Corey
Duncan

# Barbells or Broccoli?

## Web Scraping Project

Fitness Vs Nutrition

# Outline



# Problem Statement

Can we build a model that is able to tell the subreddit origin of a given post between two subreddits, using only the title?

# Objective

Develop a machine learning model that accurately predicts the subreddit category of a given Reddit post based solely on its title.

This model aims to assist in automating the categorization of posts, for use in marketing and advertising.

# Collection Process

## Fitness

- Two subreddits of choice are fitness and nutrition
- Scraped each subreddit 7 times across different days

## Nutrition

# Collection Process

## Month/Year

## Top/New

## Controversial

# Pre-Processing

## 01

### Creating Y

Added subreddit of origin column to my fitness and nutrition datasets

## 02

### Combine

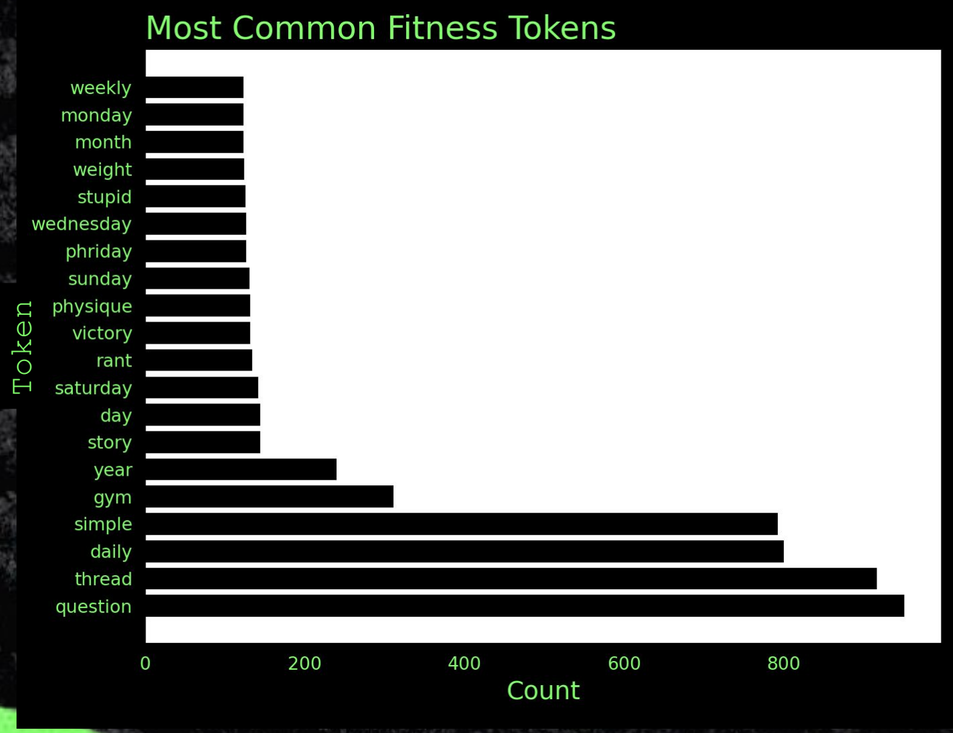
Combined my nutrition and fitness datasets into one large dataframe

