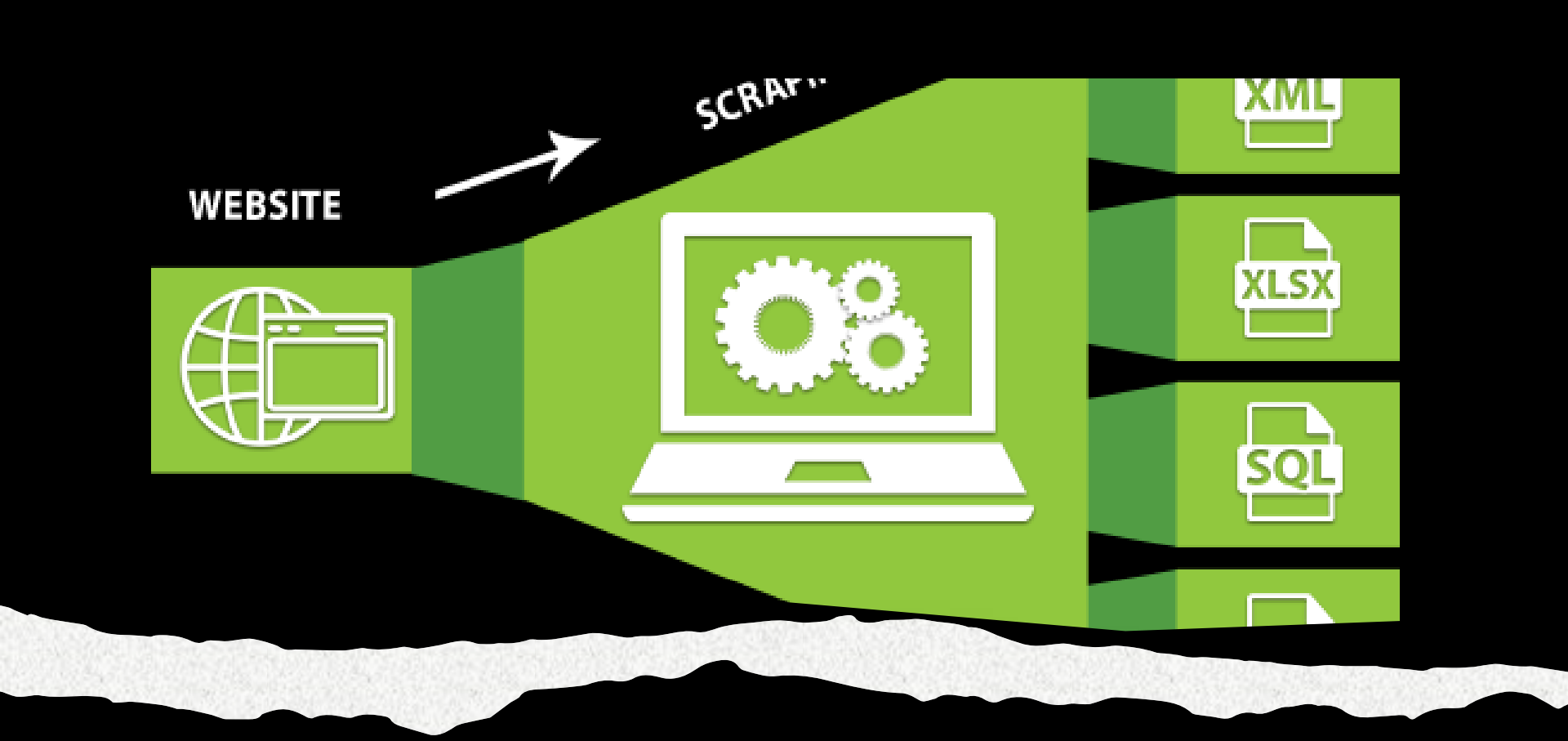


Project 3 Web APIs and NLP, Predicting Subrredit

Contents:

Problem Statement

- Predicting whether the post came from CARS subreddit or from Motorbikes subreddit.

Audience:
Companies that sell cars and motorbikes products

Cars :Toyota, Honda, GMC Sierra, Ford, etc.

Motorbikes: Nexus, Royal Enfield, Java, Yamaha, etc.

What is reddit?

- Redditt is a massive collection of forums where people can share news and content or comment on other people's posts.
- Reddit is broken up into more than a million communities known as "subreddits,"
- Registered members submit content to the site such as links, text posts, images, and videos.

