# StarWars Vs StarTrek

Presenters: Tresha, Dave and Thomas

# Content

Introduction

Problem Statements

Scraping/Cleaning

Exploratory Data Analysis

Pre-Modeling Process

Modeling Process

Evaluation

Summary

# Problem Statement

- Hired by subreddit users satisfaction team to offer an extra feature that verifies if a post belongs  to the respective subreddit
- Subreddits with a long history or one that has many posts everyday will be automatically filtered
- Save time and manpower
- More user-friendly interface

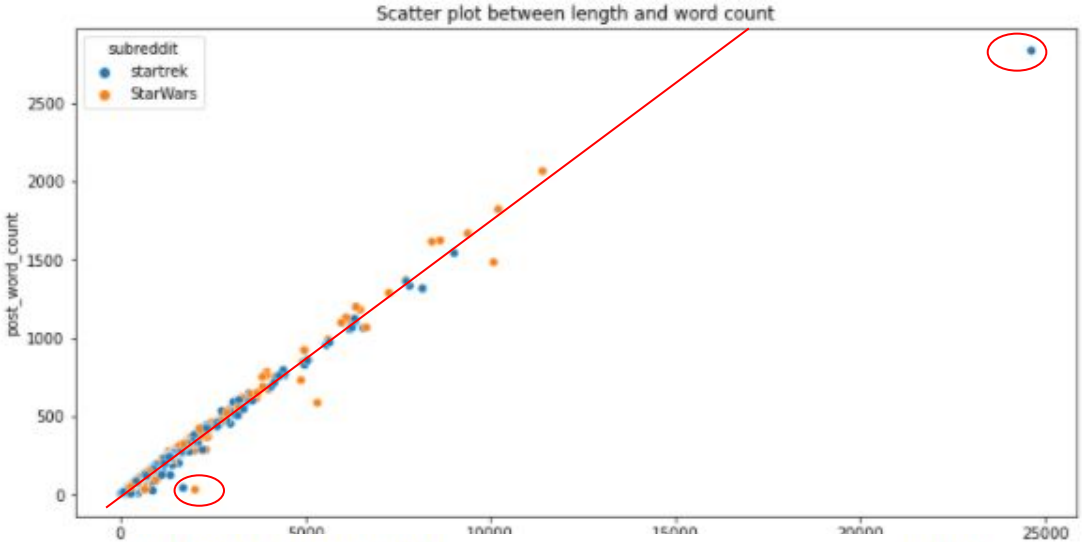
# Data Scraping/Cleaning

- Data generated using requests library
- Data was retrieved from StarWars and StarTrek subreddits
- Data in the form of  post submissions
- More than 1000 posts were retrieved from each subreddit
- Unnecessary columns were dropped  and ‘subreddit’ ‘subtext’ were kept

