Detecting the difference between two similar subreddits using Natural Language Processing

By Sean McNamara

Presentation Structure

Reddit and NLP Overview

Data Collection

Exploratory Data Analysis and Preprocessing

Model Evaluation

Conclusion

What is reddit?

Reddit - massive online forum

Subreddit - theme based section of the forum

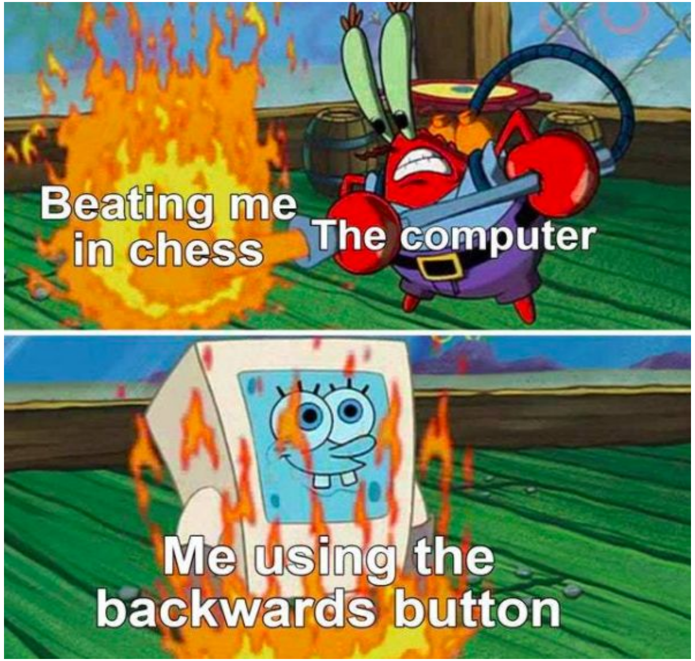
What is NLP?

The ability of a computer program to understand human language as it is spoken and written

‘Chess’ subreddit is serious

- “The home of chess on Reddit.”
- Learning
- Discussions about chess

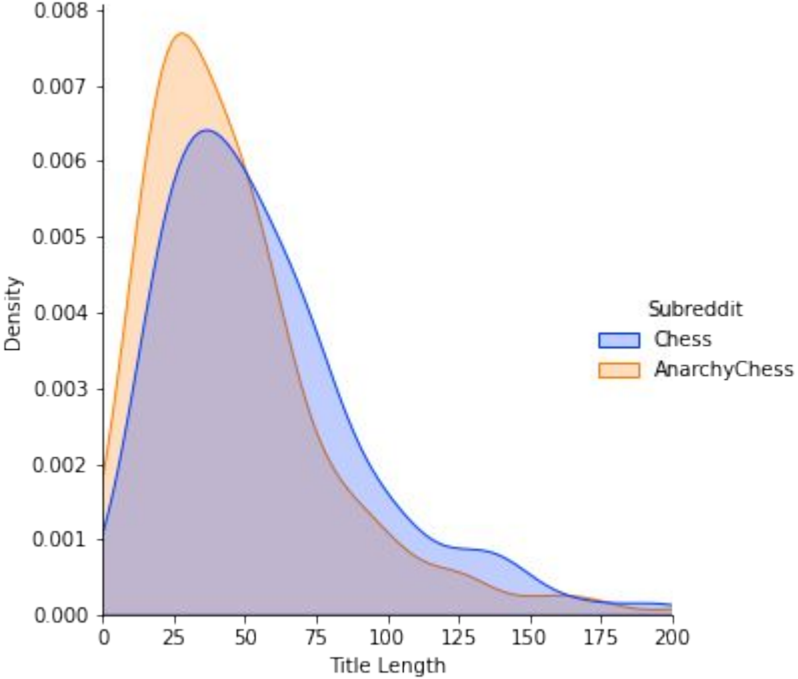
‘AnarchyChess’ not so serious...



