

INTRODUCTION

Project 3 Web APIs & Classification

Mubin Khan



Goal

- Define the problem
- Obtain the data
- Explore the data
- Model the data
- Evaluate the model
- Respond to the problem

1. Using Pushshift's API, the aim is to collect posts from any two subreddits.
2. Use NLP(Natural Language Processing) to train a classifier on which subreddit a given post came from.

AUSTRALIA

POST


A person's arm and hand are shown lifting a dumbbell against a dark background.

Define the problem

Can we leverage on machine learning to create a classifier that can accurately predict the origins of reddit posts?

Subreddit 1

Bodyweight fitness

A large, detailed gold Bitcoin coin is shown, featuring the Bitcoin symbol and binary code around its edge, set against a background of circuitry.

Subreddit 2

Crypto_com

- Define the problem
- Obtain the data
- Explore the data
- Model the data
- Evaluate the model
- Respond to the problem

Why I choose Crypto_com



familiarity



community



Volatile
market

- Define the problem
- Obtain the data
- Explore the data
- Model the data
- Evaluate the model
- Respond to the problem

Why I choose bodyweight fitness



keep fit



community



diverse

Image source:
<https://www.menshealth.com/uk/building-muscle/a756325/10-best-bodyweight-exercises-for-men/>

- Define the problem
- **Obtain the data**
- Explore the data
- Model the data
- Evaluate the model
- Respond to the problem



Crypto_com

https://api.pushshift.io/reddit/search/submission/?subreddit=Crypto_com (Submission)

b) https://api.pushshift.io/reddit/search/comment/?subreddit=Crypto_com (Comments)



Bodyweight fitness

a) <https://api.pushshift.io/reddit/search/submission/?subreddit=bodyweightfitness> (Submission)

b) <https://api.pushshift.io/reddit/search/comment/?subreddit=bodyweightfitness> (Comments)

- Define the problem
- **Obtain the data**
- Explore the data
- Model the data
- Evaluate the model
- Respond to the problem



Crypto_com

10,000 comments



Bodyweight fitness

9,898 comments

- Define the problem
- Obtain the data
- Explore the data
- Model the data
- Evaluate the model
- Respond to the problem

Why I choose comments over posts

images



Crypto_com

text



Bodyweight fitness

Approximately **10%**
of data was reduced

- Define the problem
- Obtain the data
- **Explore the data**
- Model the data
- Evaluate the model
- Respond to the problem

Data Cleaning



Dropped duplicates



Removed html, hyperlinks, punctuation

Image source:
<https://www.menshealth.com/uk/building-muscle/a756325/10-best-bodyweight-exercises-for-men/>

- Define the problem
- Obtain the data
- **Explore the data**
- Model the data
- Evaluate the model
- Respond to the problem

Preprocessing

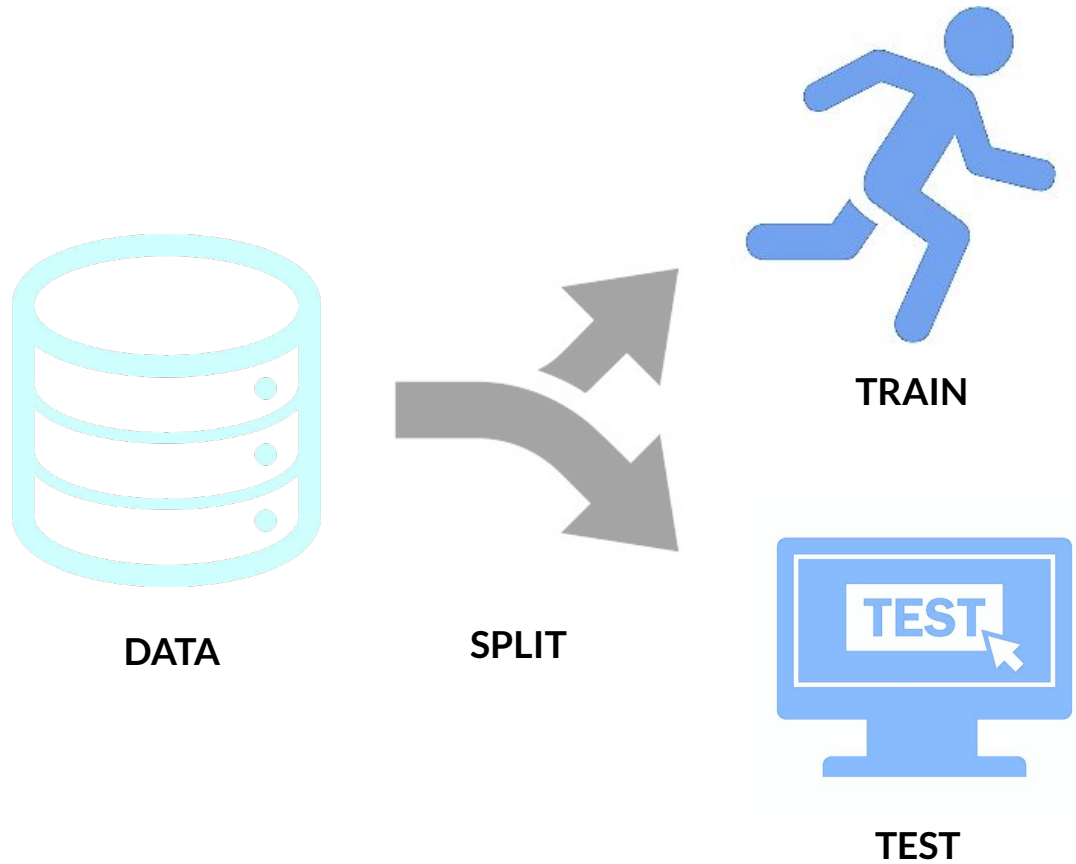
- Lemmatization (cards >> card, banks>>bank, boxes>> box)
- Added to stop words: 'com','don','got','ha','wa','going'

