Agustin Marinovic
Michael Kliger
Robin Zhang

Sentiment Analysis of r/politics

On this project we performed sentiment analysis on comments collected from Reddit's r/politics from November 2016 to February 2018. The aim was to see the sentiment distribution (positive and negative) towards the two main political parties and the President. The labeled data (obtained from Ryan R. Rosario) consists of 2000 comments tagged with +1 (positive), -1 (negative), 0 (neutral), or -99 (does not apply) for Democrats, Republicans and President Donald J. Trump. This labeled data was used to train a logistic regression model on a 5-fold cross validation.

With the trained model, all the comments towards the parties mentioned above were tagged and we performed further analysis on it.

We successfully classified the sentiment towards Republicans and President Trump; however we failed with Democrats. Almost all positive sentiments towards Democrats were labeled 0 (no positive) and almost all negative sentiments were labeled 1 (all negative). The plots and results will STILL be displayed. This is a disclaimer that their data is useless.

1. **Time series plot (by day) of positive and negative sentiment.**

The following plots show daily sentiment towards Donald J. Trump, Republicans, and Democrats. The following two plots were created:

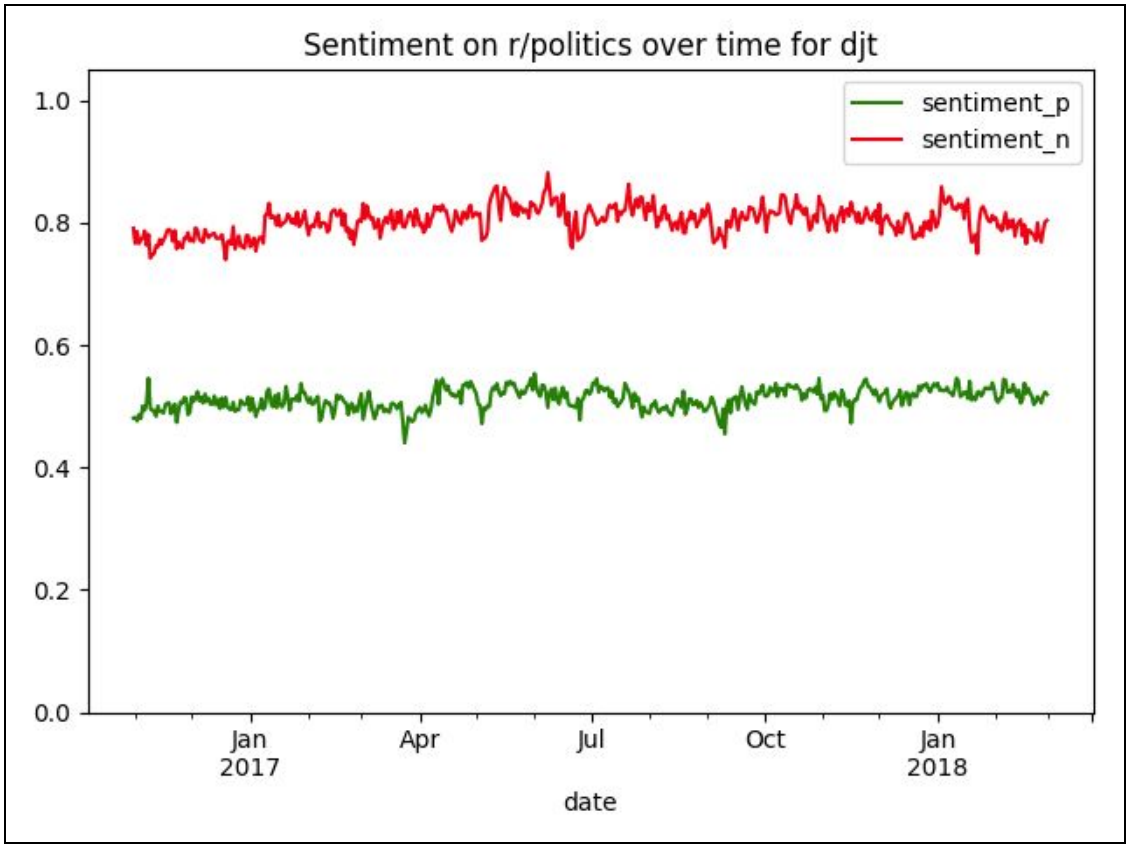

Figure 1 - Time series plot (by day) of positive and negative sentiment towards Donald J. Trump

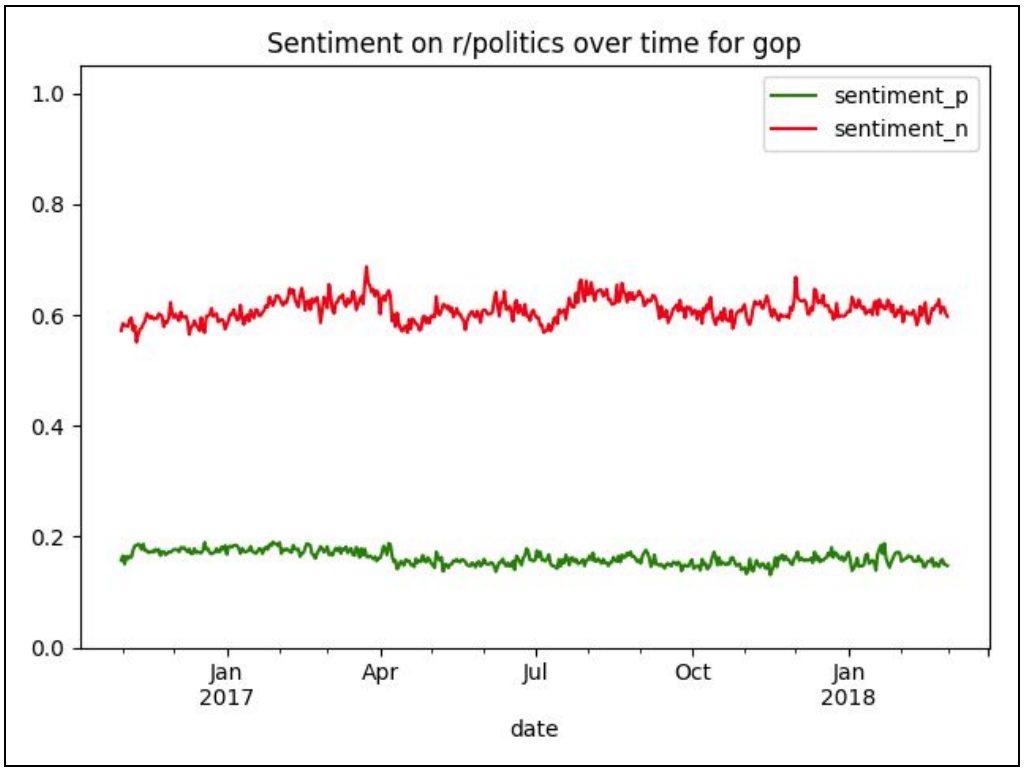

Figure 2 - Time series plot (by day) of positive and negative sentiment towards Republicans

