Project 3: Web APIs & NLP
Fallout vs. Star Trek

Regene M. DePiero

General Assembly

regenedepiero@gmail.com

October 9, 2020

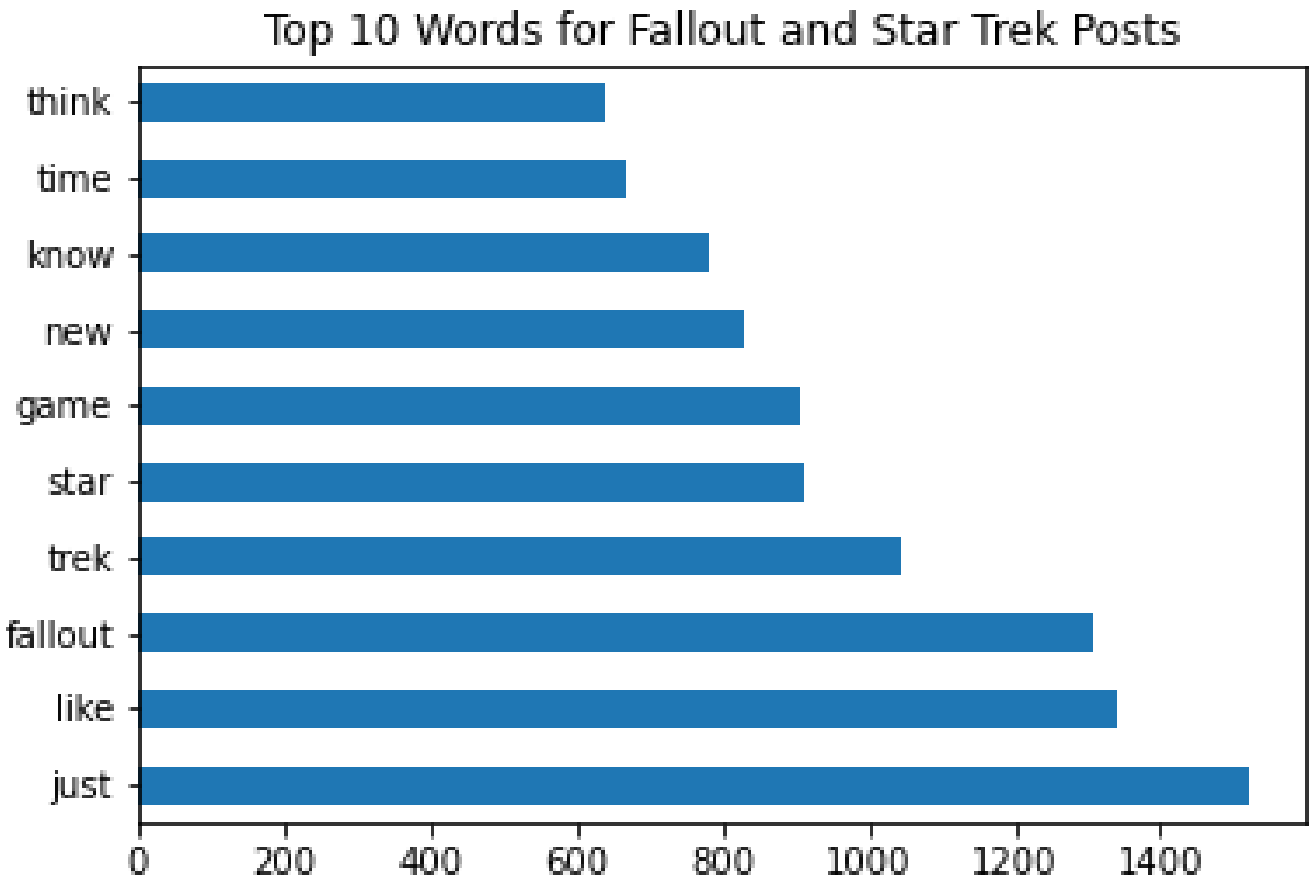
Overview

The Problem

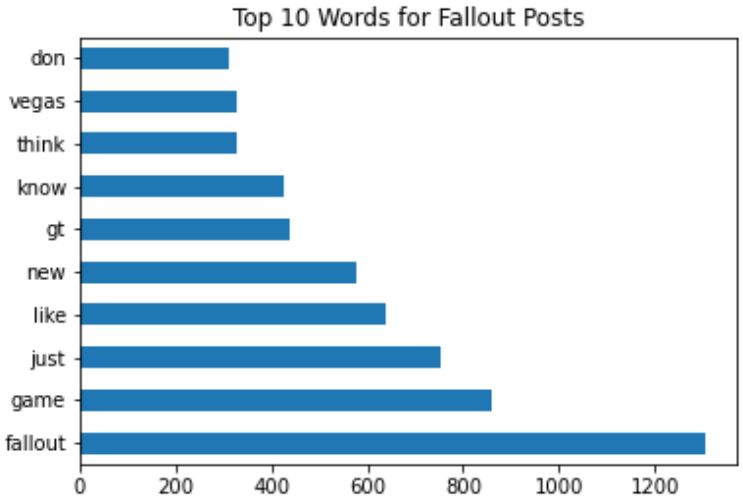
Train a classifier that can determine which subreddit a post came from.

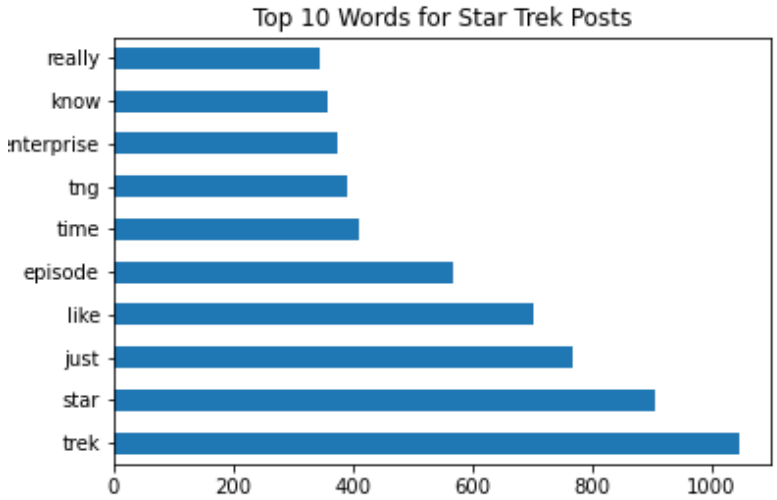
Subreddits of Interest

Top Words


Top 10 Words for Fallout and Star Trek Posts
think
time
know
new
game
star
trek
fallout
like
just
0
200
400
600
800
1000
1200
1400

Top Words by Subreddit





Naive Bayes

- Best model used $\alpha = 0.0001$
- max_df $= 0.9$

Scoring:

- F1 Score Training data: 0.9577
- F1 Score Testing data: 0.9437
- F1 Score on complete data: 0.9542

Random Forest

- No special parameter selection was done on this model.

Scoring:

- F1 Score Training data: 0.9872
- F1 Score Testing data: 0.9211
- F1 Score on complete data: 0.9704

Conclusions

- Increasing n-grams did not improve model performance.
- Use of TfidfVectorizer did not improve model performance.
- Further processing of data could be done to clean text.

The End