Or

Using Natural language Processing to distinguish reddit posts about two acclaimed video games

The Problem:

- Consider two of the most successful and critically acclaimed video game franchises of all time: God of War and The Legend of Zelda
- Both have massive cult followings and have many reddit threads devoted to them.
- There are dozens of subreddits and thousands of posts about two very different games. Let's use NLP to tell which is which.

Obtaining our Data: Scrape Reddit

url, name, n requests	API requests	Output to csv
Our method takes the desired subrredit url, name for outputted .csv, number of API requests. Headers and included in the body and after is initialized as a None type.	Request JSON and store desired properties in temporary dictionary. Retrieve subreddit name, post title, paragraph body, score, name, link flair text. After each loop delay by one second.	Finally convert post dictionaries into a data frame and give appropriate column names.

About our data

951 posts collected

	post traits scraped
0	subreddit
1	title
2	score
3	post paragraph
4	link_flair_text

725 posts collected.

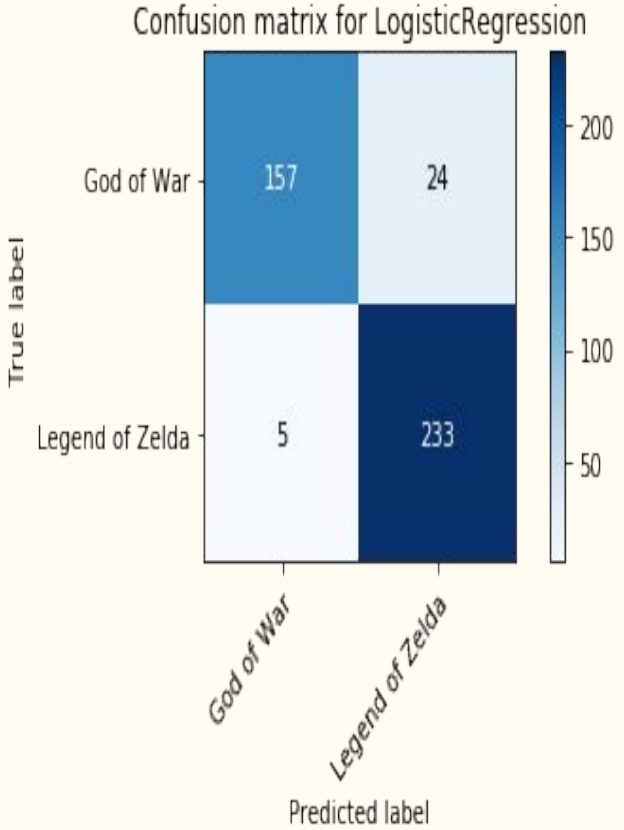
The NLP process

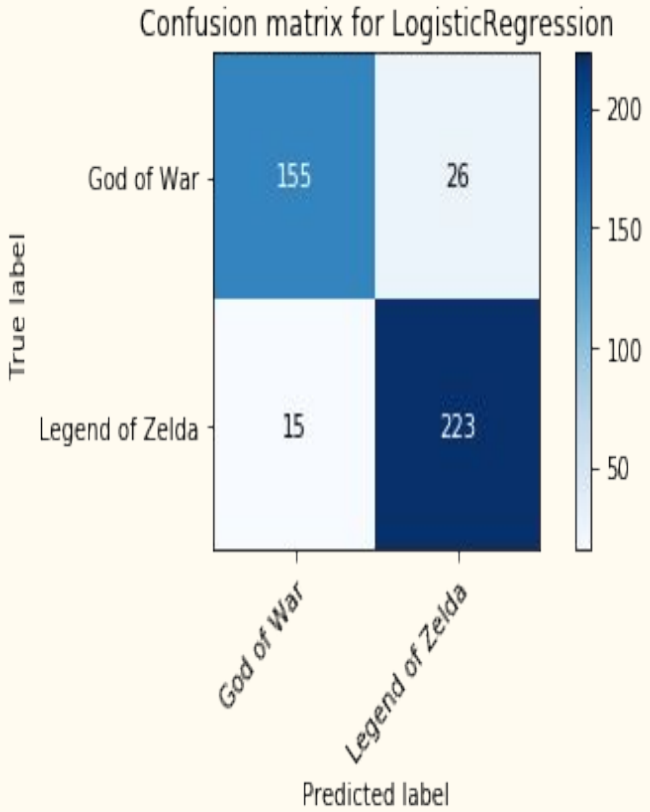
1. Gathered the Data (scrapped reddit's API)
2. Cleaned, Formatted, and Vectorized
3. Selected relevant features and target
4. Train-Test-Split
5. Ran through four models via Gridsearch
6. Analyzed different models performance

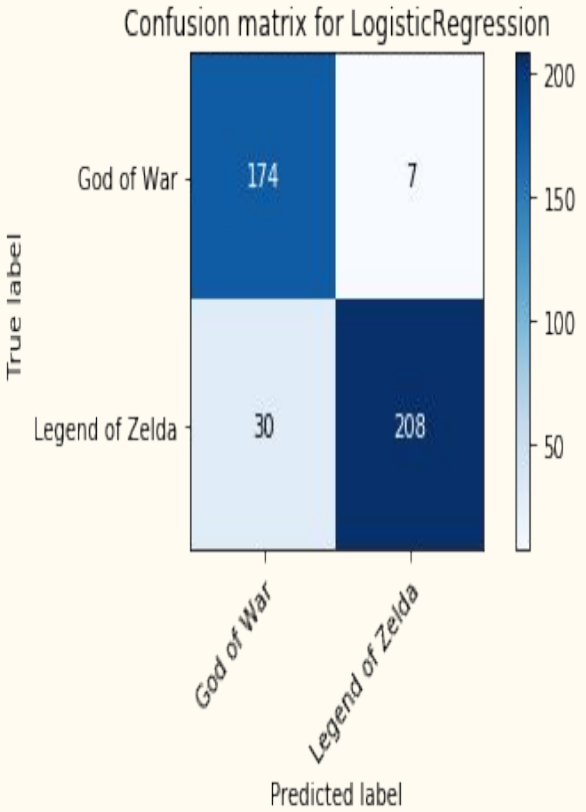
Evaluating Model Performance

Model:	LogisticRegression	**Model:**	Random Forrest	**Model:**	Naive Bayes Gauss
Vectorization:	CountVectorize	**Vectorization:**	CountVectorize	**Vectorization:**	CountVectorize
CV Folds	5	**CV Folds**	5	**CV Folds**	N/A
Parameter 1	C = 1.3	**Parameter 1**	n estimators = 106	**Parameter 1**	N/A
Parameter 2	Penalty = l2	**Parameter 2**	min sampl split = 2	**Parameter 2**	N/A
Train Set Score	1.0	**Train Set Score**	0.9992	**Train Set Score**	0.9976
Test Set Score	0.93079	**Test Set Score**	0.9141	**Test Set Score**	0.9117

Classification Confusion Matrices







Recommendations

1. Use the model r/GodofWar and r/truezelda.
2. Try term frequency-inverse document frequency
3. Try the model on different subreddit pairs