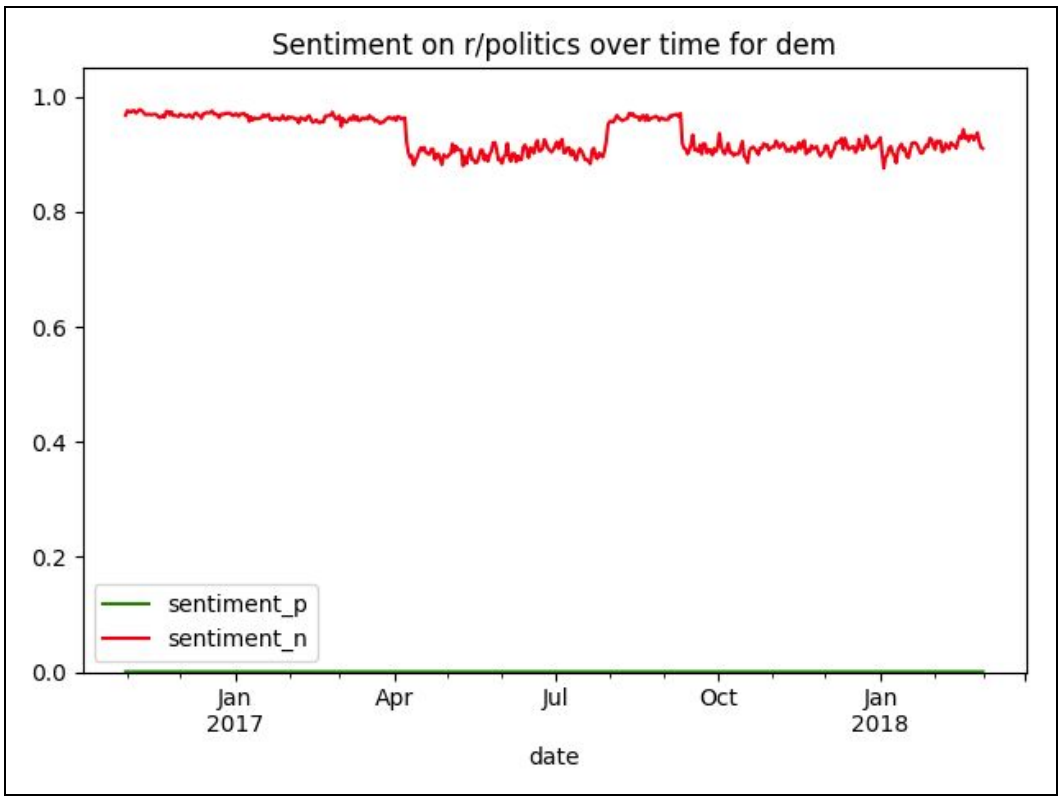

Figure 3 - Time series plot (by day) of positive and negative sentiment towards Democrats.

2. Sentiment (by percentage) towards each group, by state

The following plots are a map of sentiment across the United States by state. The states are determined by self-reported state via the user flair. Darker shades correspond to stronger sentiments.