## 03

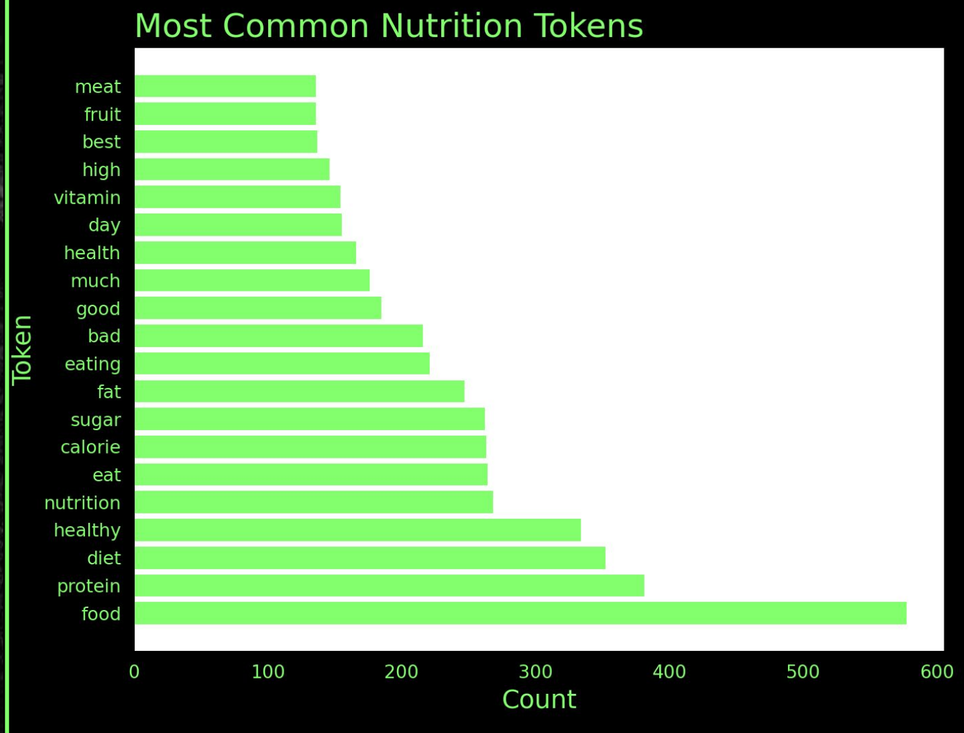
### Tokenize

Created  function return a new title column that was tokenized, lemmatized, lowercase, with removed punctuation and stop words

# Most Common Words




- Almost no matching words in the top 20 except for ‘day’
- Most common fitness words mostly focused on time related words
- Most common nutrition words focused on nutrients
- Most common words are ‘food’ and ‘question’

