WEB APIS & CLASSIFICATION

VS

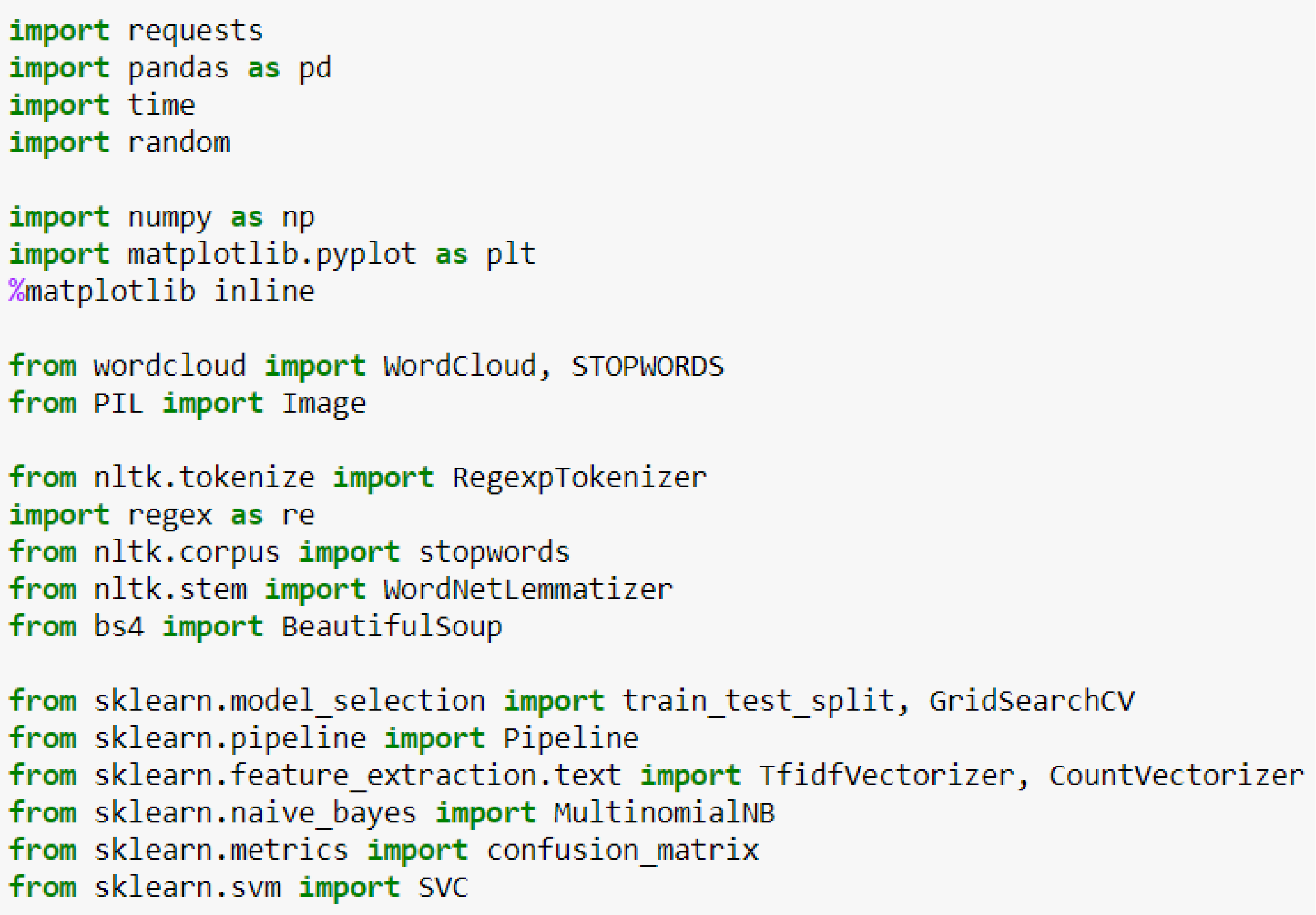
INTRODUCTION

- The first Pokemon game was dropped in 1996 and the main series of the Pokemon games is into its 8th generation
- Including other spin-offs as well as mobile apps and PC games, there are a few hundred Pokemon games today.
- Online discussions provide valuable information to developers for future updates and new game developments.
- The first step of analyzing such discussions for insights would be classifying such discussions with regards to which Pokemon game it is talking about.

