Shower Thoughts or Today I Learned?

—

By: Cassy Clark

Natural Language Processing and Reddit

- Natural Language Processing is everywhere around us
- Reddit is a popular social media app
- Subreddits of choosing: Shower thoughts and Today I Learned
- Models to be used: Multinomial Bayes Classifier and Logistic Regression

Getting Started

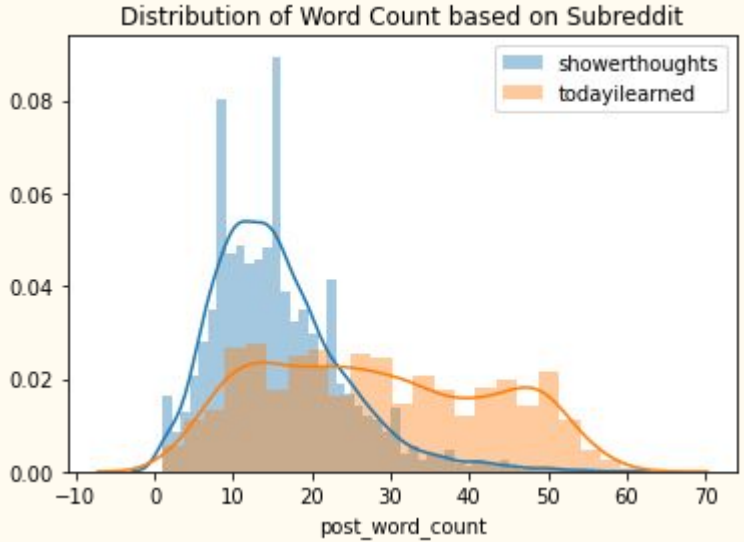
- Used pushshift API to request posts from subreddits
- Attempted 9,000 pulls from each subreddit making sure to drop duplicates
  - Try not to get kicked out, Y I K E S
- Shower Thoughts total of 8,856 posts
- Today I Learned total of 5,499 posts

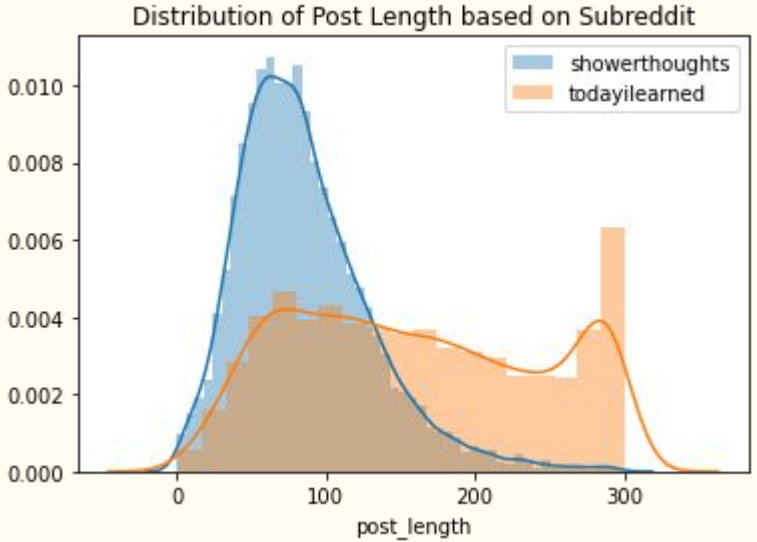
Data Cleaning

- First step combine both data frames
- No null values, yay!
- Removed punctuation from titles
- Original data frames contained: Author, title, subreddit, and created_utc
  - Combined data frame dropped author and created_utc
- Shower thoughts → 1 and Today I Learned → 0