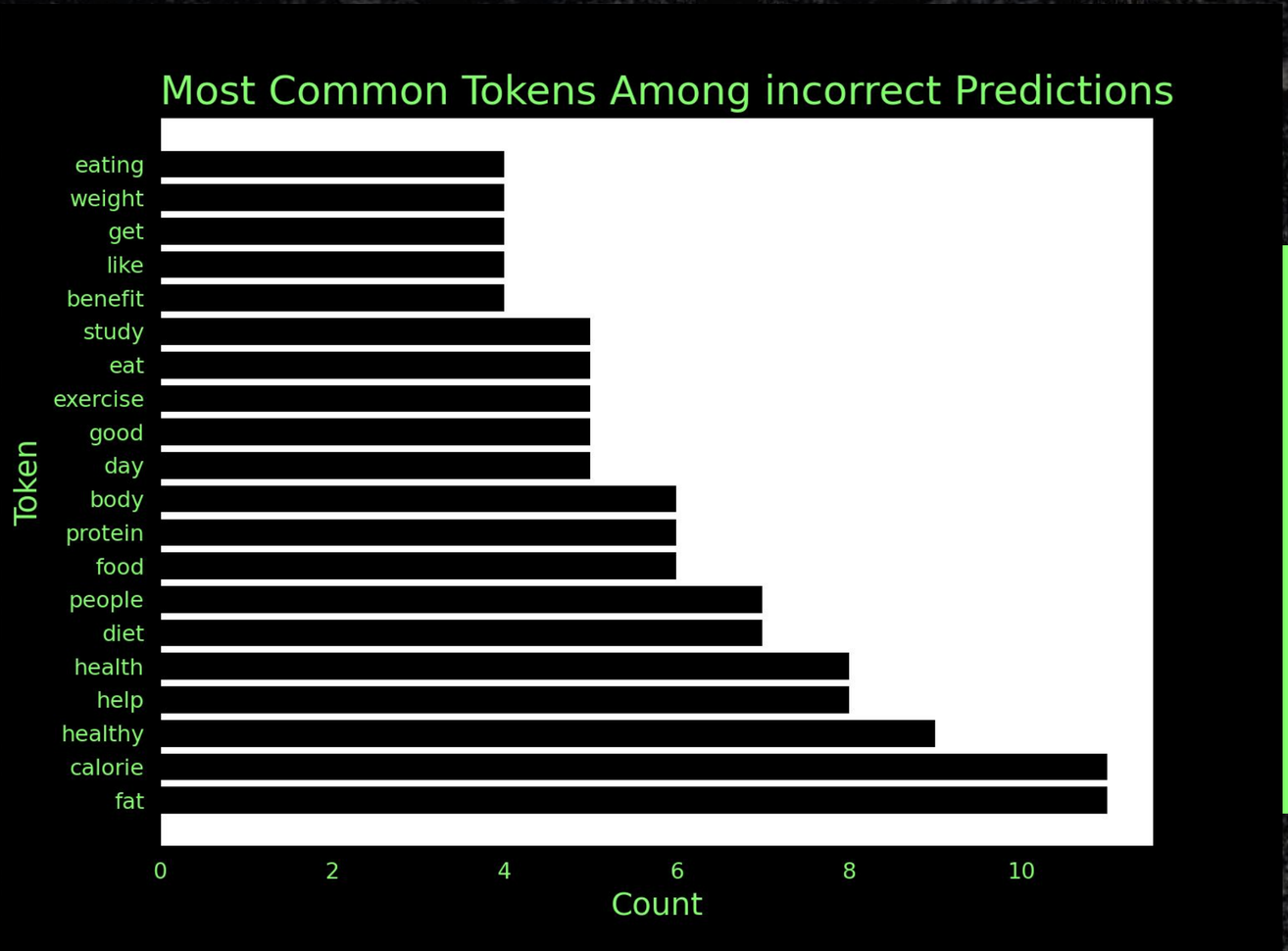


# Model 1 : CVEC

Train .957

Test .936

- No crossfold, max iterations or gridsearch
- Needed to beat a baseline accuracy of .557




## Confusion Matrix

| | Predicted Fitness | Predicted Nutrition |
|---|---|---|
| **Actual Fitness** | **646** | **81** |
| **Actual Nutrition** | **24** | **892** |

# Model 2: TF-IDF

Train .957

Test .928

- Lower performance
- None of my tokens had high TF-IDF scores
- Tried bigrams and limiting features

| Word | tfidf_score |
|---|---|
| question | 0.042156 |
| thread | 0.040173 |
| daily | 0.038872 |
| simple | 0.037905 |
| food | 0.018215 |
| protein | 0.015476 |
| gym | 0.013722 |
| diet | 0.013691 |
| healthy | 0.013351 |
| calorie | 0.011922 |

## Confusion Matrix

|  | Predicted Fitness | Predicted Nutrition |
|---|---|---|
| Actual Fitness | 629 | 98 |
| Actual Nutrition | 20 | 896 |

