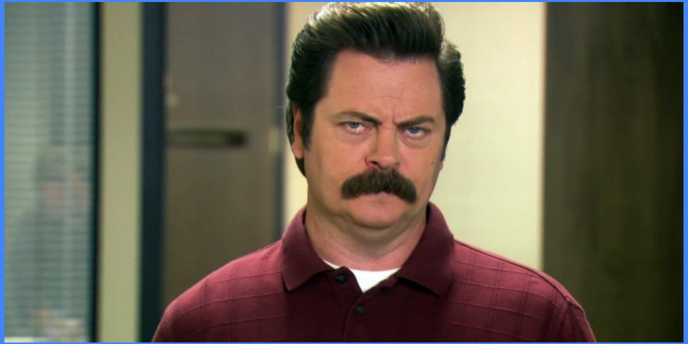
The Office vs Parks and Rec

Jeong Dam (James) Lee



Goal:

Create a model to classify subreddits using comments

1) Data Scraping - Pushshift
2) Cleaning
3) EDA
4) Modeling
5) Conclusion

Data Collection

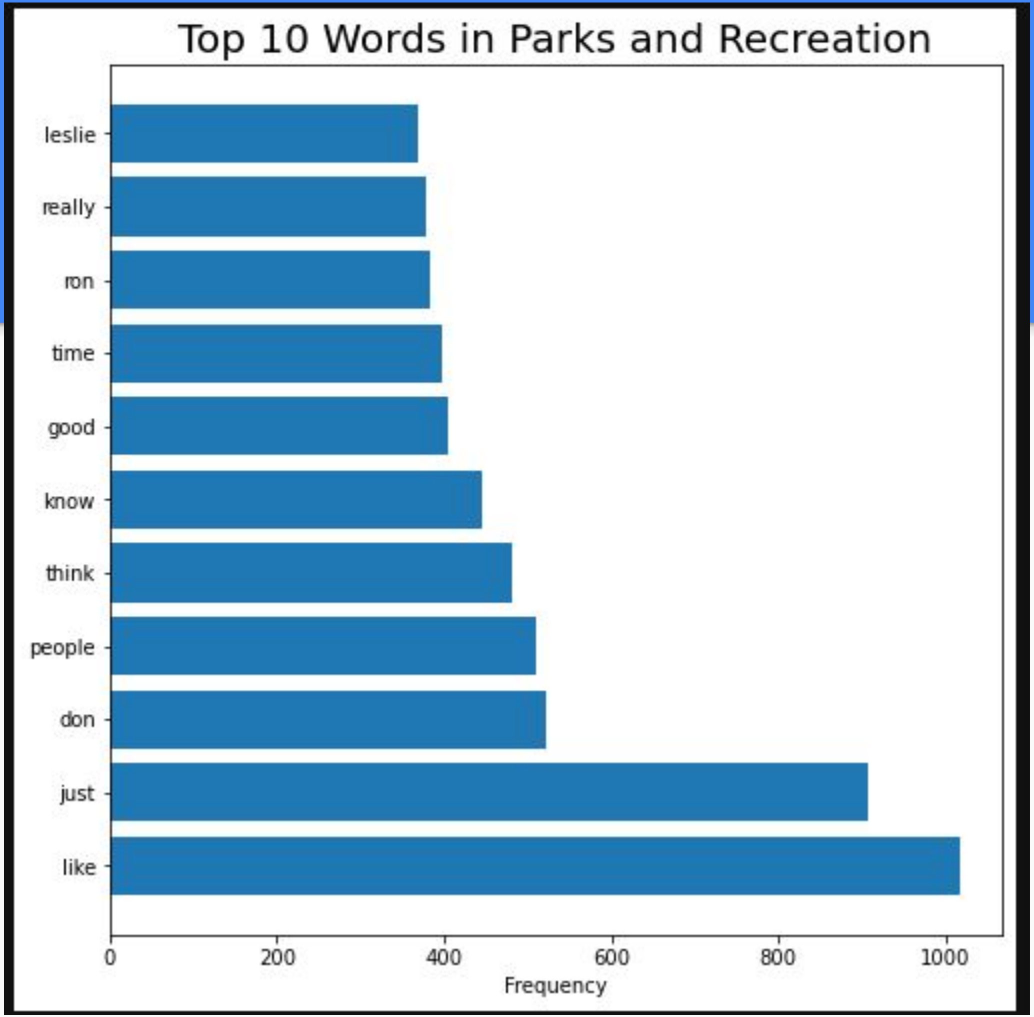
API