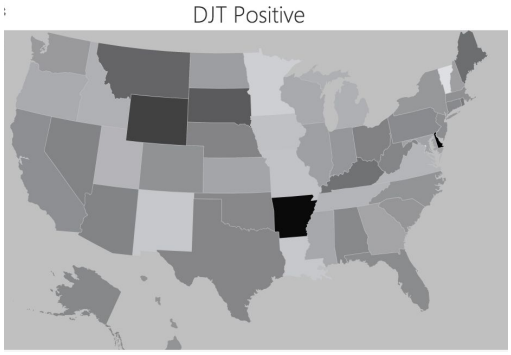

Figure 4 - Positive sentiment towards Donald J. Trump by state

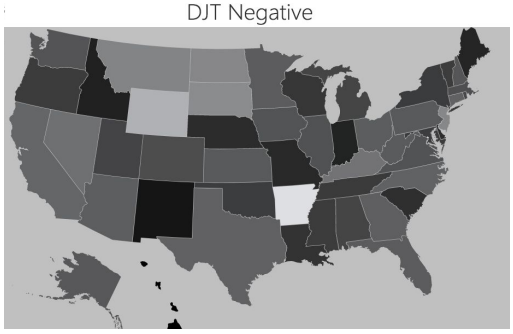

Figure 5 - Negative sentiment towards Donald J. Trump by state

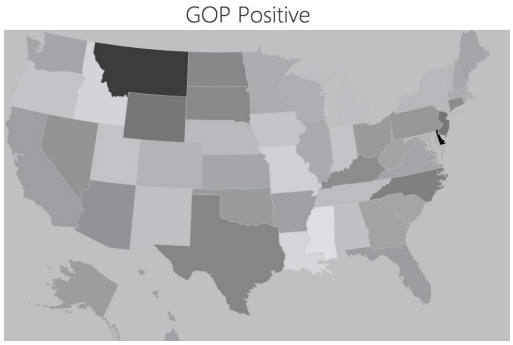

Figure 6 - Positive sentiment towards Republicans by state

Agustin Marinovic
Michael Kliger
Robin Zhang

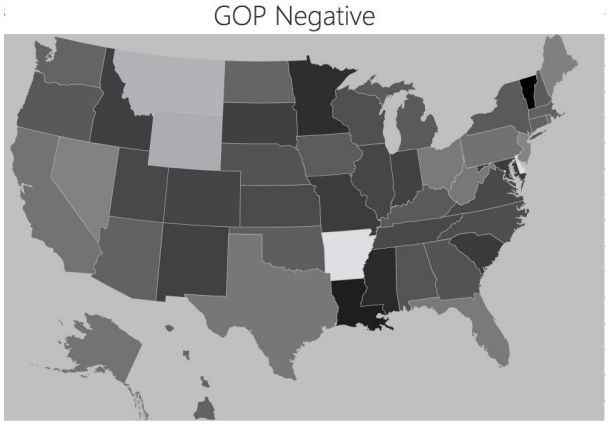

Figure 7 - Negative sentiment towards Republicans by state

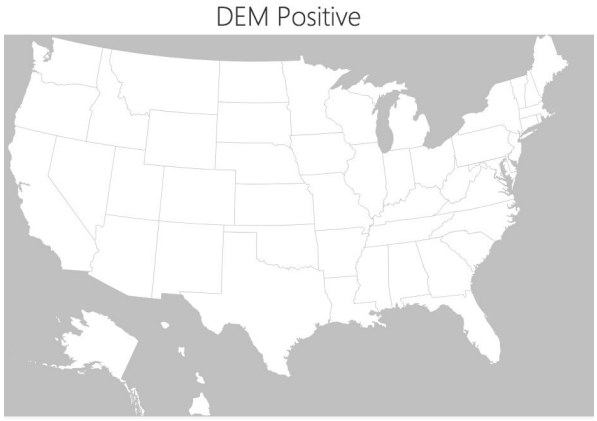

Figure 8 - Positive sentiment towards Democrats by state

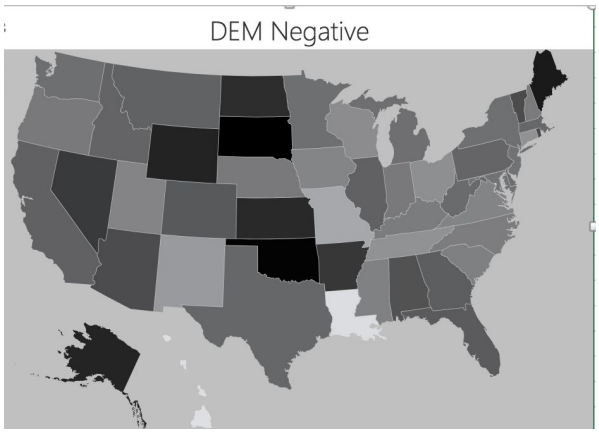

Figure 9 - Negative sentiment towards Democrats by state

Agustin Marinovic
Michael Kliger
Robin Zhang

3. Sentiment difference by state

The following plots show sentiment difference across the United States for the different entities. Darker colors represent a greater difference between positive and negative sentiments. Lighter colors represent similar positive and negative sentiments. The equation used for this is ABS(Pos-Neg).

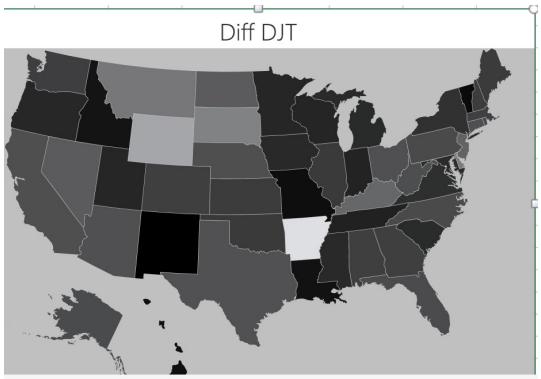

Figure 10 - Difference of sentiment towards Donald J. Trump

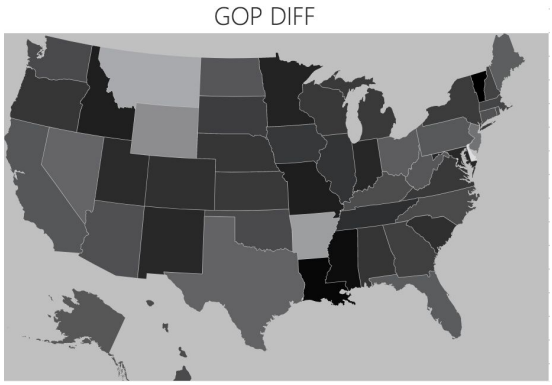

Figure 11 - Difference of sentiment towards Republicans

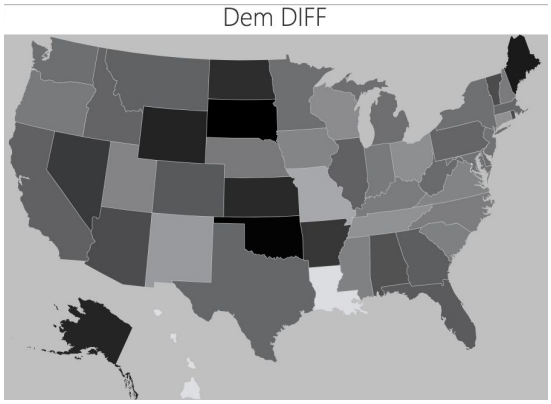

Figure 12 - Difference of sentiment towards Democrats

4. Top 10 Positive and Top 10 Negative Stories

The following tables represent the top 10 stories given positiveness and negativeness. Upon producing it, we added an extra column which gets top 10 submission with at least 10 comments.

Top Story (DJT +)	Score	Top Story (min 10 comments)	Score
Donald Trump's in-laws may have benefited from 'chain migration' process he rails against	1	Recount unlikely to change Alabama vote	1
Manafort offered to give Russian billionaire â€˜private briefingsâ€™ on 2016 campaign	1	In Trumpâ€™s America, The Statue Of Liberty Is Closed While Immigrants Are Pushed Out	1
Trump in 'excellent health,' White House doctor says after exam	1	Donald Trump just defended his response to Charlottesville. Again.	1
Is this it for Rex Tillerson	1	Trump Set to Abandon Trans-Pacific Partnership, Obama's Signature Trade Deal	1
UCLA basketball players thank Trump, apologize for actions in China	1	Trump says US culture being 'ripped apart' by Confederate memorial removals	1
GOP's Amash calls for independent commission on Russia after Comeyâ€™s firing	1	Roy Mooreâ€™s values could take Alabama back to a place many of its residents have tried to get past	1
Secret Service costs for President Donald Trumpâ€™s family mounts	1	A massive impending Bay Area ICE raid is apparently intended to send a signal to sanctuary cities	1
Mueller Seeks White House Documents Related to Trumpâ€™s Actions as President	1	Report on election hacking says Russia plans to do more	1
Trumpâ€™s Shutdown Retreat Relieves GOP, But Spending Bill Remains Secret	1	Schumer blames Trump: â€˜Great deal-making president sat on the sidelinesâ€™	1
Trump Administration: Puerto Rico Has Enough Money and Doesn't Need More Aid	1	â€˜No more!â€™ Trump tweets to Pakistan, accusing it of â€˜lies & deceitâ€™	1

