Reddit Classification

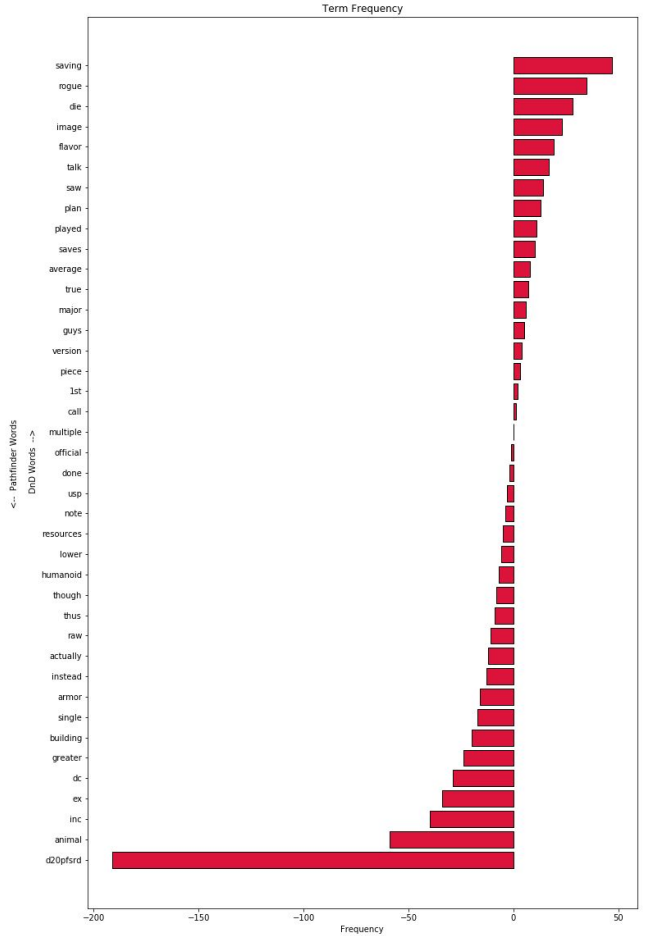
Analysis by Derek Steffan

Introduction

- DnDNext and Pathfinder_RPG
- Many common words, fewer unique words
- How to differentiate the two?

