/r/MiningReddit

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

Project Description

Our data set consists of every comment from the month of January 2017. We plan to derive interesting or useful information from these comments, by text mining for significant trends and patterns using various processes discussed in class.

Project Description

Some potential questions we could mine for are?

- Can we infer a network of connections between people within a subreddit based on their comments?
- Can we predict events (election results, natural disasters, fashion trends, etc.) by mining user comments?

Prior Work

- There is a lot of research that attempts to analyze social media data for various purposes
  - “Twitter as a Potential Risk Reduction Tool”
  - Brandseye: company that analyzes Twitter comments predicted the outcome of the 2016 election and ‘Brexit’
  - Lexicography of German Subreddits, frequently used place names, smileys, and parts of speech
  - “role of the social web in behavioral therapy” specifically using subreddits and discussions about mental health
  - Analyzing which comments get up and down votes
  - Several comment generators who learn from reddit comments

Dataset

The dataset we will use was found within Reddit itself at the following link:

https://www.reddit.com/r/datasets/comments/5t3vfd/reddit_january_2017_comments_are_available_a/

- Size: 42 GB
  - Mine out specific subreddits/timestamps to make project more manageable
- JSON format -- many available tools for analysis

Proposed Work

- Pre-processing is needed
  - Sampling to test code
  - Feature Extraction
  - Reduction of size
- Pattern Recognition/Data Mining
  - Classification, Clustering, Regression
- Possibly:
  - Visualization?

Tools

Required:

- Python
- JSON
- 60 GB of space
- Thumbdrives

Possible:

- MongoDB
- WEKA
- R

Results

- The correlation between subreddits by looking at common comment authors
- The correlation between subreddits by looking at popular words
- Trending culture/social trends by analyzing frequency of words and phrases