OUTLINE

- 975 posts are scrapped from each of the respective sub-reddits for Pokemon Go and Pokken.
- The dataset underwent exploration and cleaning to be made suitable for model fitting.
- Count Vectorizer and Tfidf Vectorizer were used with Naive Bayes classifier in a pipeline to Gridsearch for the best parameters.
- Best performing Vectorizer was then used to compare the Naive Bayes classifier against the Support Vector classifier to develop an overall best model.

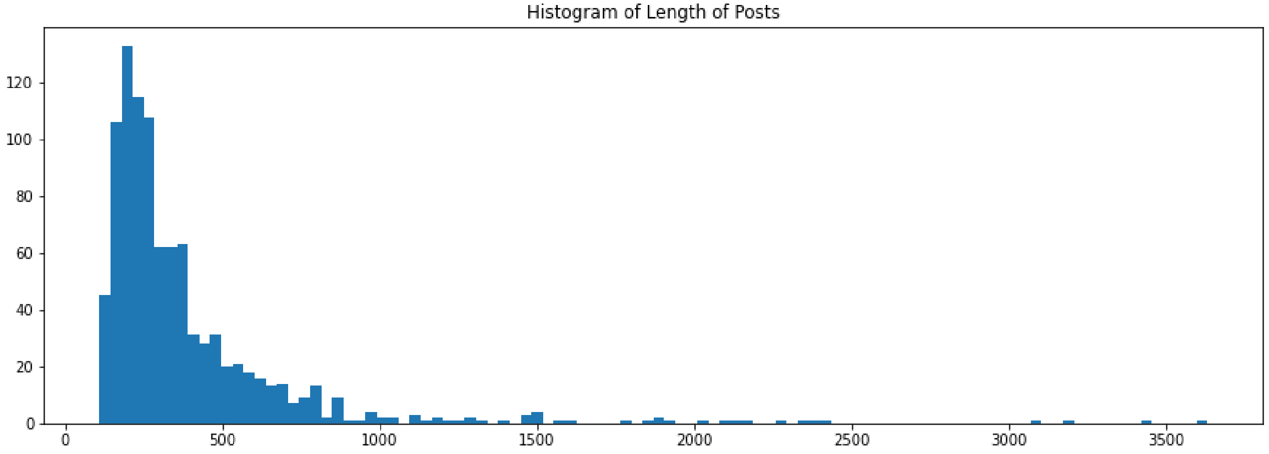
TOOLS & LIBRARIES USED



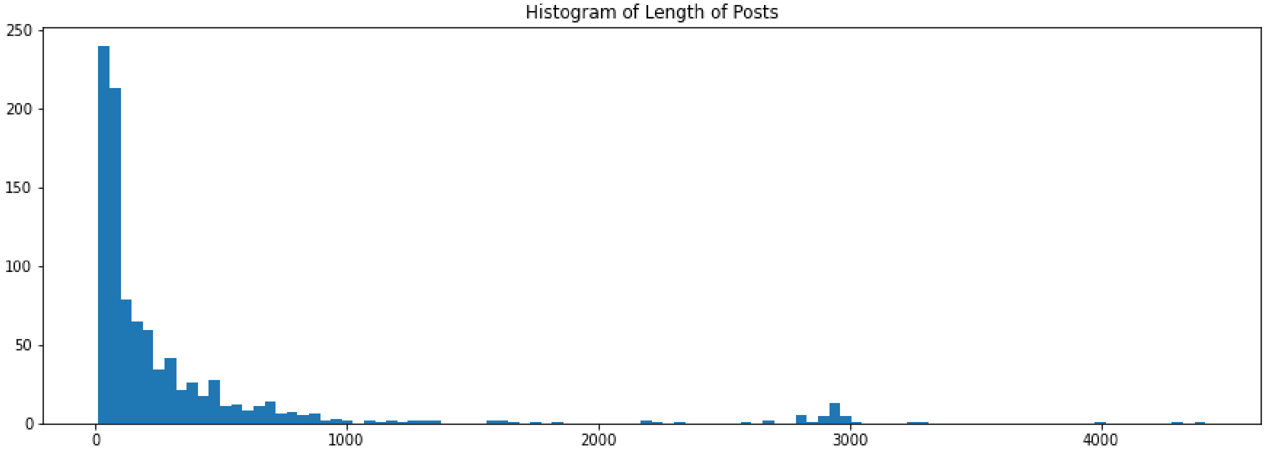
```
import requests
import pandas as pd
import time
import random

import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

from wordcloud import WordCloud, STOPWORDS
from PIL import Image

from nltk.tokenize import RegexpTokenizer
import regex as re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from bs4 import BeautifulSoup

from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import confusion_matrix
from sklearn.svm import SVC
```