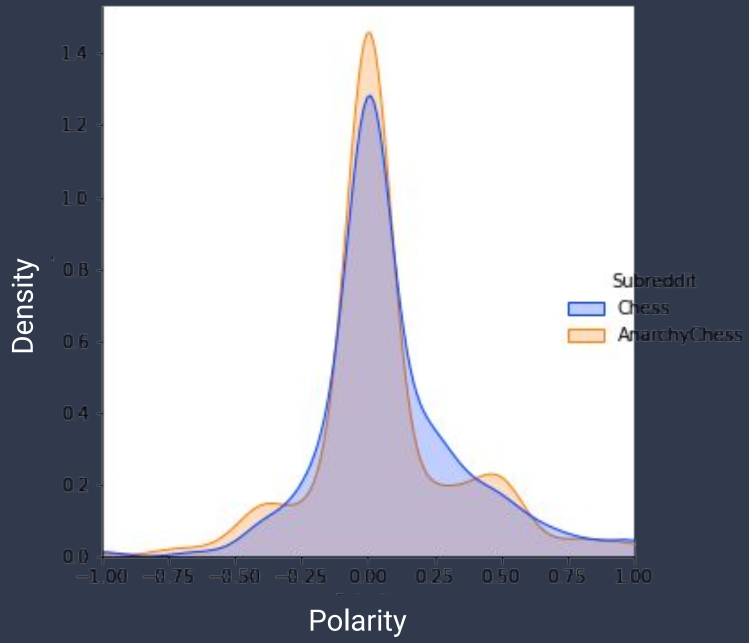
Data Collection

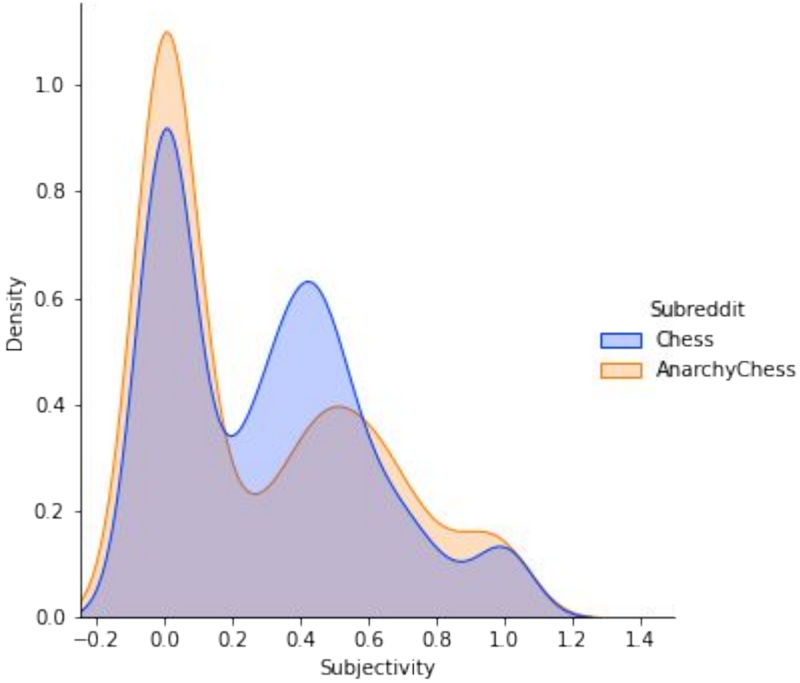
- Reddit API was used
- One hundred titles per year from 2017 to 2021
- Pulled 1000 titles