Image source:
<https://www.kenwoodworld.com/en-gb/inspiration/articles/food/8-recipes-food-processor>

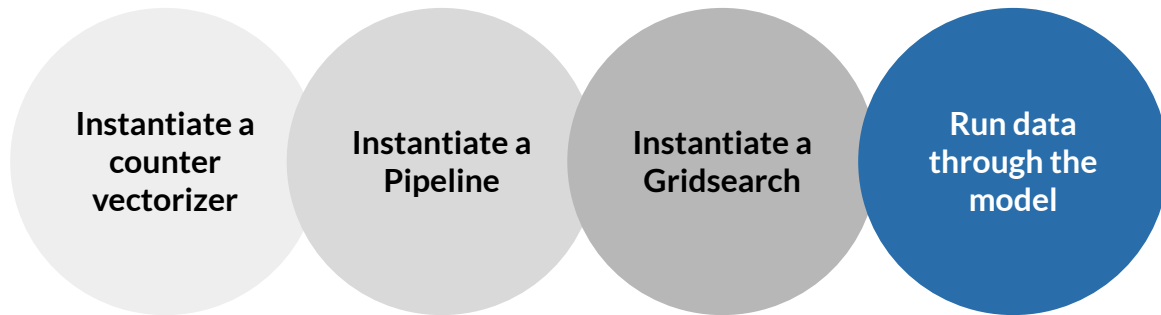
- Define the problem
- Obtain the data
- **Explore the data**
- Model the data
- Evaluate the model
- Respond to the problem



- Define the problem
- Obtain the data
- Explore the data
- **Model the data**
- Evaluate the model
- Respond to the problem



- Define the problem
- Obtain the data
- Explore the data
- **Model the data**
- Evaluate the model
- Respond to the problem



- Define the problem
- Obtain the data
- Explore the data
- Model the data
- **Evaluate the model**
- Respond to the problem

Train Accuracy Score

96.32

Test Accuracy Score

91.29

Best Model

**Logistic
Regression**

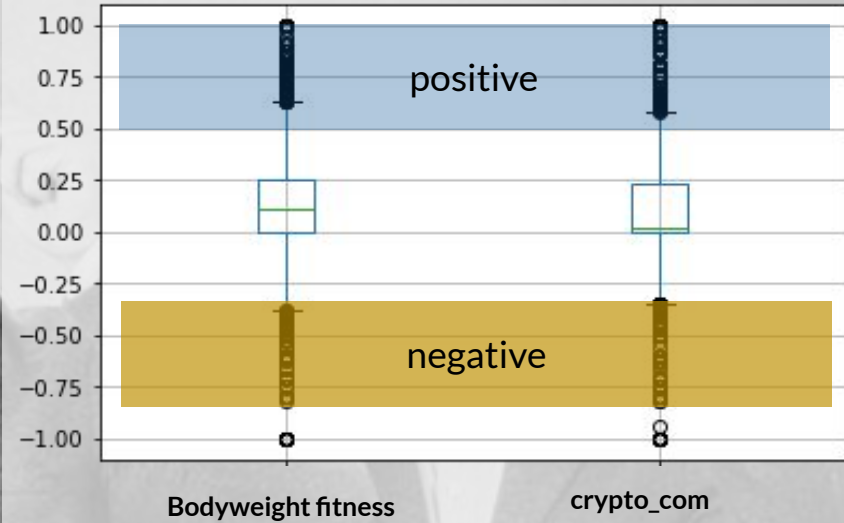
Best Max Features

10,000

Image source:
<https://www.questionpro.com/blog/respondent-sentiment-analysis/>



Boxplot grouped by SENTIMENTS



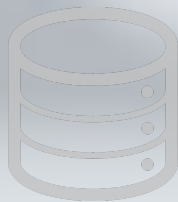
SENTIMENTAL ANALYSIS

Good balance of positive and negative
Diverse group of participants -from amateur to seasoned

Conclusion

Logistic regression model - achieved the best scores
(train / test scores: 0.9632 / 0.9129).

Potential improvements for future



More training data



More data cleaning &
preprocessing



More intensive
gridsearching

- Define the problem
- Obtain the data
- Explore the data
- Model the data
- Evaluate the model
- Respond to the problem