Project 3: Web APIs & NLP

Sophia Joseph

Table of contents

01

Problem Statement

Both fitness and nutrition are an important part of living a healthy lifestyle. Although the two can overlap in many ways, there can also be many distinguishable features about them individually. The primary goal of this project is to identify whether words are found in fitness or nutrition subreddits and to base the success off of the two model's f1 scores.

02

f1 score

What is an f1 score?

- Gives an accuracy on how many times your model gave a True Positive prediction
- Puts the precision score and recall score of a model together
  - Precision score -> number of correctly classified positive predictions
  - Recall score -> number of correctly classified actual positive predictions out of all the positive predictions
- f1 score chosen for this project to see how many times our models pick the correct subreddit for the specific word

03

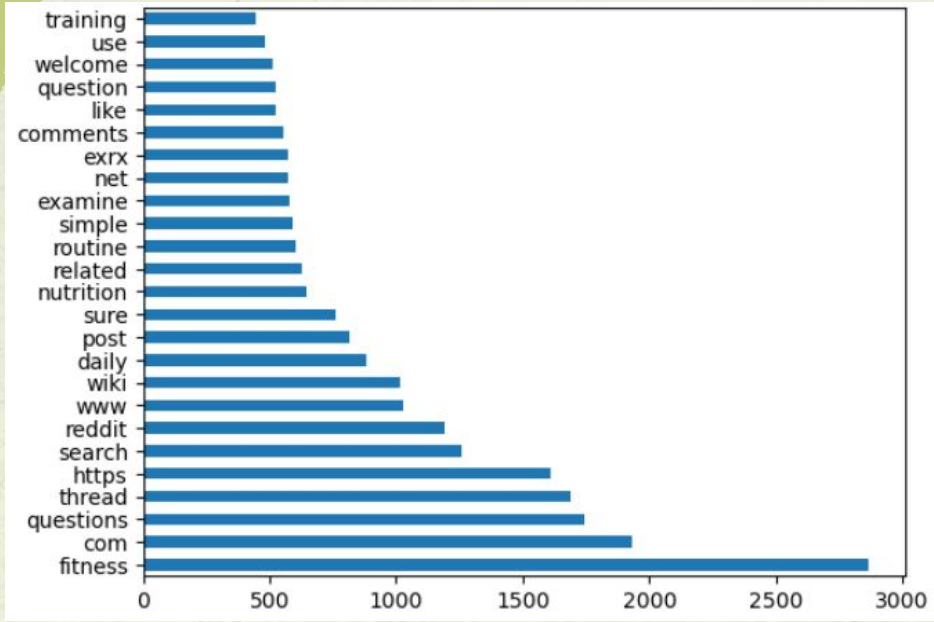
Scraping & Preprocessing

Scraping & Preprocessing

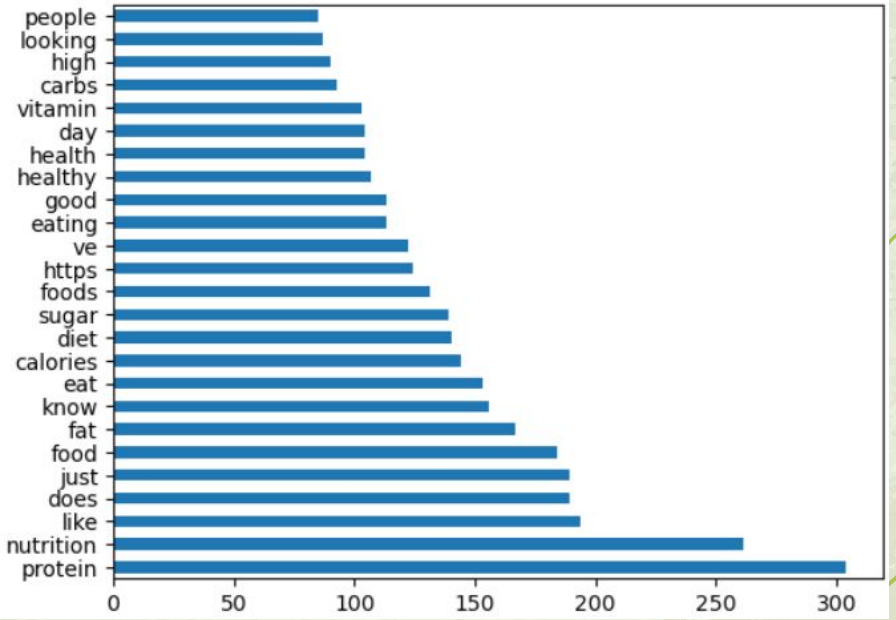
1) The Reddit API was used to scrape reddit of 1000 posts
    - Fitness
    - Nutrition
2) Preprocessing was done in a custom function
    - Tokenzing
    - Lemmatizing
    - Changing text to lowercase
    - Remove unnecessary characters
3) Count Vectorizing was performed
    - Converts text into a count of how many times a word is in a document

Most Common Words

Fitness



Nutrition



04
Models

Models

Random Forest Classifier

- Used to gather predictions from various sets of training data and put them together

Logistic Regression

- Used to distinguish whether the data belongs to the positive class

05

Metrics

	Precision	Recall	f1 score	Baseline Score
Random Forest Classifier (Fitness)	1.0	0.9016	0.9482	-
Logistic Regression (Fitness)	0.9917	0.9370	0.9636	-
Random Forest Classifier (Nutrition)	0.9004	1.0	0.9476	-
Logistic Regression (Nutrition)	0.9333	0.9912	0.9614	-
Fitness	-	-	-	0.5044
Nutrition	-	-	-	0.4956

06

Conclusions & Recommendations

Conclusions & Recommendations

Model performance

Both models performed very well on identifying subreddit various words came from

Baseline vs f1 score

The f1 scores all did better than the baseline score

Which model did better?

Logistic Regression model did slightly better (higher f1 score)

Next steps

More columns can be pulled from reddit to add more data

Next steps

Another model could be added to test more scores

Resources

- https://www.v7labs.com/blog/f1-score-guide#:~:text=for%20Machine%20Learning-,What%20is%20F1%20score%3F,prediction%20across%20the%20entire%20dataset.

Thanks

Do you have any questions?

Instructions for use

If you have a free account, in order to use this template, you must credit **Slidesgo** by keeping the **Thanks** slide. Please refer to the next slide to read the instructions for premium users.

As a Free user, you are allowed to:

- Modify this template.
- Use it for both personal and commercial projects.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute Slidesgo Content unless it has been expressly authorized by Slidesgo.
- Include Slidesgo Content in an online or offline database or file.
- Offer Slidesgo templates (or modified versions of Slidesgo templates) for download.
- Acquire the copyright of Slidesgo Content.

For more information about editing slides, please read our FAQs or visit our blog:
https://slidesgo.com/faqs and **https://slidesgo.com/slidesgo-school**