Figure 13 - Top 10 positive stories towards Donald J. Trump

Top Story (DJT -)	Score	Top Story (min 10 comments)	Score
If China wonâ€™t get tough on North Korea, Trump should get tough on China	1	State Department offers buyouts as critics charge Trump administration with destroying diplomatic corps	1
Monuments of White Supremacy	1	"The Administration steadfastly opposes legalization of marijuana and other drugs because legalization would increase the availability and use of illicit drugs, and pose significant health and safety risks to all Americans, particularly young people."	1
The US and Russia may be getting closer to a military confrontation	1	Schumer: Trump should blame himself for vacancies	1
An affair with Trump? Nikki Haley on 'disgusting' rumors and her rise to a top foreign policy role	1	No reason to believe Trump is target of any investigation: White House	1
Barry Manilow for president? Singer pledges to 'make America happy again'	1	Sessions Press Conference on CSPAN live	1
White House Announces President Donald Trump has Fired FBI Director James Comey	1	Trump blasts 9th Circuit for ruling against travel ban 'at such a dangerous time'	1
Hillary Clinton uses obscenity on TV describing reaction to Trump's inaugural speech	1	George W. Bush Bashing Trump's 'Cruelty' Is Hypocritical Bullshit	1
Netanyahu: Israel is a country that says 'Merry Christmas'	1	The Trump Doctrine Falls Flat	1
Retired generals cite past comments from Mattis while opposing Trumpâ€™s proposed foreign aid cuts	1	Kimmel: 'We're one Trump toilet tweet away from being the United States of Florida'	1
Trump Just Showed That He Is Terrified Of Facing Elizabeth Warren In 2020	1	Trump to donate $1 million to Texas recovery	1

Figure 14 - Top 10 negative stories towards Donald J. Trump

Top Story (GOP +)	Score	Top Story (min 10 comments)	Score
Trump Removes Anthony Scaramucci From Communications Director Role	1	Donald Trumpâ€™s Pollster Says the Election Came Down to Five Counties	1
Why 9/11 Victims' Families Are Upset With Donald Trump's Refugee Ban	1	Trump to sign executive order undoing Obama's clean power plan	1
Lieutenant Governor Parson calls for resignation of Senator Chappelle-Nadal over Trump-Assassination post	1	Most Democrats would consider voting for third-party presidential candidate in 2020: Poll	1
Republican activists to monitor Election Day polls	1	We Have a President Who Wonâ€™t Defend Our Nation	1
VP Pence casts historic tie-breaking vote to confirm Betsy DeVos as Education Secretary	1	Trump Must Be Respected as U.S. President, Says Germany's Merkel	1
Justice Department inspector general to investigate pre-election actions by department and FBI	1	U.S. intel report identifies Russians who gave emails to WikiLeaks -officials	1
Foreign Payments to Trump Firms Violate Constitution, Suit Will Claim	1	Debbie Wasserman Schultz: â€˜We are really being priced out of the ability to afford drugsâ€™	1
D.C. Womenâ€™s March Organizer, Linda Sarsour, Flashes The ISIS Single Fingered Salute?	1	Trump Condemns Obama for Not Intervening in Syria after Begging Obama not to Intervene in Syria	1
Hereâ€™s the statement Sean Spicer should give	1	Pelosi throws up a giant Yield sign on single-payer	1
â€˜A Day Without a Womanâ€™: The Womenâ€™s Global Strike and the Growing Movement Against Donald Trump	1	100 days in, Howard Stern was 100% right about Donald Trump	1

Figure 15 - Top 10 positive stories towards Republicans

Top Story (GOP -)	Score	Top Story (min 10 comments)	Score
Sean Spicer says US jobs figures have been 'phoney' in the past but are 'real' under Donald Trump	1	Sen. Lindsey Graham votes for Evan McMullin over Trump, Clinton	1
Behind the Quiet State-by-State Fight Over Electric Vehicles	1	Fascist Milwaukee Sheriff David Clarke Resigns	1
Philippine President Duterte: I can be friend to Trump	1	Secret Service director to step down, giving Trump chance to select his own security chief	1
San Bernandino shooter was NOT fully investigated before being granted a U.S. visa, official claims	1	GOP senator: ObamaCare repeal bill may not have votes to pass	1
Trump taps S.C. Gov. Nikki Haley as U.N. ambassador	1	Trump vows to â€˜dramatically reduce income taxesâ€™	1
Dems target Flake's seat amid GOP infighting	1	White nationalist â€˜Unite the Rightâ€™ organizer blames Charlottesville police, counter-protesters for sparking deadly violence	1
John Kasich: We're all rooting for Trump to 'get it together'	1	40 House Dems to urge Trump to suspend Flynn	1
Manafort switching legal team as feds crank up heat on him	1	At White House, Christie says no interest in joining Trump administration	1
How Bike Helmet Laws Do More Harm Than Good: They donâ€™t do much to improve safety, but theyâ€™re great at getting people to avoid cycling altogether	1	Gorka: 'Nonsensical' for Tillerson to discuss military matters	1
McCain, Flake oppose the end of DACA	1	Gay GOP group urges Trump to reinstate transgender protections.	1

Figure 16 - Top 10 negative stories towards Republicans

Top Story (DEM +)	Score	Top Story (min 10 comments)	Score
The US is turning away Mexican avocados at the border	1	Texas anti-masturbation bill moves closer to becoming law	1
Trump admin reportedly will allow 872 refugees into US despite travel ban	1	The US is turning away Mexican avocados at the border	1
Trump tries to deepen Dem divisions	1	Trump admin reportedly will allow 872 refugees into US despite travel ban	1
Fake news, Russia and Comey: all poor answers to why Donald Trump won	1	Liberal leaders call for challenge to Gabbard over Syria skepticism	1
Alabamaâ€™s Senior Republican Senator: My State â€œDeserves Betterâ€ Than Roy Moore.	1	The End Of American Prison Visits: Jails End face-To-Face Contact â€“ And Families Suffer	1
Trump, next to empty chairs, tears into Dems for skipping meeting	1	Fake news, Russia and Comey: all poor answers to why Donald Trump won	1
How the Supreme Court could limit gerrymandering, explained with a simple diagram	1	Alabamaâ€™s Senior Republican Senator: My State â€œDeserves Betterâ€ Than Roy Moore.	1
Ex-WH ethics chief: Trump has created environment of â€˜serial ethics problemsâ€™	1	Trump, next to empty chairs, tears into Dems for skipping meeting	1
Chuck Cooper Confirms: He's AG Jeff Sessions' Lawyer	1	How the Supreme Court could limit gerrymandering, explained with a simple diagram	1
18 states sue Education Secretary DeVos for delaying student protection rules	1	Chuck Cooper Confirms: He's AG Jeff Sessions' Lawyer	1

