Project 3: Web APIs & NLP

Omar Younis

Problem Statement

- Working for DreamWorks
- Finding Pixar Fans
- Targeted Advertisements

Data Dictionary

Feature	Type	Description
subreddit_name	*object*	The name of the subreddit that the comment came from.
body	*object*	The contents of the comment.
created_utc	*int*	The epcoh time stamp of when the comment was created.
comment_length	*int*	How many characters the comment contains.
word_count	*int*	How many words the comment contains.

Longest 10 Comments

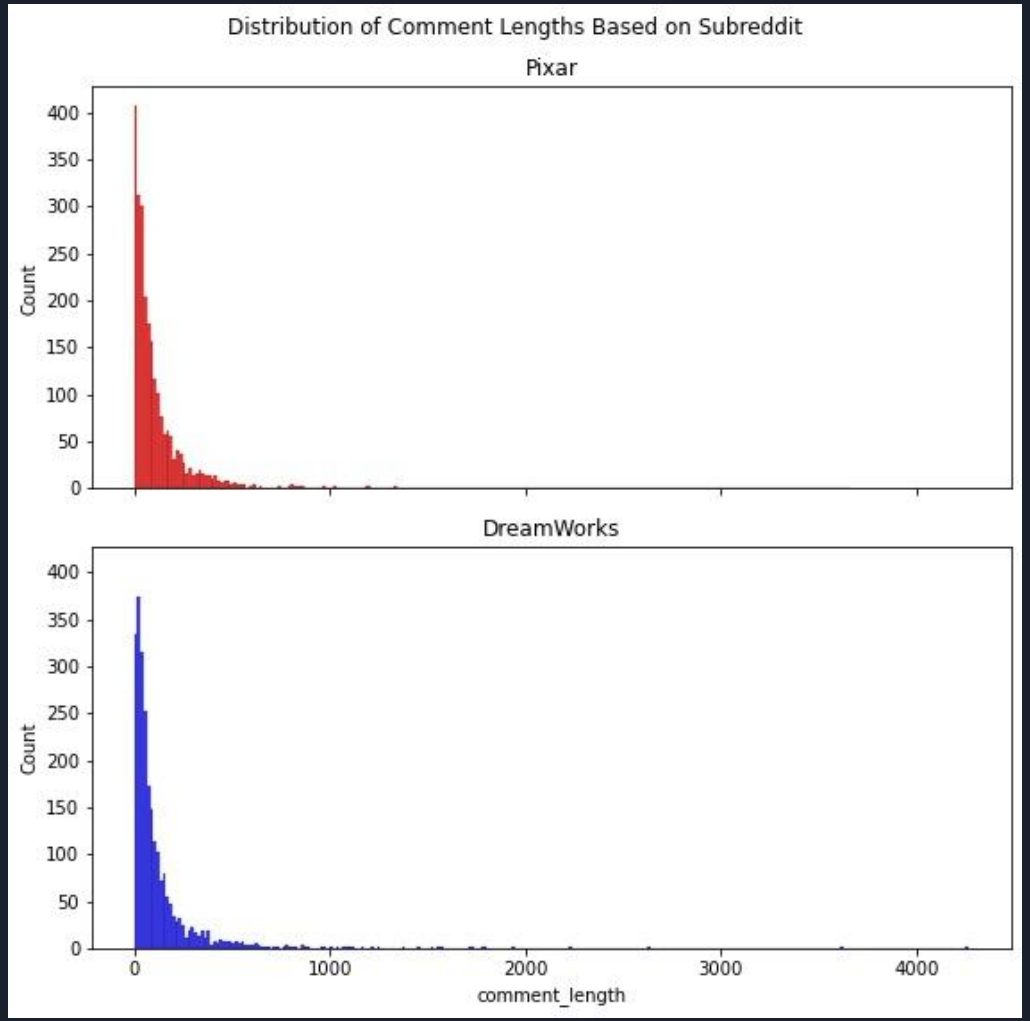
```
------------- Longest 10 Comments by Subreddit -------------
3910    DreamWorks
470           Pixar
3912    DreamWorks
1891          Pixar
2447          Pixar
3914    DreamWorks
3909    DreamWorks
1784          Pixar
961           Pixar
3957    DreamWorks
Name: subreddit_name, dtype: object
```