Anarchy Chess titles tend to be shorter

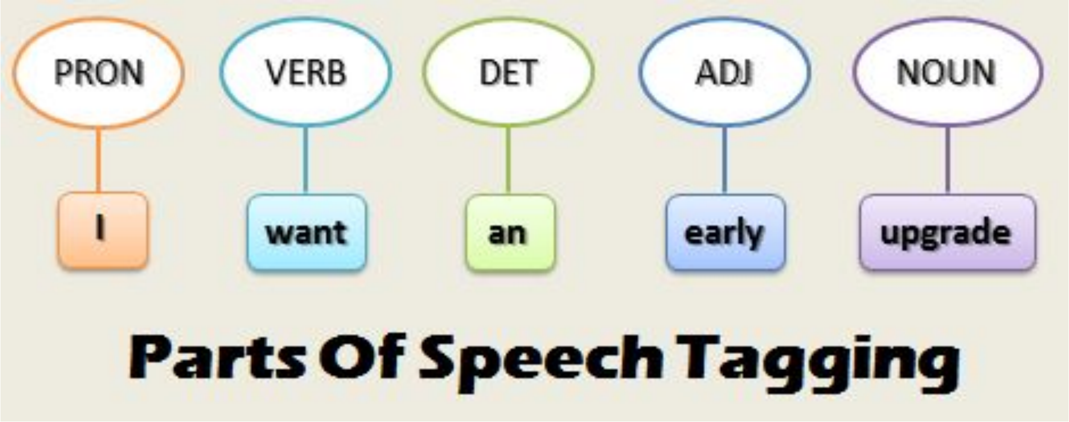


AnarchyChess titles are more emotional

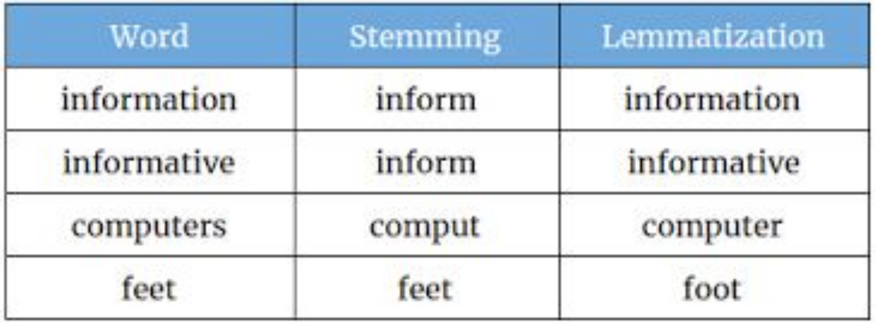




Chess subreddit is more grammatically correct



Lemmatization and Stemming

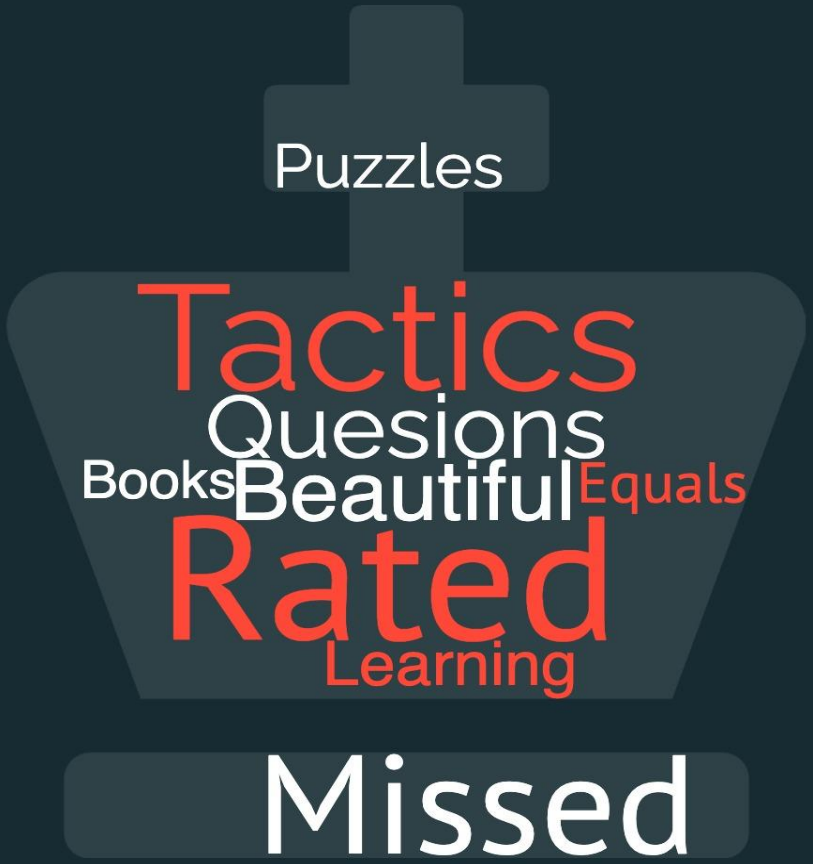


Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

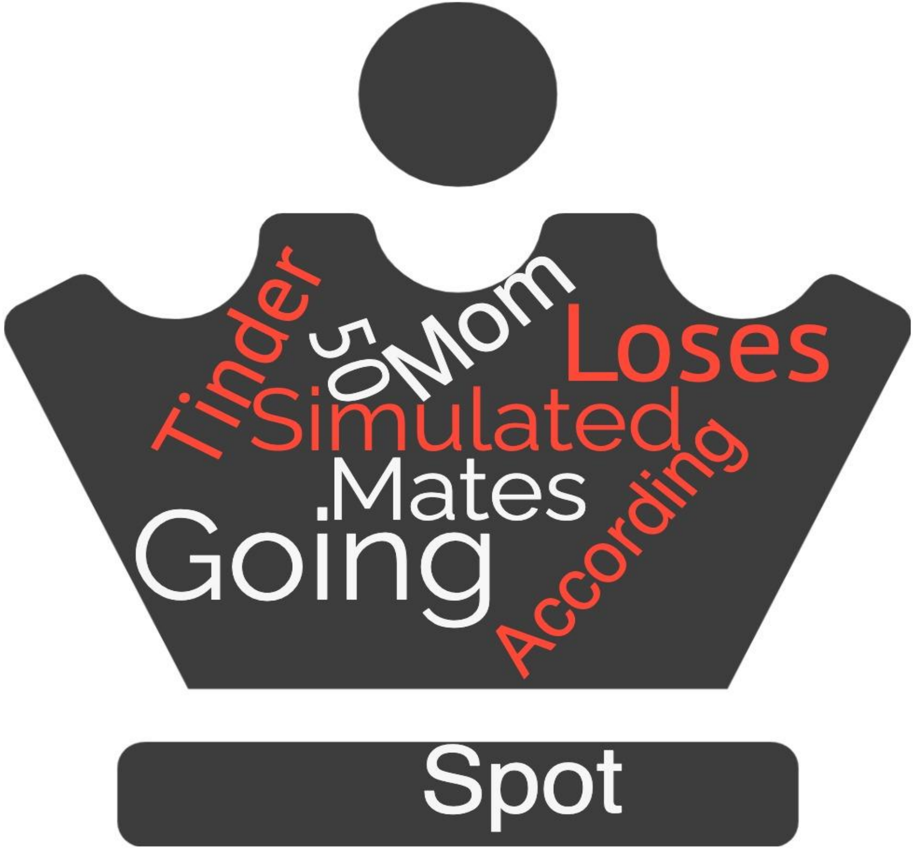
Model Evaluation

Model	Accuracy
Sean’s Brain	72%
Logistic Regression	69%
Naives Bayes	67%
Random Forest	63%
Bagging Classifier	60%
Baseline Model	50%

Typical Chess subreddit



Typical AnarchyChess subreddit



Conclusion

- 69% accuracy
- Similar subreddits
- AnarchyChess slightly random
- Comments needed

Thank you for Listening