LENGTH OF POSTS

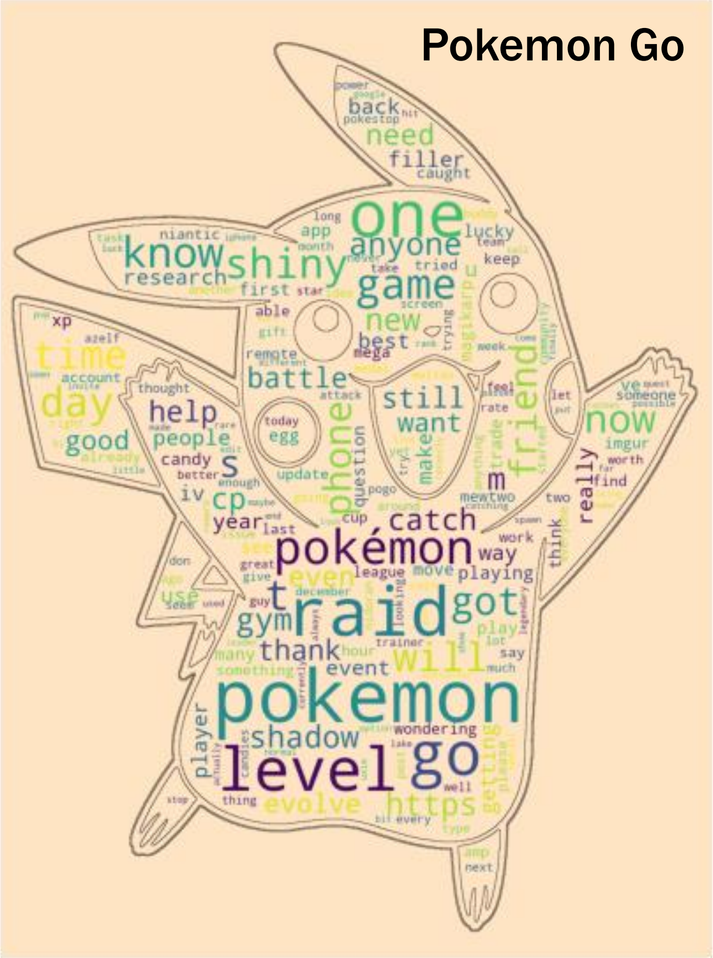
Pokemon Go



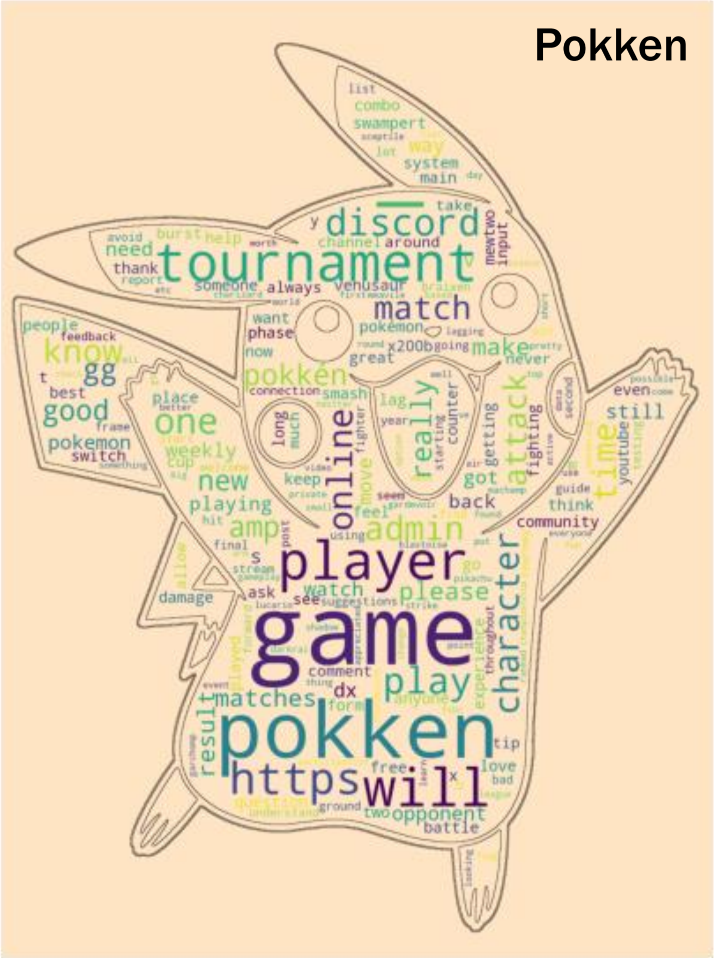
Pokken



POPULAR WORDS



Pokemon Go



Pokken

CLEANING & ESTABLISHING BASELINE

- Used beautiful soup to remove HTML
- Regex to remove non-letters
- Stopwords as well as words associated with title of sub-reddit removed
- Converted all words to lower case
- Applied lemmatization
- Baseline score after train test split: 51.59%

COUNT VECTORIZER & NB CLASSIFIER IN PIPELINE FOR GRIDSEARCH

- **Best parameters:**

```
‘cvec__max_features’: 3000,
‘cvec__ngram_range': (1, 1)
```

- **Score on train data: 97.88%**
- **Score on test data: 95.50%**

TFIDF VECTORIZER & NB CLASSIFIER IN PIPELINE FOR GRIDSEARCH

- **Best parameters:**

```
'tvec__max_features': 2000,
'tvec__min_df': 3,
'tvec__ngram_range': (1, 1)
```

- **Score on train data: 97.95%**
- **Score on test data: 96.03%**

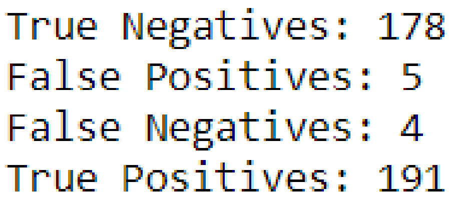
TFIDF VECTORIZER & SV CLASSIFIER IN PIPELINE FOR GRIDSEARCH

- **Best parameters:**

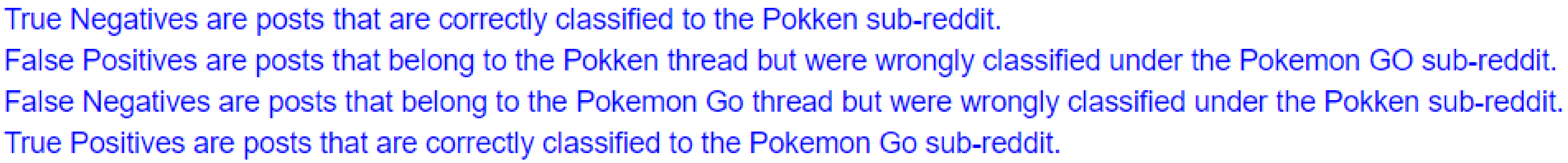
```
'svc__C': 0.5,
'svc__kernel': 'sigmoid’,
'svc__max_iter': 1100,
'tvec__min_df': 2,
'tvec__ngram_range': (1, 2)
```

- **Score on train data: 99.07%**
- **Score on test data: 97.62%**

RESULTS



```
True Negatives: 178
False Positives: 5
False Negatives: 4
True Positives: 191
```



True Negatives are posts that are correctly classified to the Pokken sub-reddit.

False Positives are posts that belong to the Pokken thread but were wrongly classified under the Pokemon GO sub-reddit.

False Negatives are posts that belong to the Pokemon Go thread but were wrongly classified under the Pokken sub-reddit.

True Positives are posts that are correctly classified to the Pokemon Go sub-reddit.

Misclassified posts generally were observed to have content that are rather general in nature and containing common words. They were then misclassified since some of these words within were more strongly associated with the wrong sub-reddit based on the classifier model.

CONCLUSION

- TFIDF Vectorizer and Support Vector Classifier gave the best score:
  - ➢ Score on train data: 99.07%
  - ➢ Score on test data: 97.62%
  - ➢ Posts are highly distinguishable between the two sub-reddits
- Fulfils purpose of being the first step in classifying posts for further analysis of online discussions regarding the two games

RECOMENDATIONS

- Model built may be used to classify posts from other forums that may be talking about these games as well.
- Model is a binary classifier which is rather limited.
  - ➢ New model classifying posts into the various Pokemon game titles across the various generations and platforms would be more useful
- For project purposes, words associated with the title of the sub-reddits were removed.
  - ➢ In an actual classifier model, it would be better to keep such words since they would be the clearest identifiers possible.

Q&A