

Reddit: Analysis of “Suspicious” Accounts

STAT 418 Project Proposal

Janella Shu

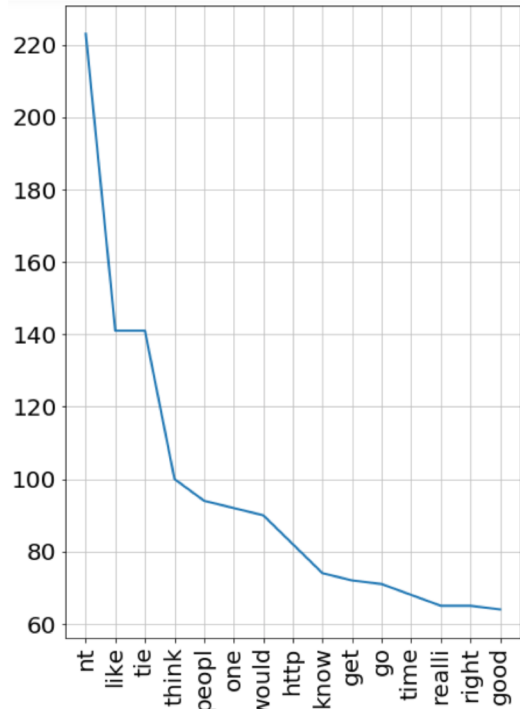
Dataset

- 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/)
- 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)