Figure 17 - Top 10 positive stories towards Democrats

Top Story (DEM -)	Score	Top Story (min 10 comments)	Score
Trump orders EPA contract freeze and media blackout	1	'I'm With You': David Brock Pledges Allegiance To Sanders	1
LÃ¼genpresse. The Lying Media.	1	Trouble in Trumpland: The president's core supporters begin to worry	1
Senator suggests FBI has transcripts that might point to collusion between Trump and Russia	1	Poll: Trump leads Clinton by 7 points in battleground North Carolina	1
Dwindling Odds of Coincidence	1	Senator suggests FBI has transcripts that might point to collusion between Trump and Russia	1
Trump Breaks Down During Press Conference And Complains That All He Does Is Work	1	America has elected its own Berlusconi. Now itâ€™s about to repeat Italyâ€™s biggest mistake	1
Hillary Clinton blames Bernie Sanders and â€˜Bernie Brosâ€™ for campaign problems in excerpt from memoir â€˜What Happenedâ€™	1	Dwindling Odds of Coincidence	1
Can Democrats Turn Activism Into Votes? Special Elections Might Be A Clue	1	Trump Breaks Down During Press Conference And Complains That All He Does Is Work	1
Retired generals cite past comments from Mattis while opposing Trumpâ€™s proposed foreign aid cuts	1	Hillary Clinton blames Bernie Sanders and â€˜Bernie Brosâ€™ for campaign problems in excerpt from memoir â€˜What Happenedâ€™	1
Trump Just Showed That He Is Terrified Of Facing Elizabeth Warren In 2020	1	Fox News guest claims rainbow flag is as divisive as Confederate flag	1
Boom! Consumer confidence surges in March to highest level since December 2000	1	GOP Wants To Punish Filming, Photography On The House Floor.	1

Figure 18 - Top 10 negative stories towards Democrats

5. Submission score vs Sentiment, and Comment score vs Sentiment

The following graphs show the submission score and comment score against the sentiment for each party and Trump.