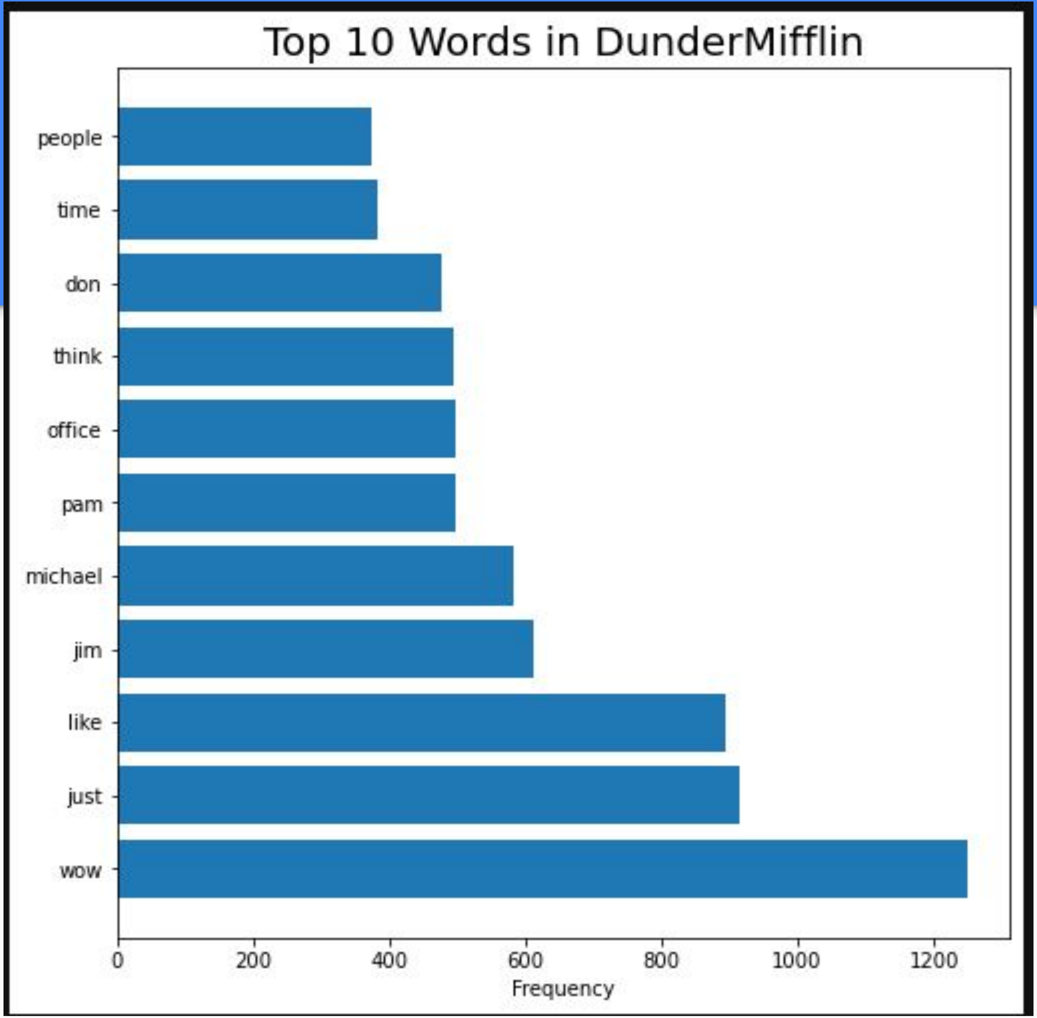
- 10,000 submissions from r/PandR and r/DunderMifflin
- 10,000 comments from r/PandR and r/DunderMifflin

Focused on the comments

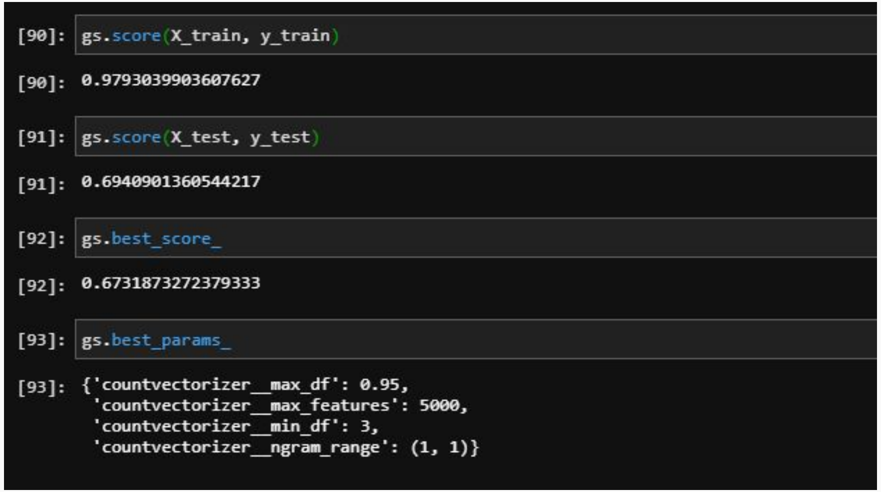
Included in Data

- Author
- Created Time
- Username
- ‘Body’ (the comment itself)
- Link
- Subreddit

