WHAT MAKES A BOOK SUCCESSFUL

Mark Sammartino
Executive Publisher
HarperCollins Publishers

Fantasy Tropes
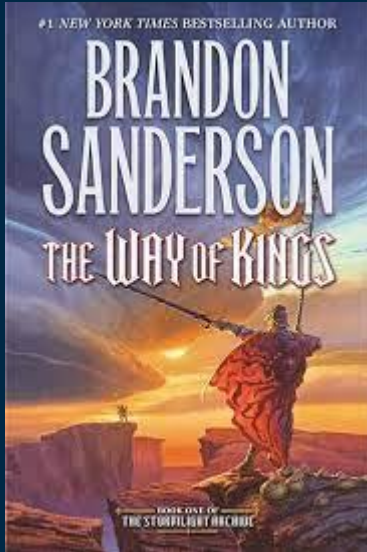
Swords

Magic

Dragons

A Bunch of Made Up
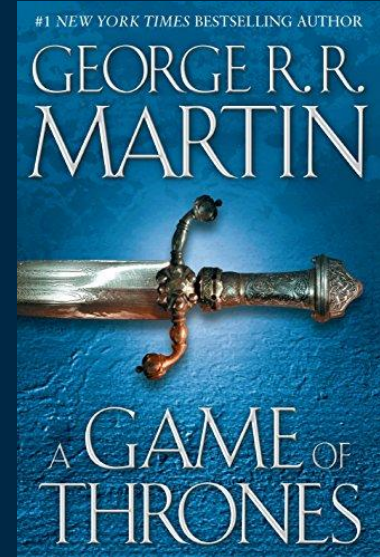Nonsense Words

What are Customers Reading?

Commonalities

Series specific jargon

Unique Character names and locations

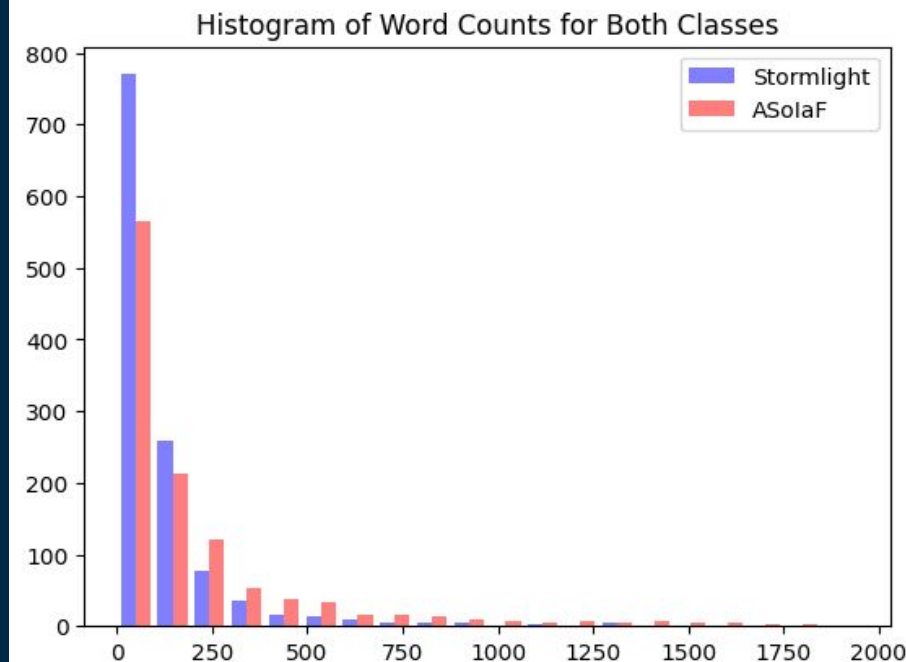Made up magic, powers and prophecies
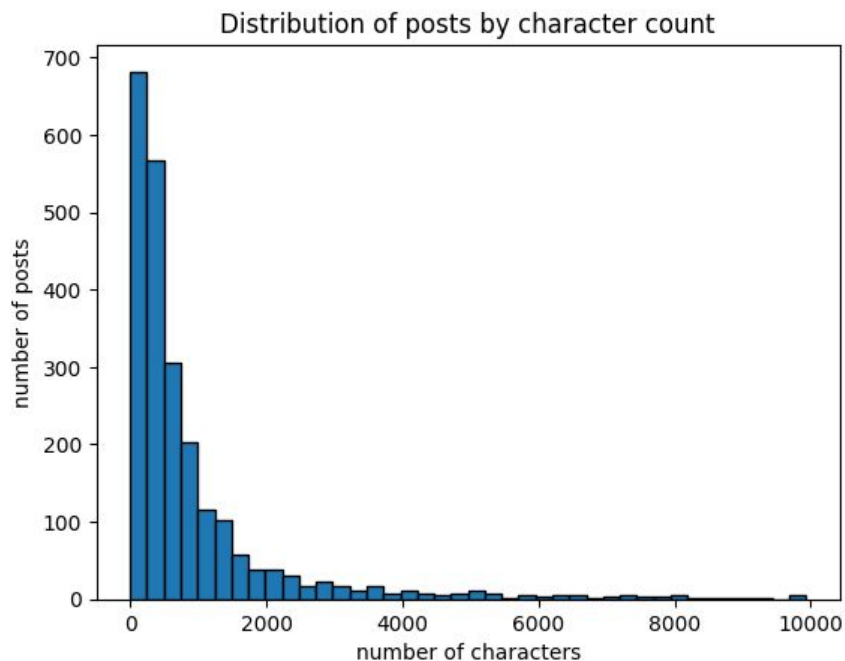
r/Stormlight_Archive