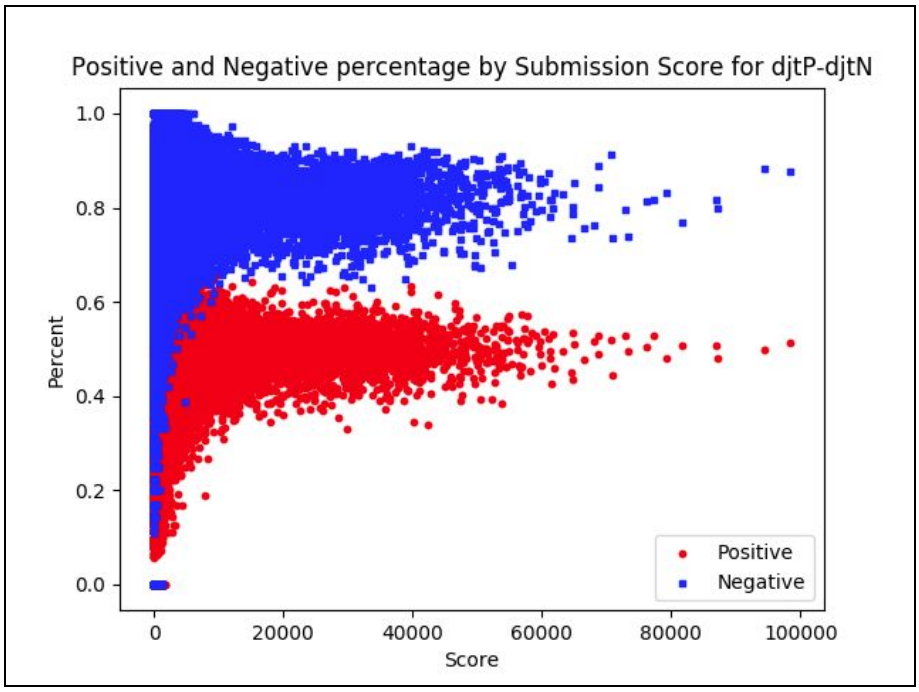

Figure 19 - Submission score vs Sentiment for Donald J. Trump

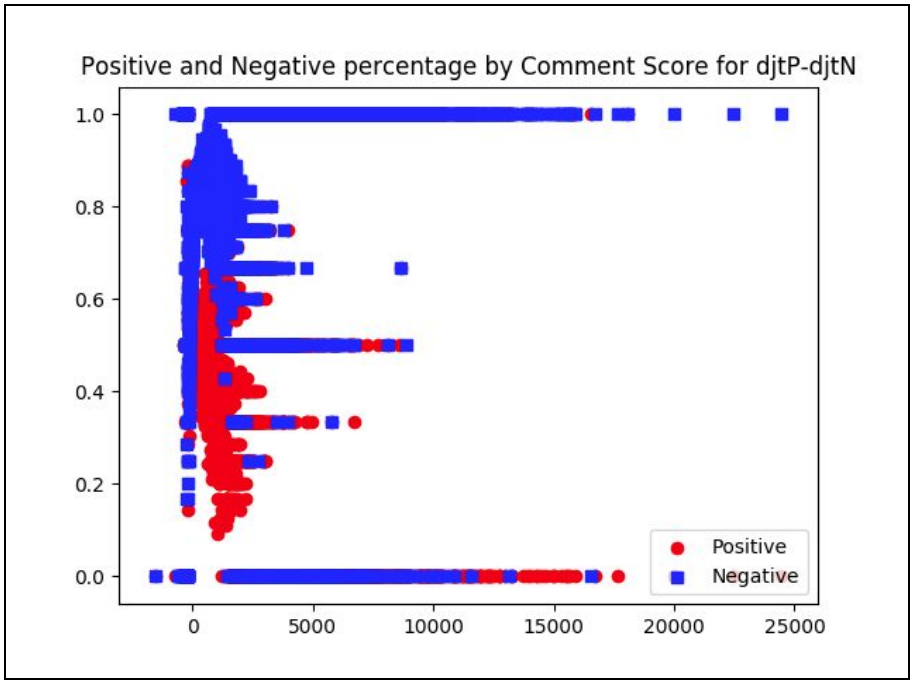

Figure 20 - Comment score vs Sentiment for Donald J. Trump

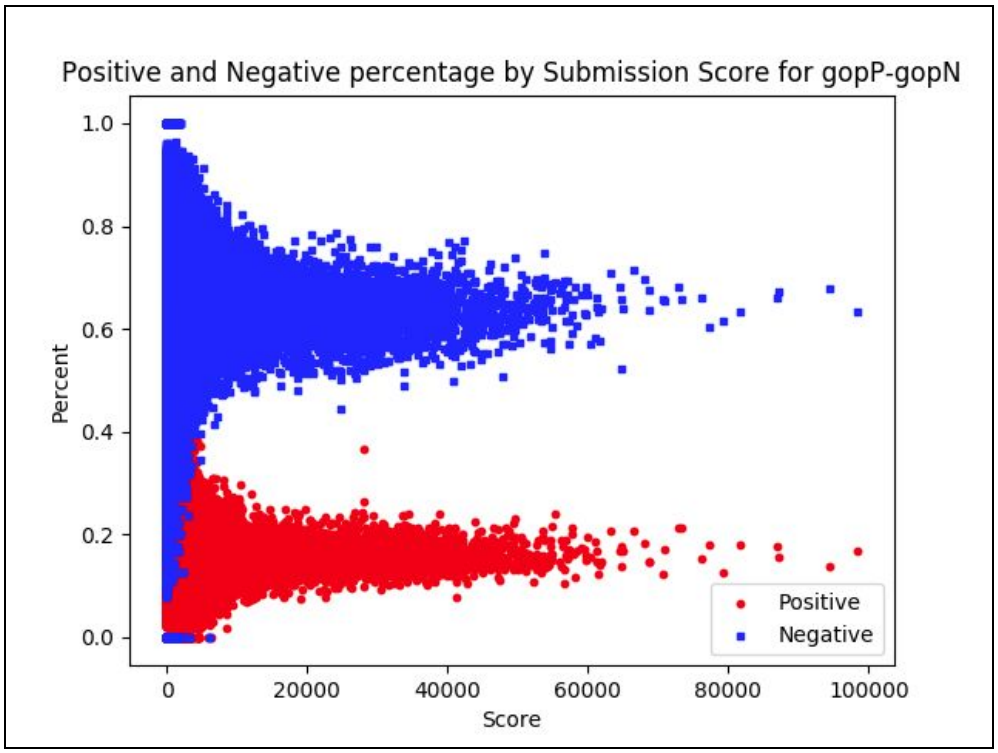

Figure 21 - Submission score vs Sentiment for Republicans

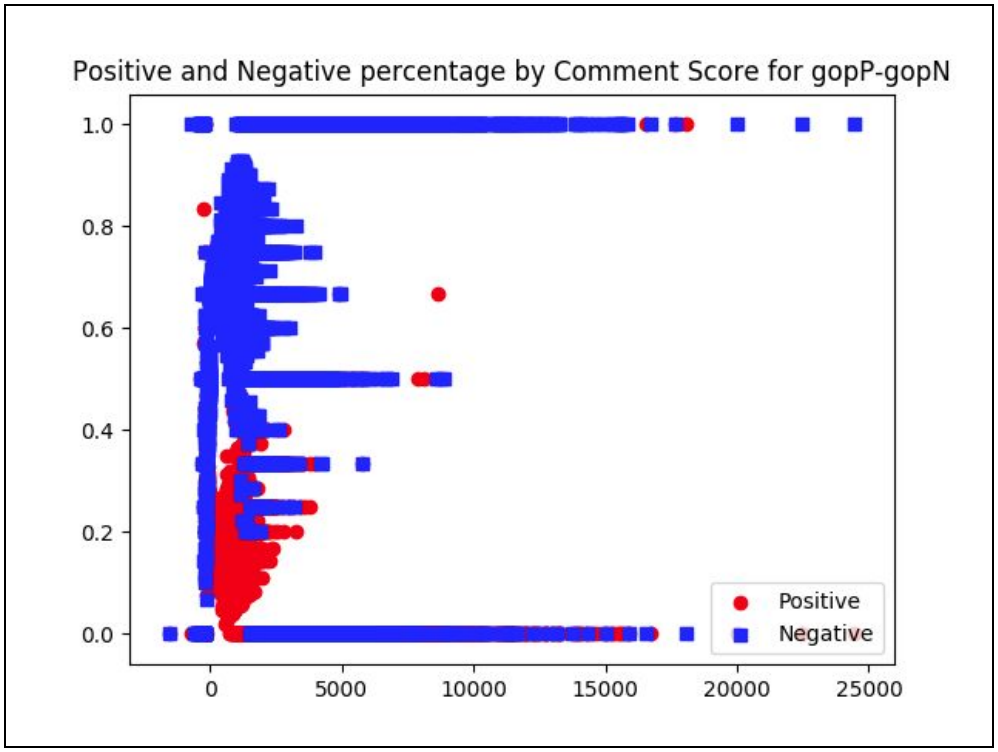

Figure 22 - Comment score vs Sentiment for Republicans

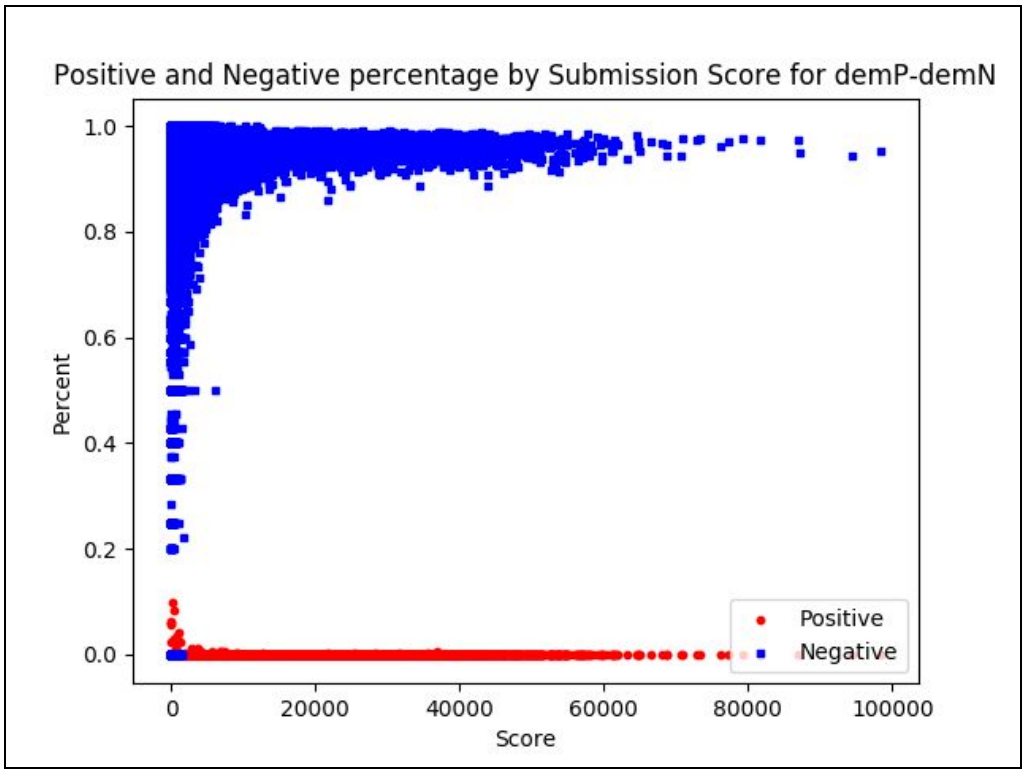

Figure 23 - Submission score vs Sentiment for Democrats

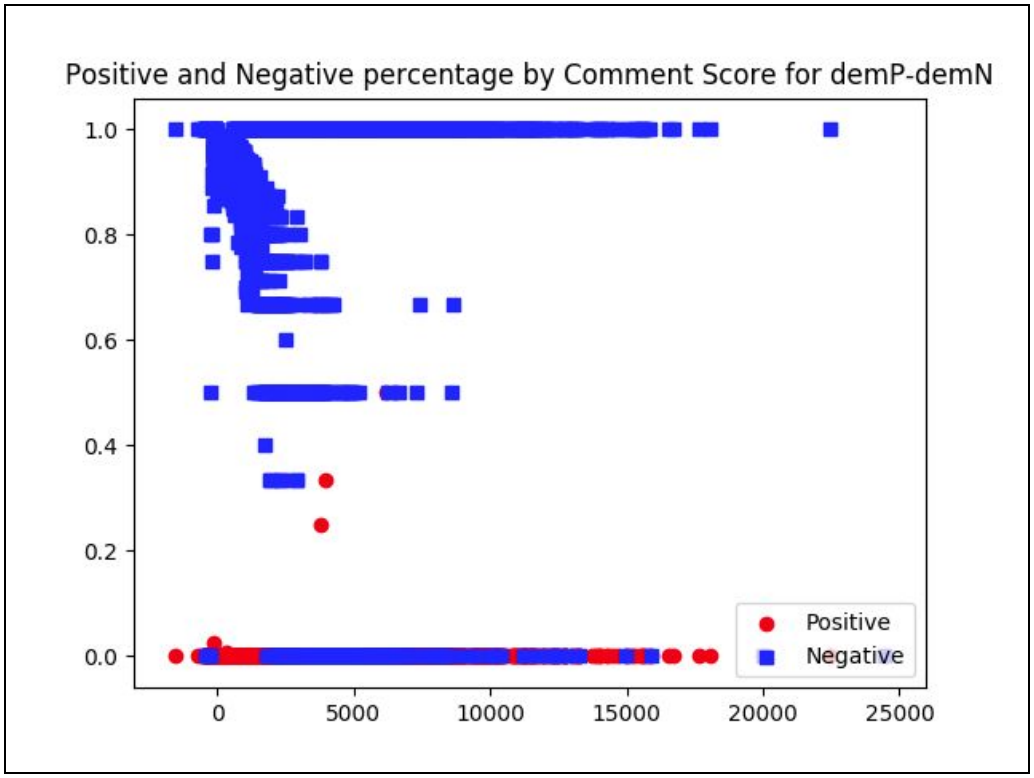

Figure 24 - Comment score vs Sentiment for Democrats

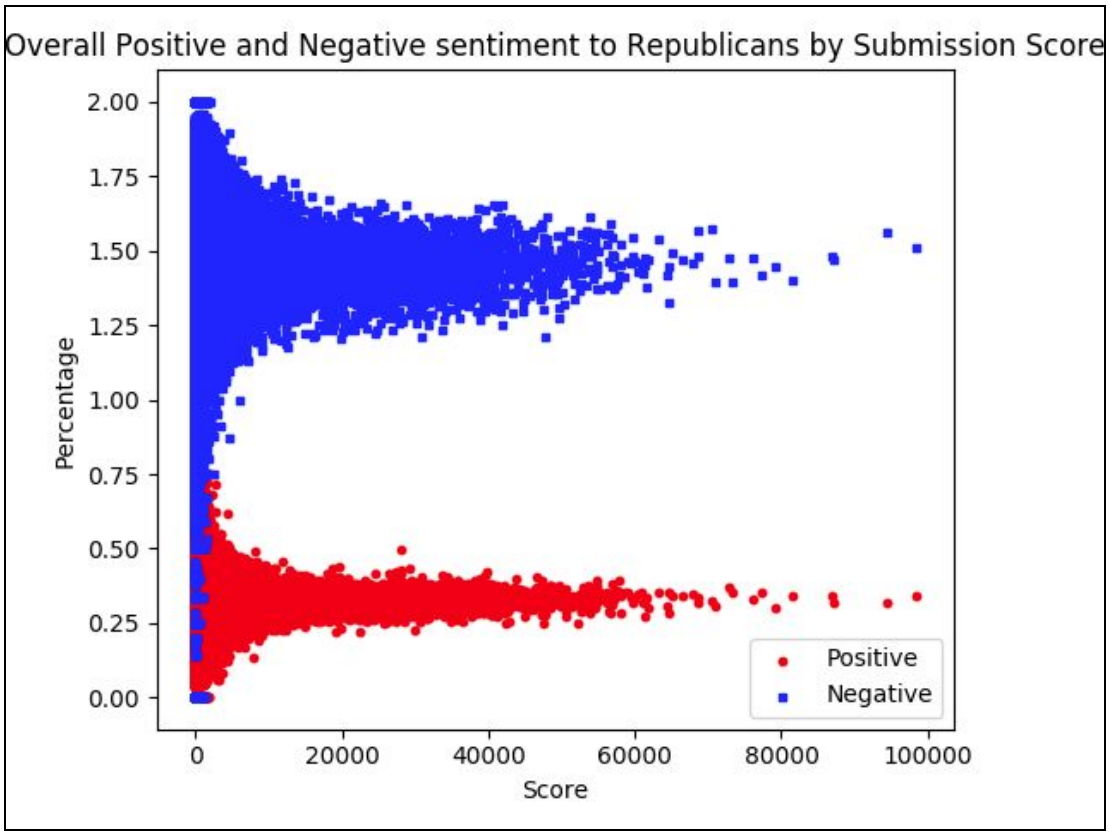

Figure 25: Average sentiment scatterplot of Republicans+Donald J. Trump vs Submission Score

6. Summary

We can see immediately that there is a general negative sentiment towards both Donald J. Trump and Republicans on Reddit's r/politics in Figures 1 through 3. The time plot of sentiment shows that both negative sentiment towards both are higher than the positive sentiment towards both. And for both graphs, the difference between positive and negative are both approximately 0.3 to 0.4. However, we can see a slight difference between the two on *how much* each sentiment holds for both the GOP and Trump. It appears that the negative sentiment towards Trump is much higher than the negative sentiment towards Republicans. Thus, we can also infer, and we can see, that the positive sentiment for Trump is higher than the positive sentiment for Republicans. We can also see that neither positive nor negative sentiment towards either Republicans or Trump change that much over time.

Our model prediction could not correctly evaluate the sentiment towards Democrats. As seen in the sentiment over time plot for Democrats, the positive sentiment is almost always zero, and negative sentiment nears 80% constantly. Even though that might be possible, these results are not close in any degree to our expectations.

We can see in the maps (Figures 4 through 9) a similar trend. These maps depict high levels of what we are showing as dark and low levels as light. For example, a state shaded completely dark for a positive sentiment graph has a high percentage of the users with the state's flair as being positive towards whichever group is being represented. The positive sentiment towards Republicans and Donald J. Trump are low, except for certain states, most notably

