

# Reddit: Analysis of “Suspicious” Accounts

STAT 418 Project Proposal

Janella Shu

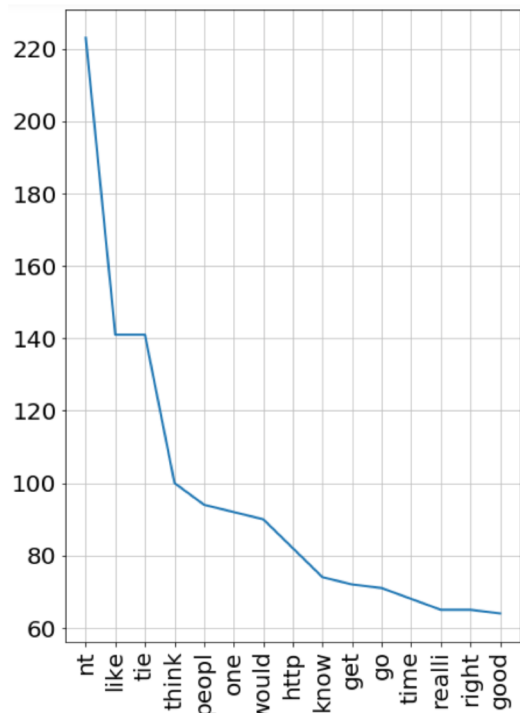
# Dataset

- Pushshift.io
  - Reddit API has a 1,000 item limit
  - Searched for submissions and comments made by 944 suspicious accounts ([https://www.reddit.com/r/announcements/comments/8bb85p/reddits\\_2017\\_transparency\\_report\\_and\\_suspect/](https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/))
- Comments
  - Variables: author, body, created\_utc, id, link\_id, parent\_id, subreddit, subreddit\_id
  - 1,620 comments (word count: 14,793)
- Submissions
  - Variables: author, created\_utc, domain, subreddit, subreddit\_id, title, url
  - 3,567 posts (word count: 17,713)

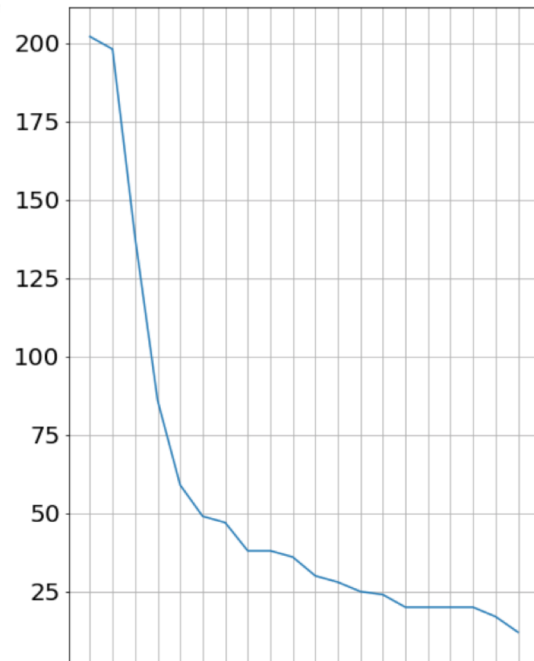
# Exploratory Data Analysis

Comments

Body

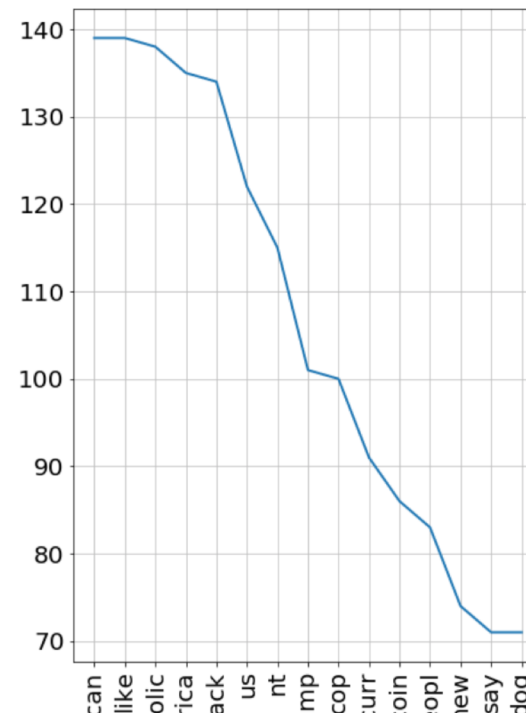


Subreddit

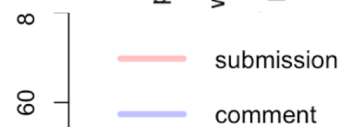
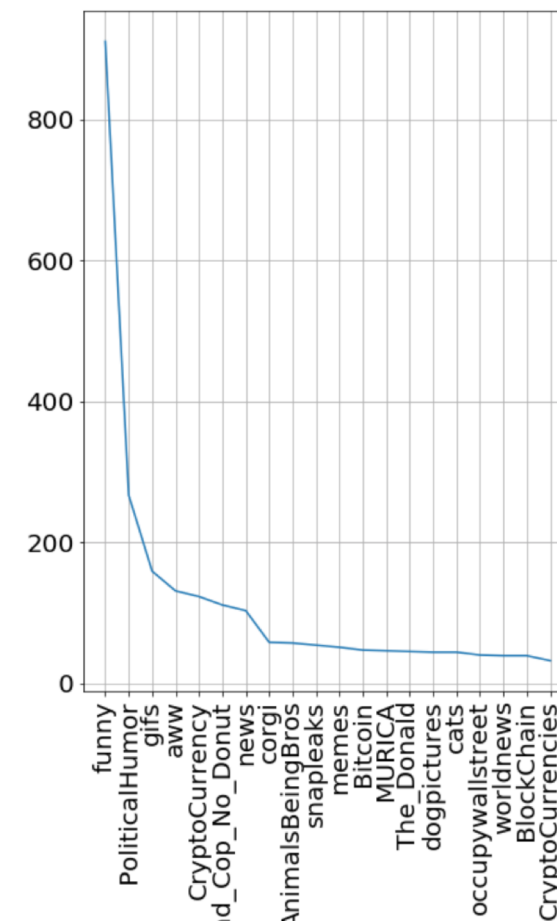


Submissions

Title



Subreddit



2011-09-02 2015-06-30 2015-07-31 2015-09-04 2015-10-08 2015-11-11 2016-03-12 2016-06-18 2016-09-04 2016-10-14 2016-12-05 2017-09-10 2017-11-26 2018-01-09 2018-02-15 2018-03-26

# Next Step

- “There were about 14k posts in total by all of these users”
  - Continue with collecting data and EDA
- Select a few subreddits to collect all comments and posts from 2015-2016
- Create model to predict if comment or post is from a “suspicious” account or a normal account