Exploratory Data Analysis

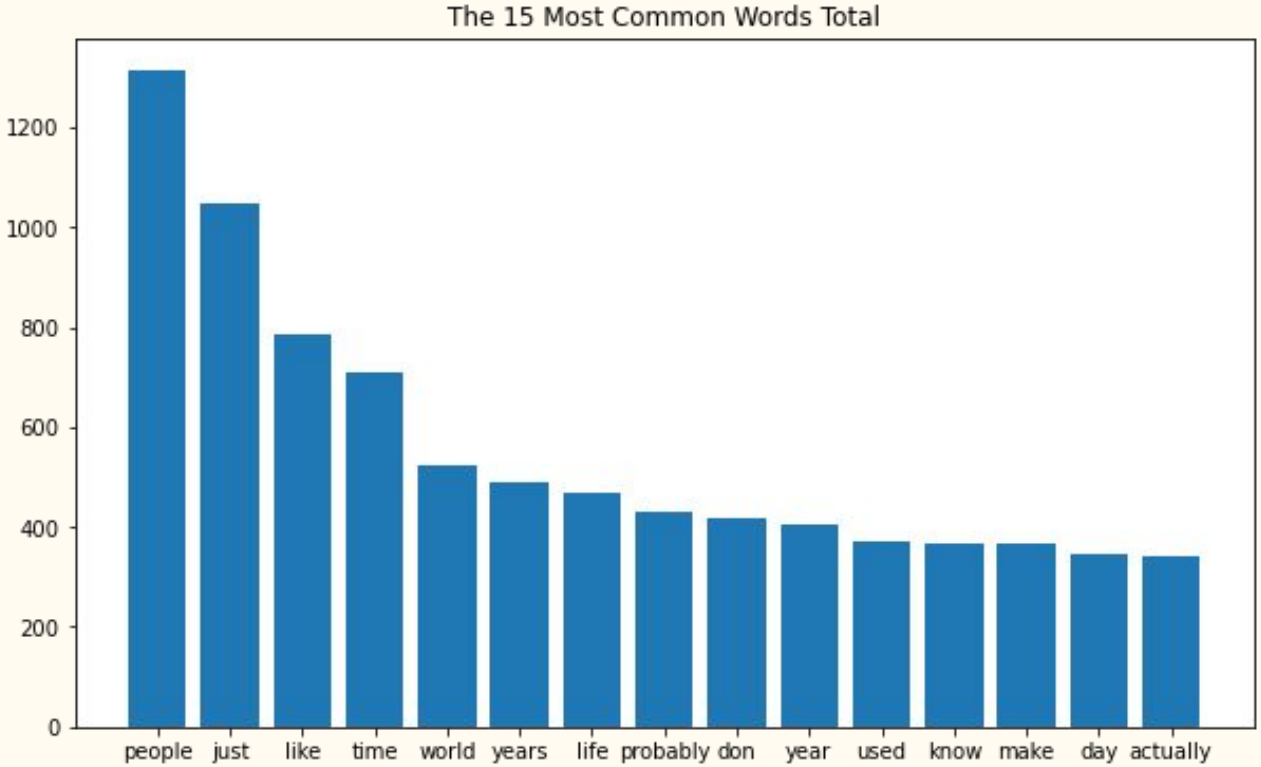
- Post length and word count added to combined data frame





EDA cont.

- Most common 15 words and least common 15 words
- Most common 15 word pairs and least common 15 word pairs



Sentiment Analysis

- Ran a sentiment analysis to check the positive, negative, and neutral emotion of each post
- Overall, bot subreddits generally had a neutral sentiment

subreddit	positive	neutral	negative
0	0.069555	0.863174	0.066721
1	0.078210	0.842359	0.078641

Multinomial Bayes Model

- Baseline score: 61.69%
- Tested using count vectorizer and tf-idf
- Stop words included the standard english as well as words/abbreviations related to the subreddit name
- CV scores
  - 0.833 - Train
  - 0.831 - Test
  - 0.743 - Specificity
- Tf-idf scores
  - 0.821 - Train
  - 0.824 - Test
  - 0.663 - Specificity

Logistic Regression Model

- Set up the same as Bayes Model by using various parameters that were run through a gridsearch and using count vectorizer and if-idf
- CV scores:
  - 0.832 - Train
  - 0.834 - Test
  - 0.703 - Specificity
- Tf-idf scores:
  - 0.827 - Train
  - 0.818 - Test
  - 0.683 - Specificity

Conclusions

- Try not to get kicked out and pull more post requests to create a more balanced class
- Use NLP modeling to help Reddit and other social media platforms find trends in postings
- Produced a relatively good model that was neither under or overfit.

ANY QUESTIONS? :)