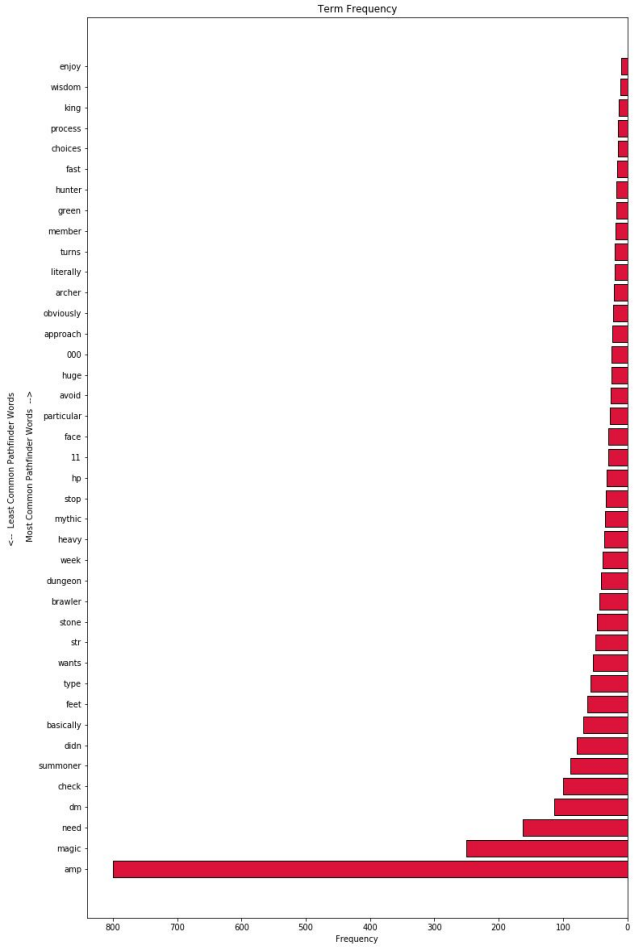
Vectorizers

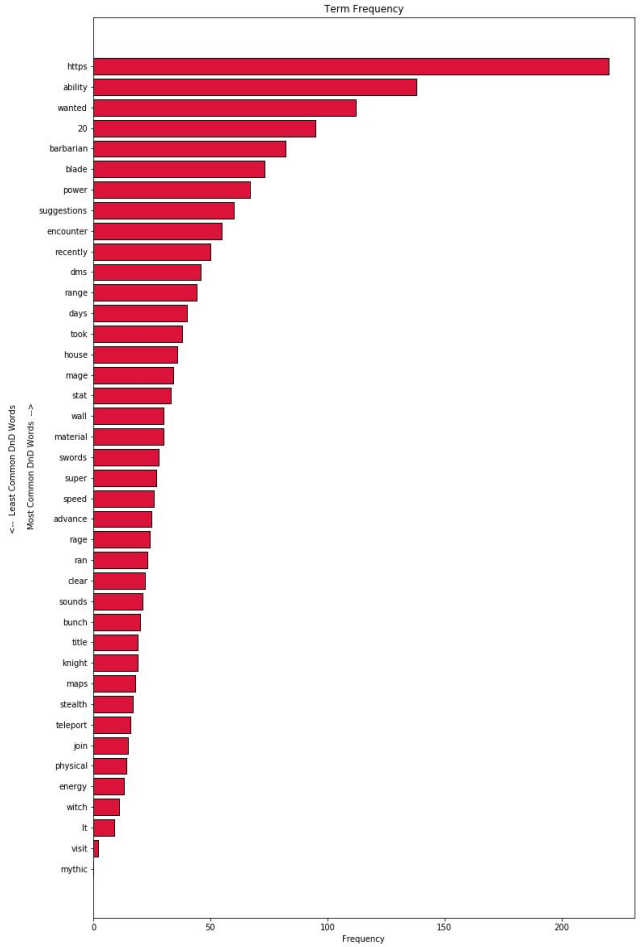
- Count preferred over TF/IDF
- Fit to entire corpus
- Sum up all features to find most common words
- Subtract the vectors to mimic a Venn Diagram



Raw Count Vectors

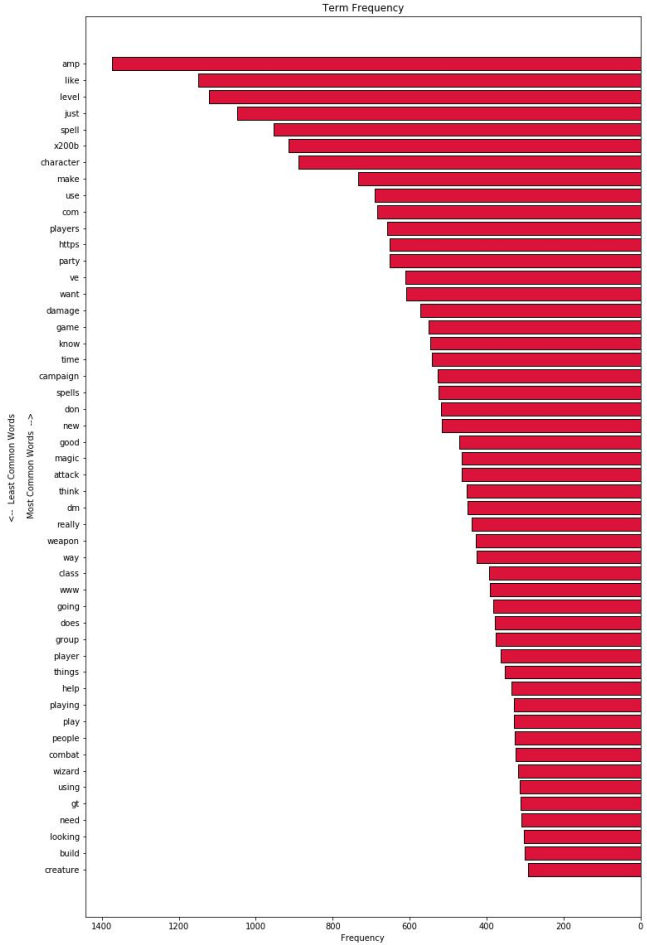
- Note the most common word for each subreddit





Making the model worse...

- Amp, x200b, https…
- Links are very common in r/Pathfinder_RPG
- Bias up, variance down



Making the model better...

- More stop words → Less features
- Less features → Less variance
- Finding the middle words from our Venn diagram
- Take the 500 most common words from the corpus
- Compare these against a slice from the middle of the diagram

Results

- Testing set

  → Accuracy: 83.4%

- Scrape 500 more posts

  → Accuracy: 82.4%

- Encouraging results...

Results

- Scrape 500 posts from r/Pathfinder and r/DnDBehindTheScreen

  → Accuracy: 68.8%

- With English stop words…

  → Accuracy: 55.2%

Conclusions and Caveats

- Don’t overfit…
- Hyperlinks
- Daily/weekly automated threads