Comparing Machine Learning Model Coefficients Between Reddit Posts

Robert Corbett

Goals

- Use machine learning models to discover tokens used to predict the subreddit thread and the score of a post
- Compare the tokens to find which ones are both used to predict the thread and a high score for the post

Data

- Reddit makes data available for fair use
- Link to data is available on my Github
- Repository has every post
  - Beginning December 2005
  - Json format
  - A lot of data

Data

- Post Information
  - parent_id
  - author
  - distinguished
  - body
  - gilded
  - score
  - author_flair_css_class
  - stickied
  - retrieved_on
  - author_flair_text
  - id
  - subreddit

Data Curation

- Data is from June, August and September
  - 41.0GB June
  - 48.4GB August
  - 47.2GB September
  - July was corrupted

Data Curraion

- List of subreddit threads

  -First 100,000 posts in August

  - Produced almost 10,000 unique tag
- Picked 24 that looked promising
  - Political/Biased
  - Political/No bias
  - Control

Data Curation

- Activity of subreddit threads
  - Can I retrieve enough posts?
- Length of posts
  - Are the bodies of the posts substantial?
  - Greater than 150 characters

Final Data

- 8 subreddits
  - Android
  - boardgames
  - Conservative
  - hockey
  - Libertarian
  - neoliberal
  - politics
  - worldnews

Jupyter Notebook