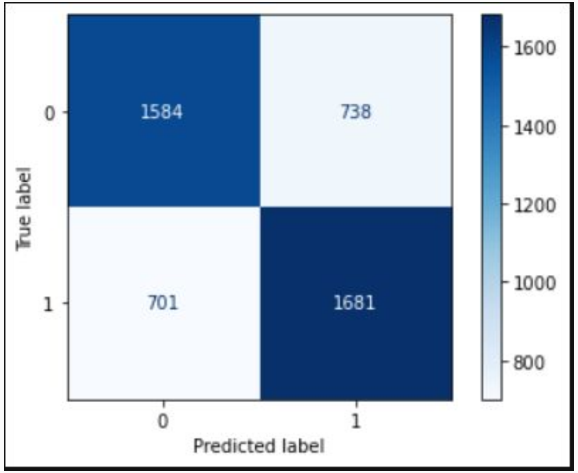

Top 10 Words in Parks and Recreation
leslie
really
ron
time
good
know
think
people
don
just
like
0
200
400
600
800
1000
Frequency


Top 10 Words in DunderMifflin
people
time
don
think
office
pam
michael
jim
like
just
wow
0
200
400
600
800
1000
1200
Frequency

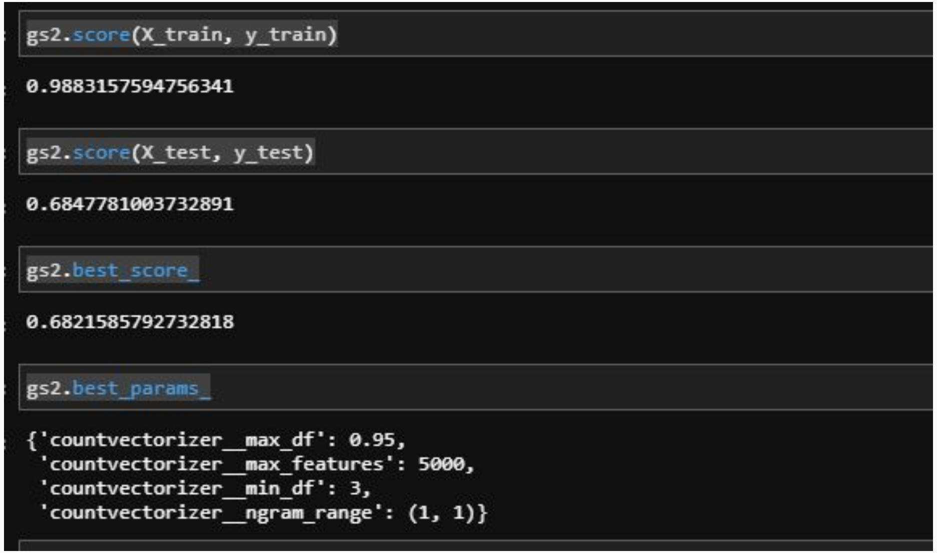
Logistic Regression (acc = .69)



```
[90]: gs.score(X_train, y_train)

[90]: 0.9793039903607627

[91]: gs.score(X_test, y_test)

[91]: 0.6940901360544217

[92]: gs.best_score_

[92]: 0.6731873272379333

[93]: gs.best_params_

[93]: {'countvectorizer__max_df': 0.95,
       'countvectorizer__max_features': 5000,
       'countvectorizer__min_df': 3,
       'countvectorizer__ngram_range': (1, 1)}
```



Random Forest(acc = .69)



```
gs2.score(X_train, y_train)
```

```
0.9883157594756341
```

```
gs2.score(X_test, y_test)
```

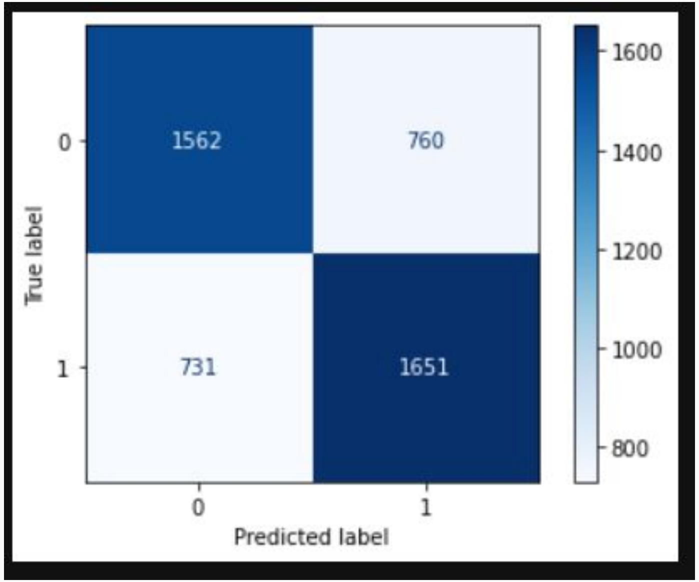
```
0.6847781003732891
```

```
gs2.best_score_
```