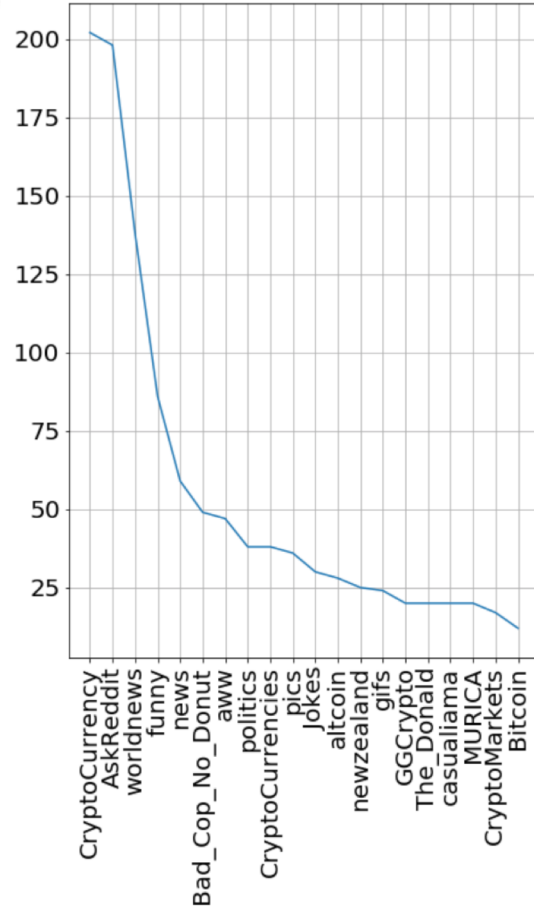
Exploratory Data Analysis

Comments

Body

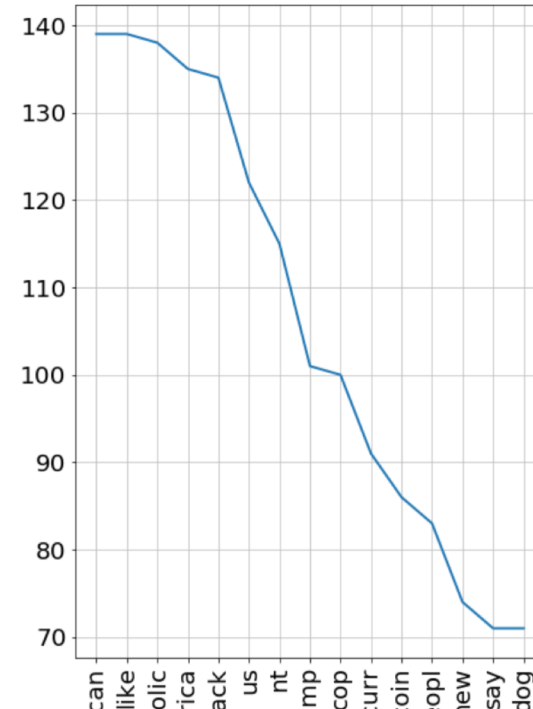


Subreddit

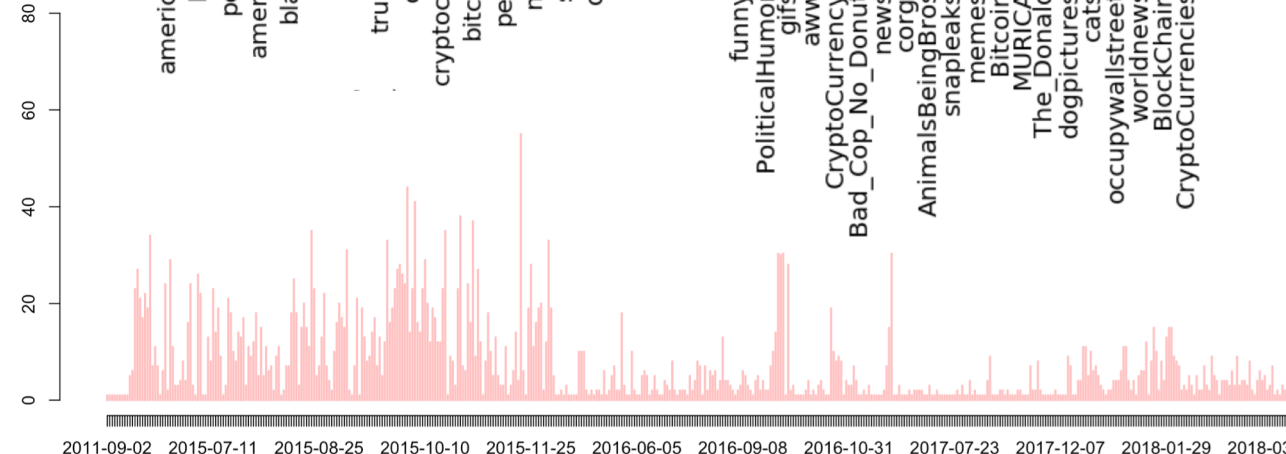
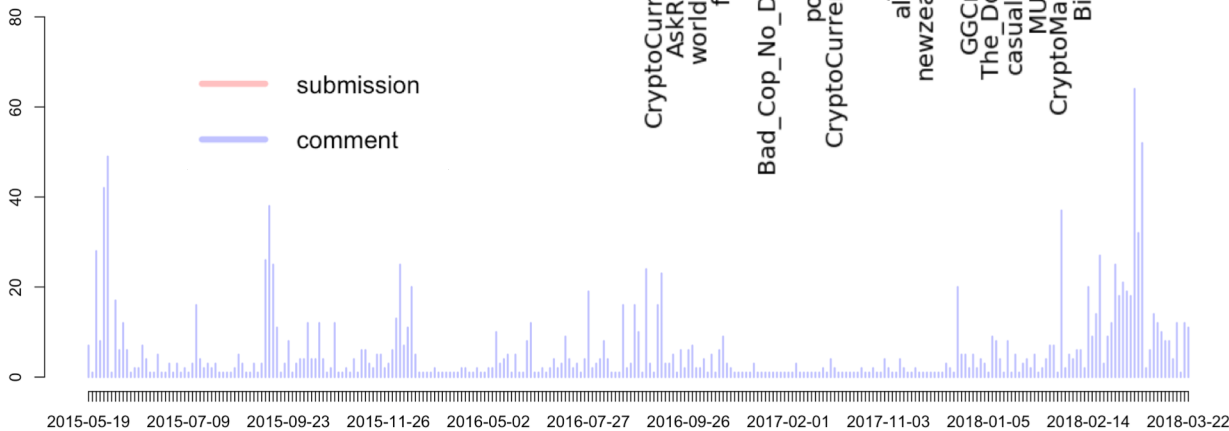
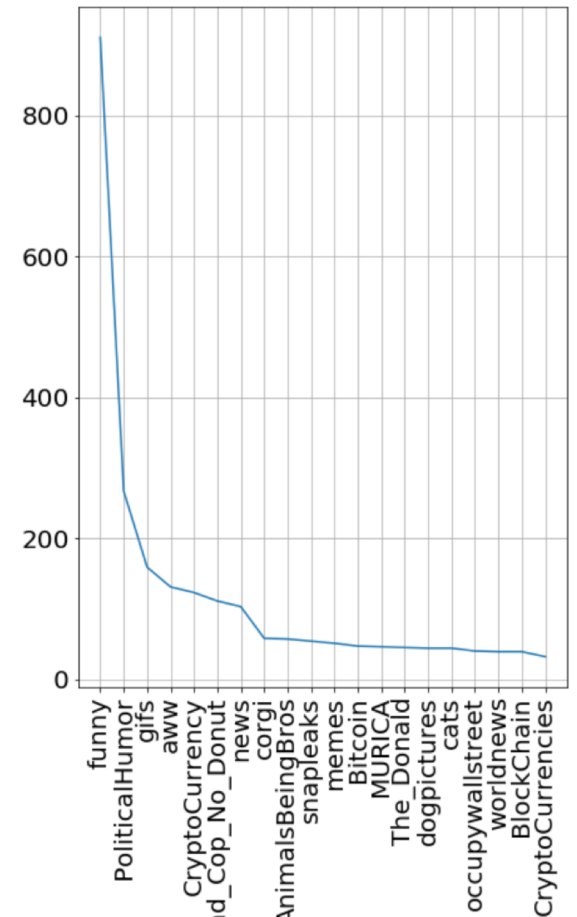


Submissions

Title



Subreddit



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