Montana and Arkansas. In fact, the negative sentiment towards these two are much greater than the positive sentiment that we see a darker overall map towards the GOP and Trump, whereas the positive sentiment is very light. Although the opposite of positive sentiment isn't negative sentiment (there is neutral sentiment), we are able to see an inverse in the positive sentiment in both maps depicted as the negative sentiment, and vice versa. The anomalies of Louisiana and Vermont being extremely supportive of Trump may be from a small sample of authors with those flairs.

We can see the difference in Figures 10 through 12. These differences are calculated using the negative sentiment minus the positive sentiment, because for all of our models, the negative sentiment was higher than the positive sentiment. With all of our data, only negative sentiment shows up for Democrats (and shows up as very high values), so our maps here may also be incorrect.

Figures 19 through 25 show the submission and comment score versus the sentiment. On the overall Republican scatterplot, sentiment was averaged for Republicans and Donald J. Trump. The plot shows that negative comments are more prevalent than positive comments for all submission scores. One thing to notice might be that for comments, the cluster of positive comments are higher for Trump than for the GOP, which is also related to the results we gathered in our time graphs in the beginning - the percentage of positive sentiment for Trump was greater than the positive sentiment for the GOP.

We can conclude that the general idea of more negative sentiment towards Trump and the GOP shows a bias against conservative thinking in r/politics. We can't say anything about the Democrats as there's no trustworthy data.

Also, on the submission and comment score scatter plots we can see the distribution of sentiment and score. The submission score is a very effective scatter plot which offers varied ranges of sentiment. The comment scatter plot does not offer much information. This is due to the binning process. A submission will have multiple comments which will return an average sentiment. For comments, the chance of overlap is small at high scores. Therefore the scatter plot shows mostly 0 or 1. We attempted binning the data in different ranges, but high scores are too sparse to obtain meaningful results from them. If we used a small range (100) high scores would not be affected. If we used a large range (1000) low scores (which have a lot more data points) would not correctly represent the variance of low scores.