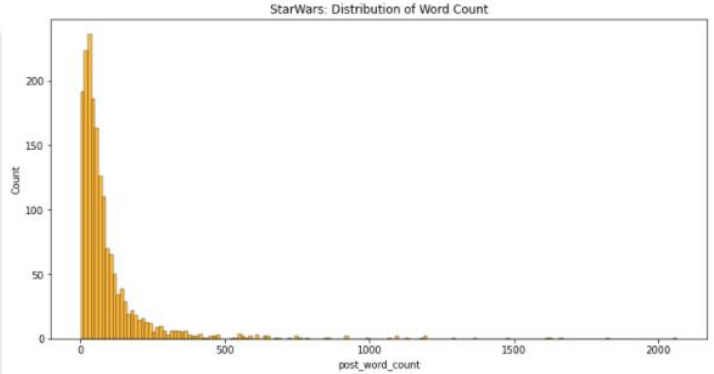
# Data Scraping/Cleaning

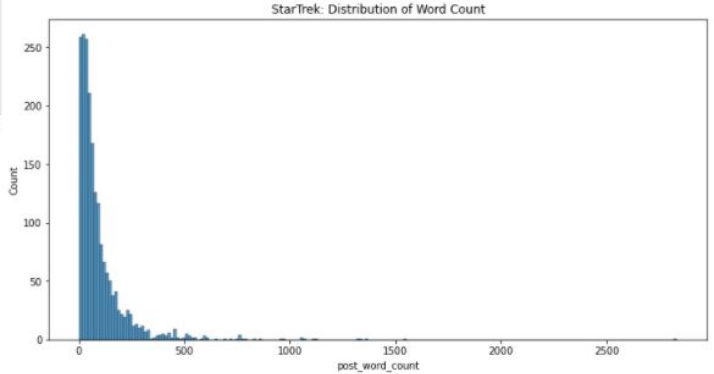
- Punctuations, white spaces, numbers and other unwanted characters were removed
- Text data converted to lower case
- Two columns for number of characters and number of words in subtext were added
- Extra data retrieved and used as Test set to try on models
- Steps taken for EDA in later sections

# EDA: Scatterplot of Post Length vs Word Count & Distribution of Word Count

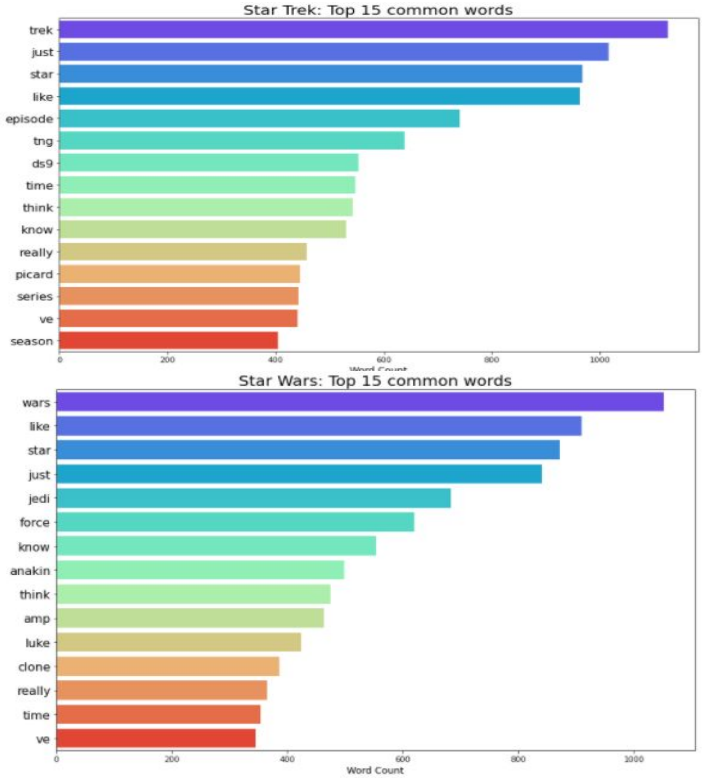



| _word_count | post_length | selftext |
|---|---|---|
| 1 | 149 | [https://www.youtube.com/watch?v=Gy3UbVi3p9o&a... |
| 1 | 166 | [https://www.facebook.com/JohndeLancie/photos/... |
| 1 | 190 | [https://hallmarkstartrekornamentsdotcom.wordp... |
| 1 | 304 | [https://prosportsextra.com/that-time-william-... |



# EDA: Top 15 Common Words & 2/3-gram




```
====== Both =======
['just', 'star', 'like', 'time', 'think', 'know', 'really', 've']
=====Star Trek=====
['trek', 'episode', 'tng', 'ds9', 'picard', 'series', 'season']
=====Star Wars=====
['wars', 'jedi', 'force', 'anakin', 'amp', 'luke', 'clone']
```

| | Count |
|---|---|
| star trek | 872 |
| https www | 129 |
| amp x200b | 94 |
| imgur com | 75 |

| | Count |
|---|---|
| star wars | 752 |
| clone wars | 292 |
| obi wan | 277 |
| amp x200b | 152 |

| | Count |
|---|---|
| https imgur com | 62 |
| youtube com watch | 34 |
| https www youtube | 34 |

| | Count |
|---|---|
| amp auto webp | 62 |
| https preview redd | 62 |
| auto webp amp | 62 |