r/asoiaf

Reddit is Messy!

- Tokenized
  - Removed punctuation, symbols, links and unicode
- Removed Stopwords
- Lemmatization
  - Reduce words to root form
- Counted characters and words

Let's Take a Look

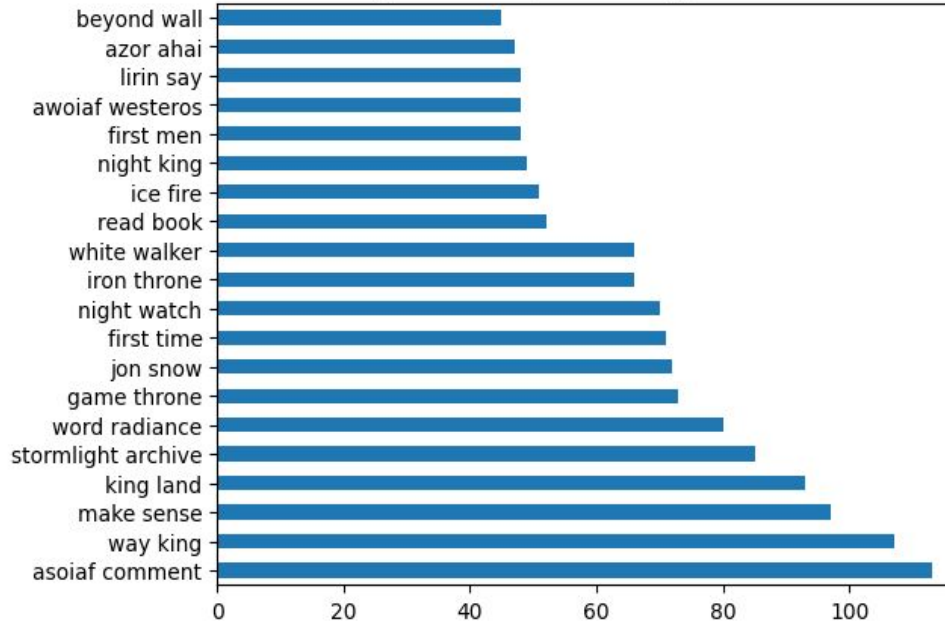Distribution of posts by character count

Histogram of Word Counts for Both Classes

	char_count	word_count
subreddit		
Stormlight_Archive	829.953393	145.645135
asoiaf	1629.948962	285.720588

Terms, Names, and Titles

Top 20 words pairs that appear in text

Top 20 original word pairs that appear in text
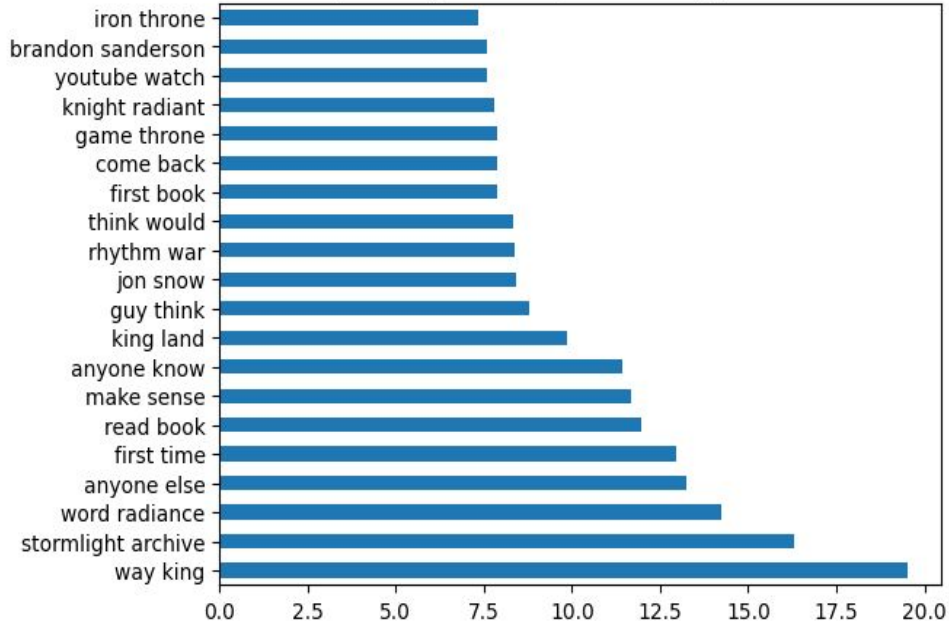
Can We Distinguish?