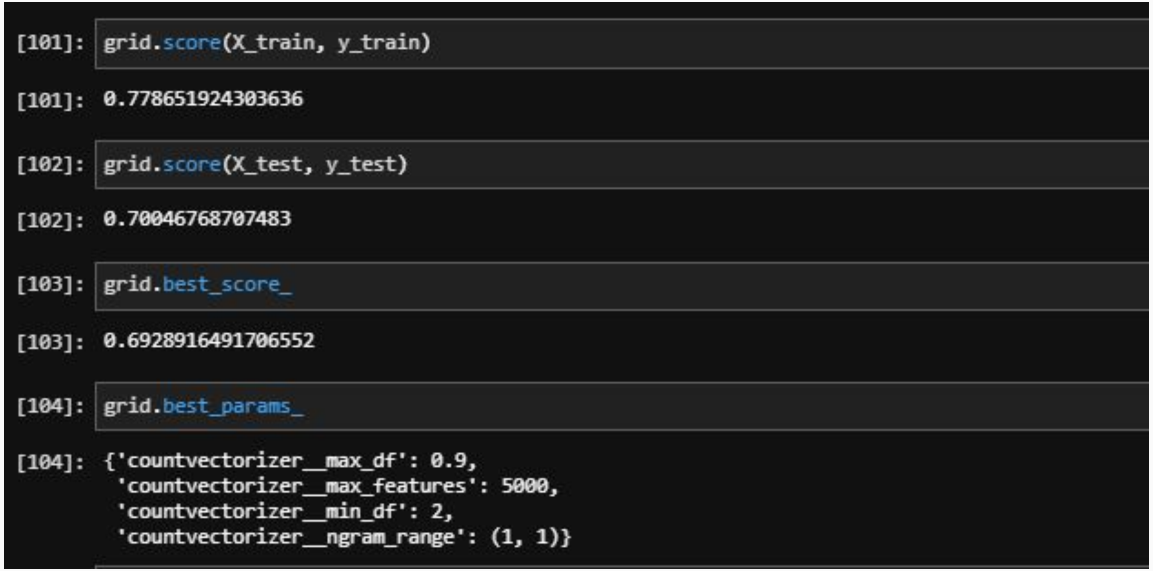
```
0.6821585792732818
```

```
gs2.best_params_
```

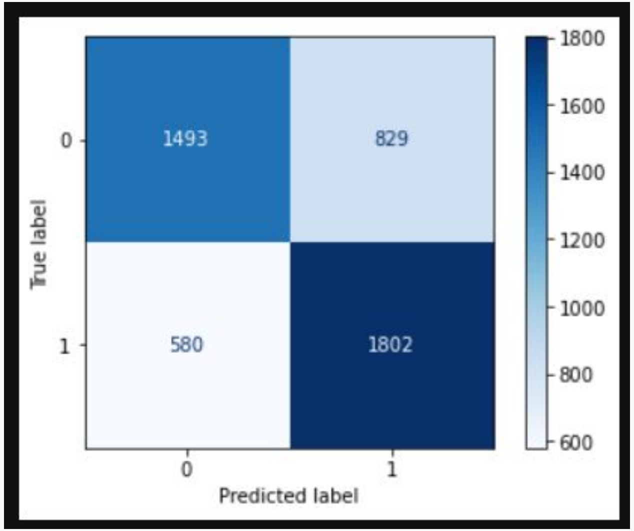
```
{'countvectorizer__max_df': 0.95,
 'countvectorizer__max_features': 5000,
 'countvectorizer__min_df': 3,
 'countvectorizer__ngram_range': (1, 1)}
```



Naive-Bayes - Best Model(acc = .700)



```
[101]: grid.score(X_train, y_train)

[101]: 0.778651924303636

[102]: grid.score(X_test, y_test)

[102]: 0.70046768707483

[103]: grid.best_score_

[103]: 0.6928916491706552

[104]: grid.best_params_

[104]: {'countvectorizer__max_df': 0.9,
        'countvectorizer__max_features': 5000,
        'countvectorizer__min_df': 2,
        'countvectorizer__ngram_range': (1, 1)}
```



The Office vs Parks and Rec

Best Scoring Model:

Naives-Bayes - Train/Test Scores: 0.779/.700

Things to think about:

More/different stop words, better data cleaning, different models (SVM, KNN)

Real Conclusion

The Office is just a better show.