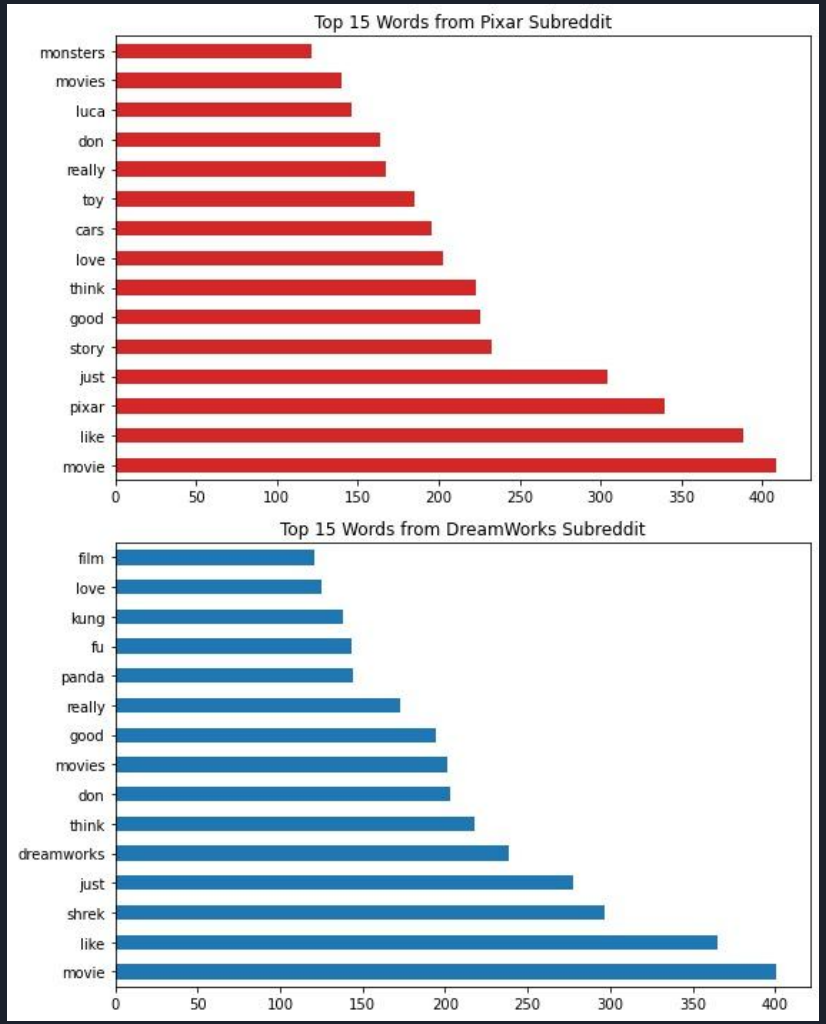
Shortest 10 Comments

```
------------ Shortest 10 Comments by Subreddit -------------
4075    DreamWorks
2568    DreamWorks
3394    DreamWorks
4485    DreamWorks
483          Pixar
2567    DreamWorks
2086         Pixar
4999    DreamWorks
3020    DreamWorks
1354         Pixar
Name: subreddit_name, dtype: object
```

Average Comment Lengths

subreddit_name	word_count	comment_length
DreamWorks	22.83854	123.903448
Pixar	22.71204	123.093504


Distribution of Comment Lengths Based on Subreddit
Pixar
Count
400
350
300
250
200
150
100
50
0
DreamWorks
Count
400
350
300
250
200
150
100
50
0
0
1000
2000
3000
4000
comment_length


Top 15 Words from Pixar Subreddit
monsters
movies
luca
don
really
toy
cars
love
think
good
story
just
pixar
like
movie
0
50
100
150
200
250
300
350
400
Top 15 Words from DreamWorks Subreddit
film
love
kung
fu
panda
really
good
movies
don
think
dreamworks
just
shrek
like
movie
0
50
100
150
200
250
300
350
400

Model Scores

```
------------- Model 1 MNB --------------
Train Score:    0.8686
Test Score:     0.7682
```

```
------------- Model 2 Logr -------------
Train Score:    0.9088
Test Score:     0.7592
```

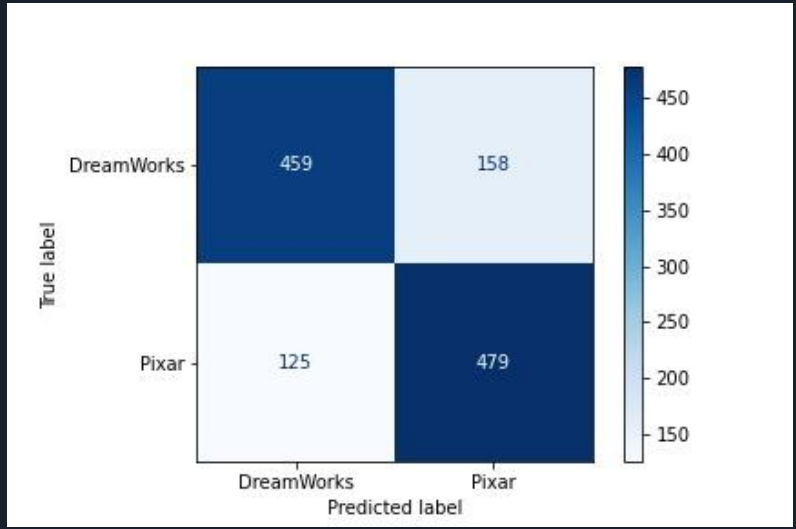
```
--------- Model 3 RandomForest ---------
Train Score:    0.9776
Test Score:     0.7453
```

```
------------- Model 4 KNN --------------
Train Score:    0.7667
Test Score:     0.5913
```

```
------ Model 5 RandomForest Grid -------
Parameter:      {'rf__max_depth': None, 'r
mators': 100}
Train Score:    0.9776
Test Score:     0.7445
```

Precision

Precision of 0.7520



Conclusion & Recommendations

- Deployment
- MultinomialNB() is Best
- More Grid Searches with Pipelines
- Increase Precision Score