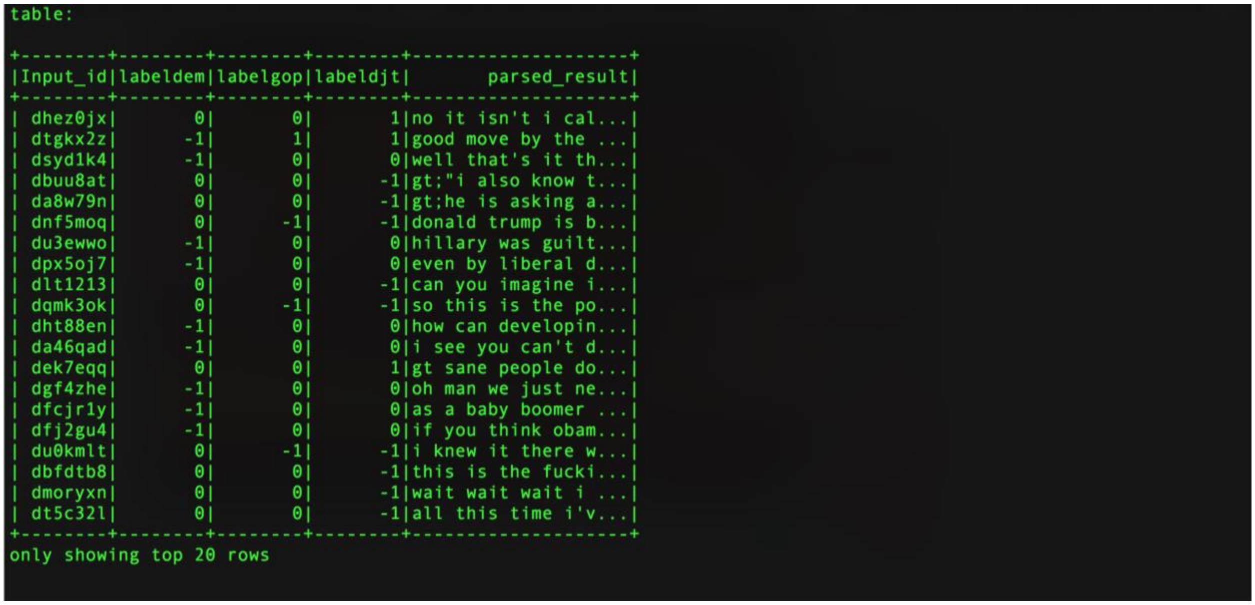
```
# Task 1:
# Load the comments (BZ2 JSON), submissions (BZ2 JSON) and labeled data (CSV) into
PySpark.
comments = context.read.json("comments-minimal.json.bz2")
labels = context.read.load("./labeled_data.csv",
                           format="csv", sep=",", inferSchema="true", header="true")

# Task 4, 5:
cleanpy_udf_int = udf(lambda z: sanitize_and_concat(z), StringType())

# Task 2:
# Only have data associated with the labeled data.
result = labels.join(comments, labels.Input_id == comments.id).select(labels.Input_id,
    labels.labeldem,
    labels.labelgop, labels.labeldjt, cleanpy_udf_int(comments.body).alias('parsed_result
        '))

print("explain: \n")
result.explain()

print("table: \n")
result.show()
```

result of running .explain():



```
== Physical Plan ==
*(3) Project [Input_id#66, labeldem#67, labelgop#68, labeldjt#69, pythonUDF0#167 AS parsed_result#161]
+- BatchEvalPython [<lambda>(body#10)], [Input_id#66, body#10, labeldem#67, labeldjt#69, labelgop#68, pythonUDF0#
67]
   +- *(2) Project [Input_id#66, body#10, labeldem#67, labeldjt#69, labelgop#68]
      +- *(2) BroadcastHashJoin [Input_id#66], [id#20], Inner, BuildLeft
         :- BroadcastExchange HashedRelationBroadcastMode(List(input[0, string, true]))
         :  +- *(1) Project [Input_id#66, labeldem#67, labelgop#68, labeldjt#69]
         :     +- *(1) Filter isnotnull(Input_id#66)
         :        +- *(1) FileScan csv [Input_id#66,labeldem#67,labelgop#68,labeldjt#69] Batched: false, Format: 
SV, Location: InMemoryFileIndex[file:/media/sf_vm-shared/project2b/labeled_data.csv], PartitionFilters: [], Pushe
Filters: [IsNotNull(Input_id)], ReadSchema: struct<Input_id:string,labeldem:int,labelgop:int,labeldjt:int>
         +- *(2) Project [body#10, id#20]
            +- *(2) Filter isnotnull(id#20)
               +- *(2) FileScan json [body#10,id#20] Batched: false, Format: JSON, Location: InMemoryFileIndex[fi
e:/media/sf_vm-shared/project2b/comments-minimal.json.bz2], PartitionFilters: [], PushedFilters: [IsNotNull(id)],
ReadSchema: struct<body:string,id:string>
```

resulting table:



```
table:

+--------+--------+--------+--------+--------------------+
|Input_id|labeldem|labelgop|labeldjt|       parsed_result|
+--------+--------+--------+--------+--------------------+
| dhez0jx|       0|       0|       1|no it isn't i cal...|
| dtgkx2z|      -1|       1|       1|good move by the ...|
| dsyd1k4|      -1|       0|       0|well that's it th...|
| dbuu8at|       0|       0|      -1|gt;"i also know t...|
| da8w79n|       0|       0|      -1|gt;he is asking a...|
| dnf5moq|       0|      -1|      -1|donald trump is b...|
| du3ewwo|      -1|       0|       0|hillary was guilt...|
| dpx5oj7|      -1|       0|       0|even by liberal d...|
| dlt1213|       0|       0|      -1|can you imagine i...|
| dqmk3ok|       0|      -1|      -1|so this is the po...|
| dht88en|      -1|       0|       0|how can developin...|
| da46qad|      -1|       0|       0|i see you can't d...|
| dek7eqq|       0|       0|       1|gt sane people do...|
| dgf4zhe|      -1|       0|       0|oh man we just ne...|
| dfcjr1y|      -1|       0|       0|as a baby boomer ...|
| dfj2gu4|      -1|       0|       0|if you think obam...|
| du0kmlt|       0|      -1|      -1|i knew it there w...|
| dbfdtb8|       0|       0|      -1|this is the fucki...|
| dmoryxn|       0|       0|      -1|wait wait wait i ...|
| dt5c32l|       0|       0|      -1|all this time i'v...|
+--------+--------+--------+--------+--------------------+
only showing top 20 rows
```

From the output of explain we can see that the default inner join is separated into several steps. First, it loads the left table, filters out the rows which do not have the attribute we wish to join the tables on, and selects the columns we specified. Then it repeats the same procedure on the right table. We could see that it saves time and memory by filtering the non-valid rows and useless columns before join. Then it performs a hash-join on the join keys and select the specified columns again. In the end it just calls the UDF function on the resulting table which is not related to the join operation.