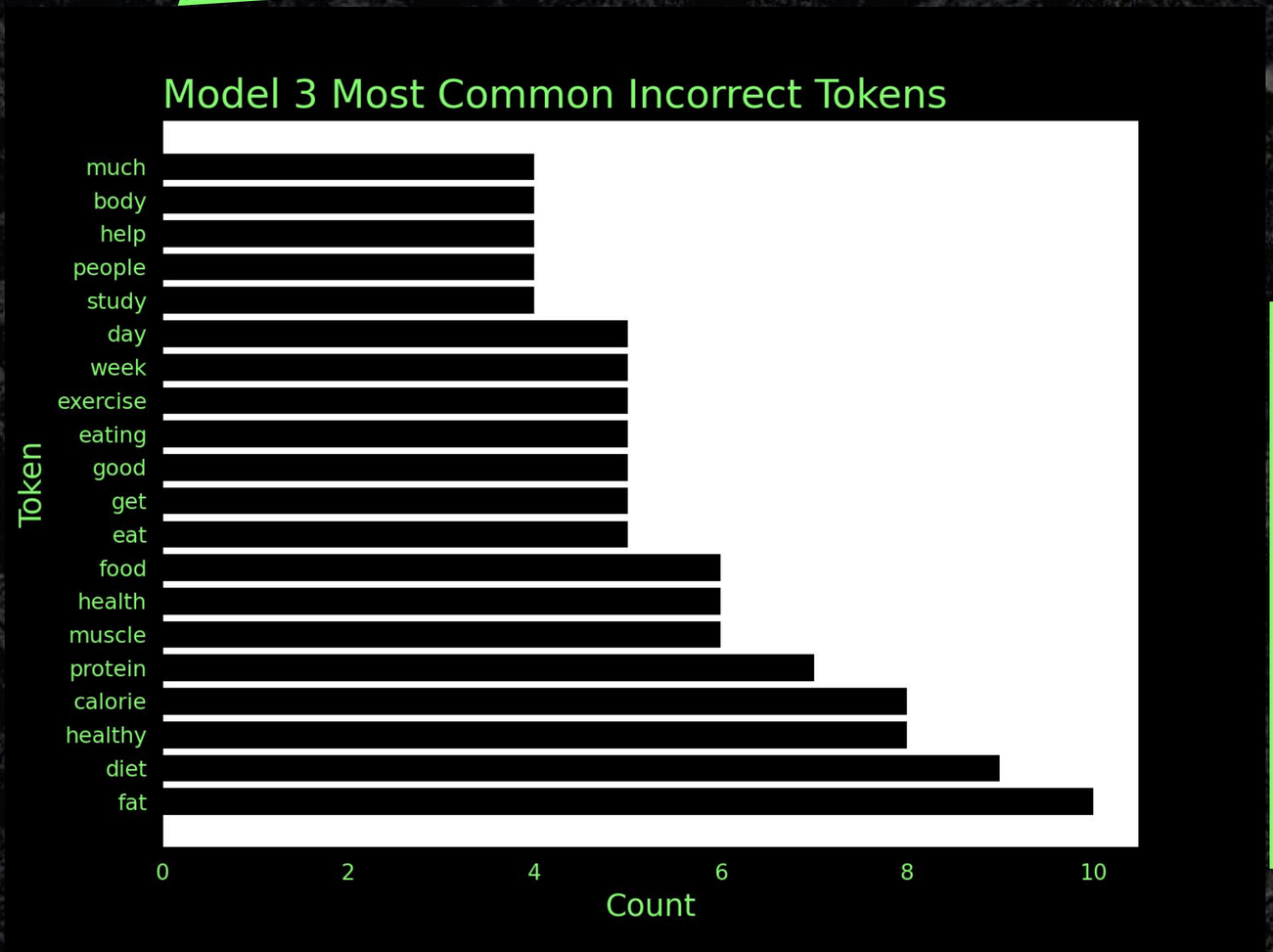
# Model 3: CVEC

Train .995

Test .939

## Bigram & GridSearch

- Highest test score, highest variance
- Also tried trigrams
- Incorrect tokens were all single words




## Confusion Matrix

|  | Predicted Fitness | Predicted Nutrition |
|---|---|---|
| Actual Fitness | 666 | 61 |
| Actual Nutrition | 40 | 876 |

# Conclusions

## Conclusion 01

The best model I was able to build was one that used CountVectorizer as it had higher scores than Model 2 and less variance than Model 3. TF-IDF, Bigrams and Trigrams seemed to hurt my model in this case.

## Conclusion 02

The project successfully developed a machine learning model capable of accurately predicting the subreddit of origin for posts related to fitness and nutrition. This model can be used for various applications, such as content filtering, targeted advertising, and enhancing user experience on Reddit

# Next Steps

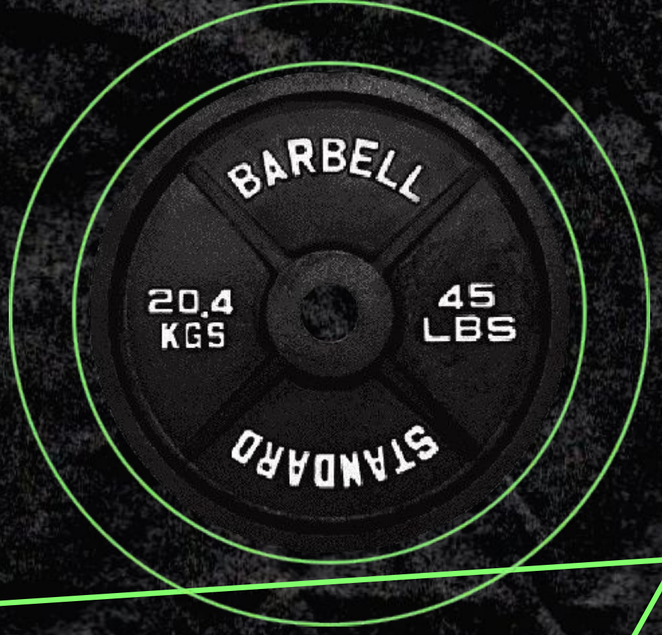
If I were to continue working on this project, I might try working on these ideas next:

Remove most common incorrect words to see if it improves the model

Random Forests

Study Neural Networks

# Works Cited




Fitness Subreddit: https://www.reddit.com/r/Fitness/
Nutrition subreddit: https://www.reddit.com/r/nutrition/
Reddit PRAW API: https://www.reddit.com/prefs/apps
PRAW Documentation: https://praw.readthedocs.io/en/latest/code_overview/models/subreddit.html#praw.models.Subreddit.top

# Thank you!

Questions?