



r/MiningReddit

Roy De Jesus | Sarah Barili | Kathryn Gray | Jia Lin

Initial Questions



- What words were used most in r/politics?
- How did r/politics users comment in other subreddits?
- Who were the most influential users in the r/politics?
- What does the comment distribution through the month of January look like?
- What are user sentiment on specific political topics?

Tools



- JSON
 - Initial format of the data
 - Easy for machine to read
 - Can be parsed into another format like CSV
- Python
 - Code is easy to read and understand
 - Large library
- WEKA
 - Text mining of Single User and r/politics
 - Quick results once the format is learned
- CSV
 - This format has more order
 - Graphing and plotting is easier with this format

Data Preprocessing

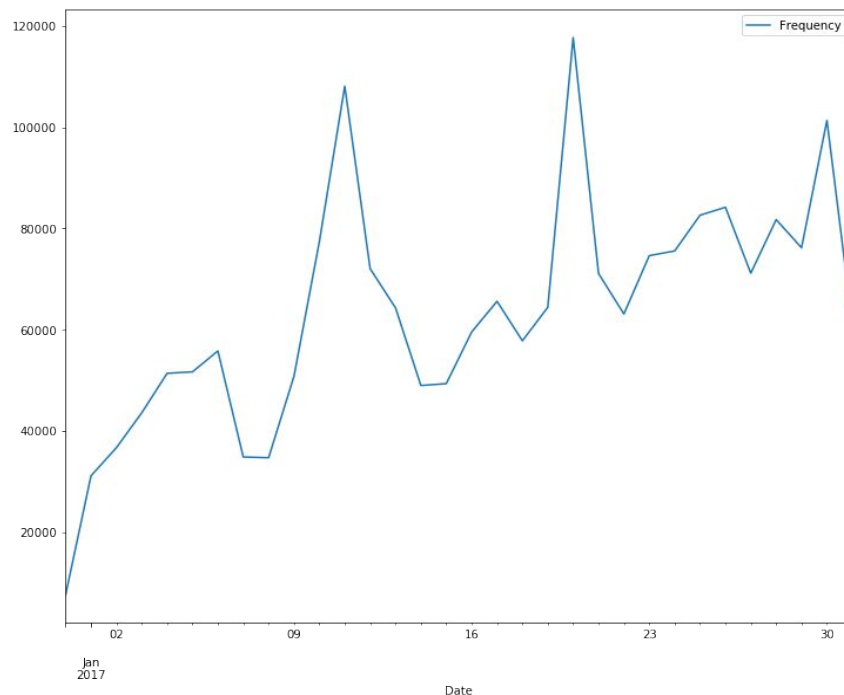


- Sampling
 - Our dataset was a massive 42GB
 - Sampling allowed us to test our source code on a manageable size
- Data Cleaning
 - Dates has to be cleaned
 - Non-ascii characters
 - Unix timestamps conversion
- Data Reduction
 - Reduced our large JSON file to multiple smaller JSON files

Results

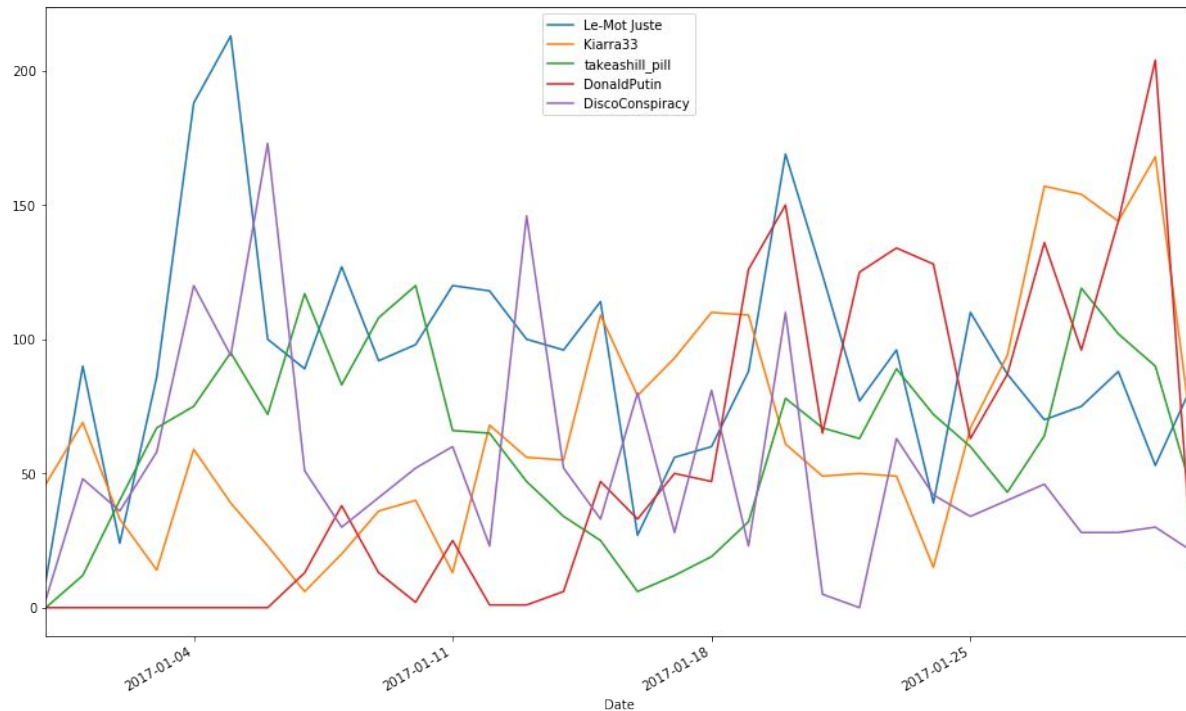
- Timeseries of the all the comments on Politics subreddit
 - January 11-12
 - Christopher Steele's Dossier
 - January 20
 - Inaguration

r/ Politics Comments Timeseries

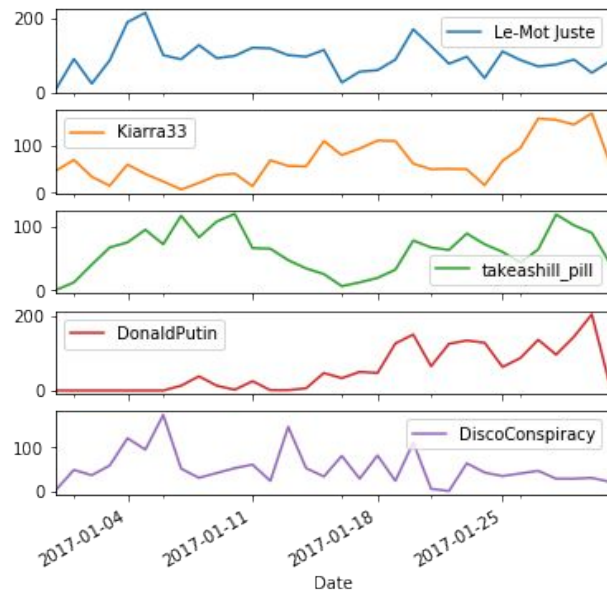


Behaviour of Top Influencers in r/ Politics

Timeseries of comments per user(s)



Time Series of comments per user



Findings and Applications



- Top influencers post times are unique compared to whole dataset
- Certain patterns were noticed for individual influencers
 - “Assange, Russia, News”, Correlation
 - “Russia, Russian, Election, Free”
 - Worried that the election was influenced by Russia
- Top influencers usually only have one highly score comment
 - Constantly changing influencers
- We can look at the subreddits influencers use and make assumptions
 - Kiarra33 as a Bernie supporter