# Pre-modeling Process

**Remove "Word"**

- html entities (& &lt)
- www, #x200B, https, youtube, com, html

**Remove "Special Characters"**

- []{}& etc

**Lemmatize or ~~Stemming~~**

- Actual words
- E.g. Marketing vs Market

```
'post_token','tokens_lem','tokens_stem'

('criticism', 'criticism', 'critic')
('intellectual', 'intellectual', 'intellectu')
('questions', 'question', 'question')
('exploration', 'exploration', 'explor')
('comes', 'come', 'come')
('references', 'reference', 'refer')
```

## Binary

Star Trek : 0 vs Star Wars : 1

## Baseline Accuracy

```
y.value_counts(normalize=True)
```

```
0    0.524798
1    0.475202
```

## CountVectorizer

- stop_word = 'english'

# Modeling Process

Naive Bayes

- Binomial Classifier
- Suitable for simple Classification
- Easily scalable

GridSearch CountVector Params

- Max_feature : 100, 250, 1000, None
- ngram _range : (1,1), (2,2), (3,3)

Random Forest

- Does not suffer from overfitting.
- Get relative feature importance

GridSearch Random Forest Params

- N_estimators : 200,400,600
- Max_depth : None, 5, 10

# Evaluation

## Cvec with Naive Bayes

```
Best Score : 0.9306306306306306
Train Score : 0.9851351351351352
Test Score : 0.9406207827260459
```

| | ngram_range | max_feature | mean_score |
|---|---|---|---|
| **9** | (1, 1) | None | 0.930631 |
| **6** | (1, 1) | 1000 | 0.914865 |
| **3** | (1, 1) | 250 | 0.884685 |
| **10** | (2, 2) | None | 0.860360 |
| **0** | (1, 1) | 100 | 0.846847 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.96 | 0.93 | 235 |
| 1 | 0.96 | 0.91 | 0.94 | 265 |
| accuracy | | | 0.93 | 500 |
| macro avg | 0.93 | 0.94 | 0.93 | 500 |
| weighted avg | 0.94 | 0.93 | 0.93 | 500 |

## Cvec with Random Forest

```
Best Score : 0.9202702702702703
Train Score : 0.9405405405405406
Test Score : 0.9176788124156545
```

| | ngram_range | max_feature | n_estimators | max_depth | mean_score |
|---|---|---|---|---|---|
| **88** | (1, 1) | None | 400 | 10 | 0.920270 |
| **89** | (1, 1) | None | 600 | 10 | 0.918018 |
| **86** | (1, 1) | None | 600 | 5 | 0.913063 |
| **85** | (1, 1) | None | 400 | 5 | 0.912613 |
| **87** | (1, 1) | None | 200 | 10 | 0.911261 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.83 | 0.89 | 235 |
| 1 | 0.87 | 0.98 | 0.92 | 265 |
| accuracy | | | 0.91 | 500 |
| macro avg | 0.92 | 0.90 | 0.91 | 500 |
| weighted avg | 0.92 | 0.91 | 0.91 | 500 |

# Summary & Recommendations

Accurate but not all powerful. Able to reduce workload of users.

“Saw it last night and I wa very meh about it after There was some really funny moment and some awesome action scene and a few holy shit moment But the pacing wa all over the place the storyline itself seemed last minute slapped together and I didnt get emotionally involved in almost any of it Despite the fact that I went into it extremely excited to see it”

Could make use of information to direct advertisement to related subreddit.

Identify problem users, users who post irrelevant content on the site.