QUESTIONS

1) We can infer each of the labels from the id. That is, id -> label for each label.
2) There are a few redundancies. For example, subreddit_id implies subreddit_name, while author(username) implies both can_gild and author_cakeday. Furthermore, author_flair_Text is implied by username and subreddit_id. Trivially, comment_id can identify any of these fields. A join across any number of these separated and normalized graphs would create too many tuples, meaning the original relation is redundant.

Agustin Marinovic
Michael Kliger
Robin Zhang

3) PySpark's .explain() function was used on a SQL query join. The join is done between comments and submission to get top stories. The result is the following:

```
== Physical Plan ==
*(4) Project [title#82, average#157]
+- *(4) BroadcastHashJoin [id#45], [link_id#5], Inner, BuildRight
  :- *(4) Project [id#45, title#82]
  +- BroadcastExchange HashedRelationBroadcastMode(List(input[0, string, false]))
    +- *(3) Filter isnotnull(link_id#5)
      +- TakeOrderedAndProject(limit=10, orderBy=[average#157 DESC NULLS LAST],
output=[link_id#5,average#157])
        +- *(2) HashAggregate(keys=[link_id#5], functions=[avg(demP_pred#7)])
          +- Exchange hashpartitioning(link_id#5, 200)
            +- *(1) HashAggregate(keys=[link_id#5], functions=[partial_avg(demP_pred#7)])
              +- *(1) FileScan parquet [link_id#5,demP_pred#7] Batched: true, Format: Parquet,
Location: InMemoryFileIndex[file:/Users/agustin/vm-shared/p2b/data/full_comments.parquet],
PartitionFilters: [], PushedFilters: [], ReadSchema: struct<link_id:string,demP_pred:double>
```

The planner does the following key steps:

1) Project both tables to retrieve needed columns
2) The columns (id and link_id) are hashed for the join
3) A Hash aggregation (average) is performed on the second table on column link_id
4) Data is retrieved from file

References:

- Pyspark's functions: http://spark.apache.org/docs/2.1.0/api/python/pyspark.html
- SQL usage in pyspark: https://spark.apache.org/docs/latest/sql-programming-guide.html
- PLT tight layout: https://matplotlib.org/users/tight_layout_guide.html