Random Forest Classifier

- High Classification Accuracy
- Good with multidimensional data
- Creates many trees trained on random rows and columns
- Merges for better model

Support Vector Machine

- Good with high dimensional and unstructured data
- Creates a Hyperplane to separate classes
- Can add dimensionality to separate hard to find clusters of a class

Random Forest Wins!!

- Test Classification accuracy of 0.935
- Balanced classes

Top 20 Important Features from Random Forest
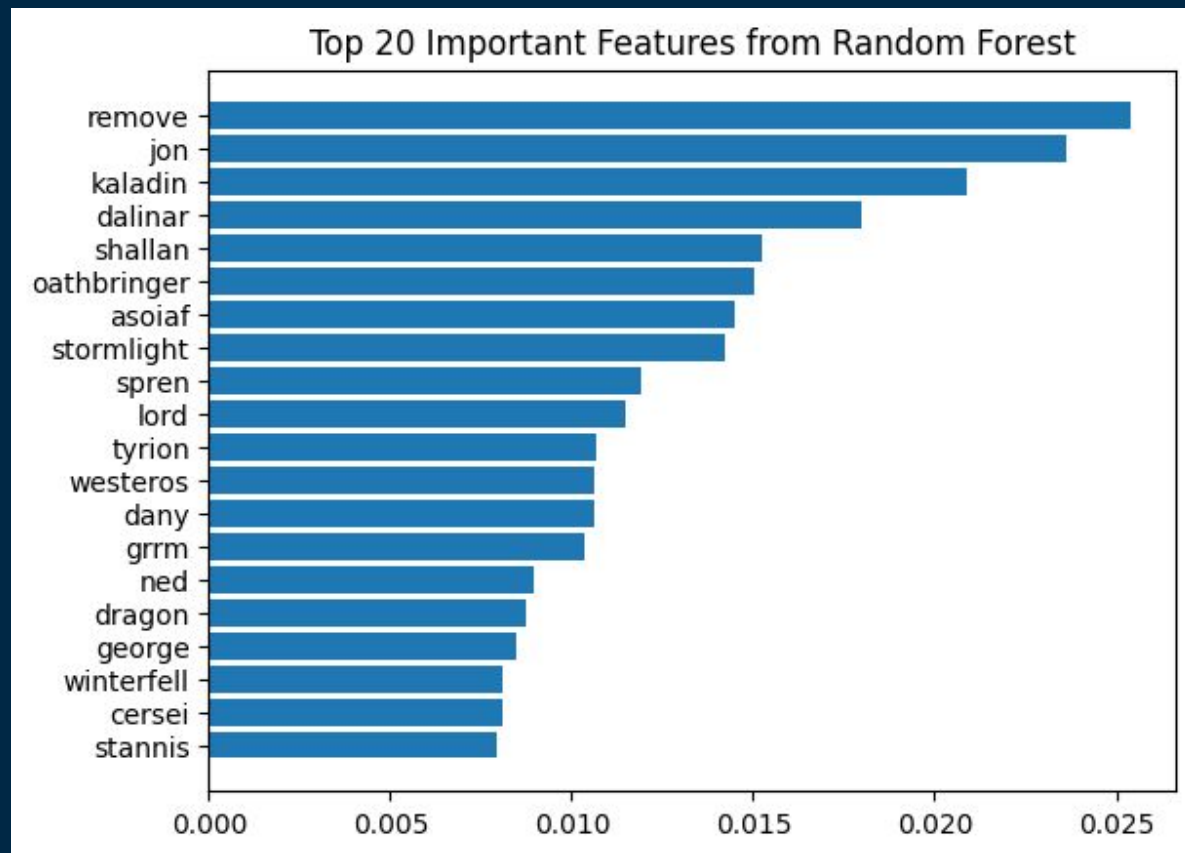
CONCLUSIONS

- Interesting names and locations are important
- Fantasy readers are nerds, they like jargon
- Regular publications keep the fanbase interested
- TV or Movie adaptations result in a spike of activity

Thank You All For Listening!

Are there any questions?

Mark Sammartino
DSI-919
10/28/22