Data Collection and Cleaning

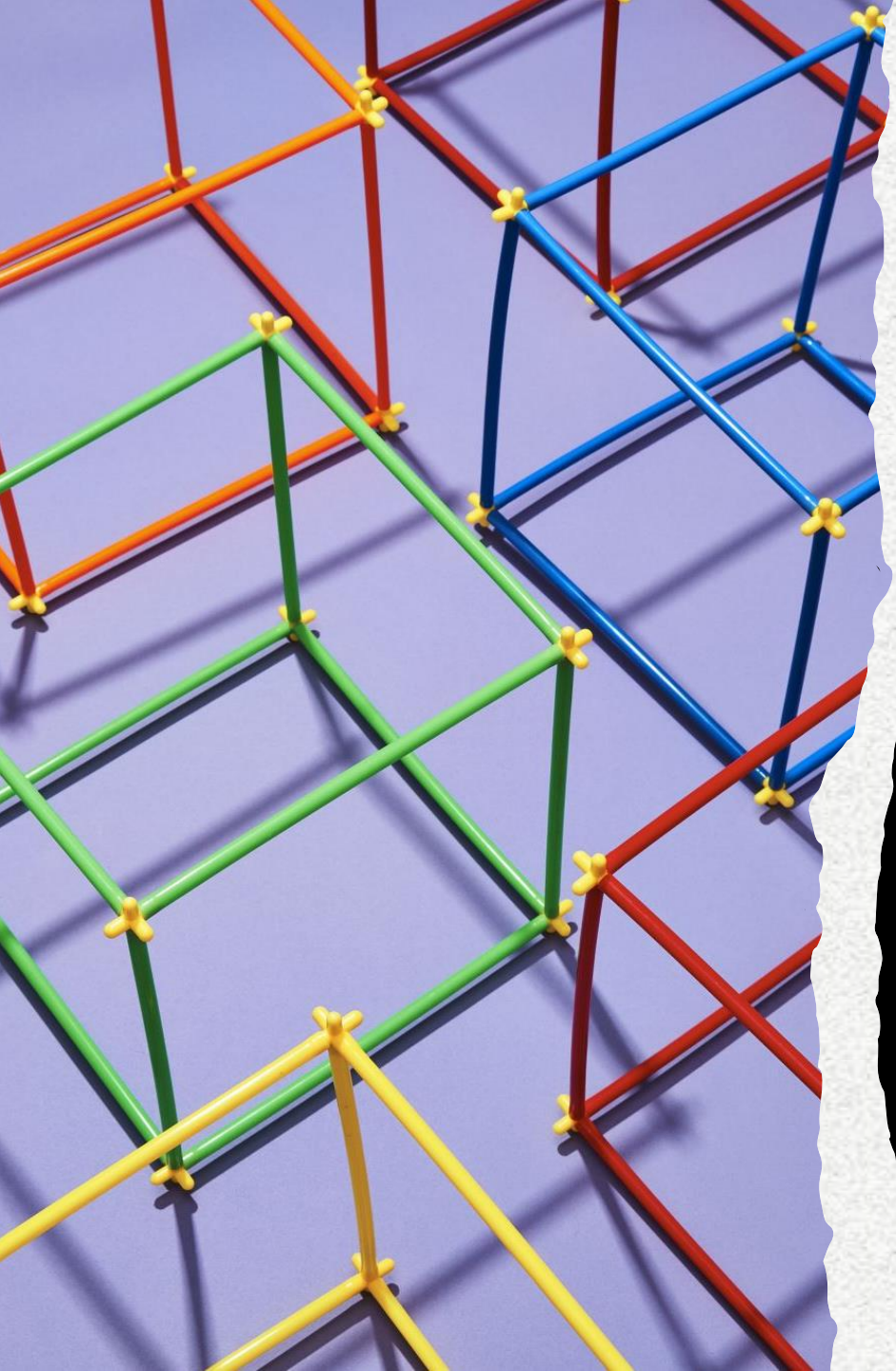
- **API Scrapping**

  Scrap the desired reddit posts using api

- **Extract the needed data**

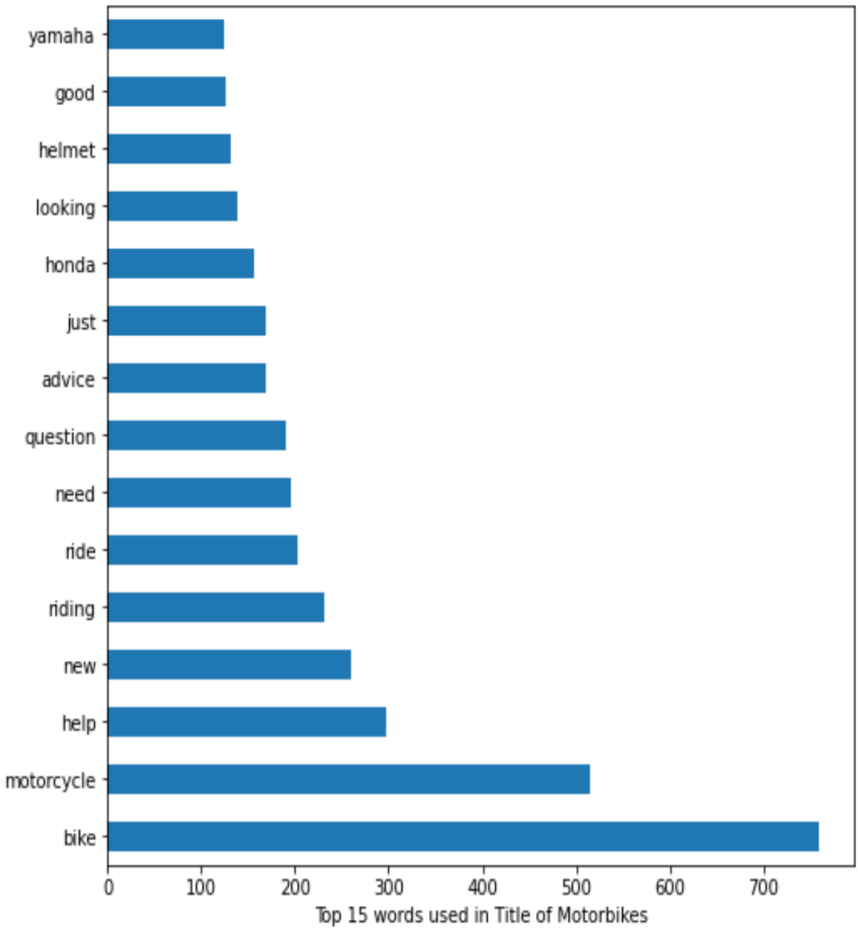
  Title and subreddit

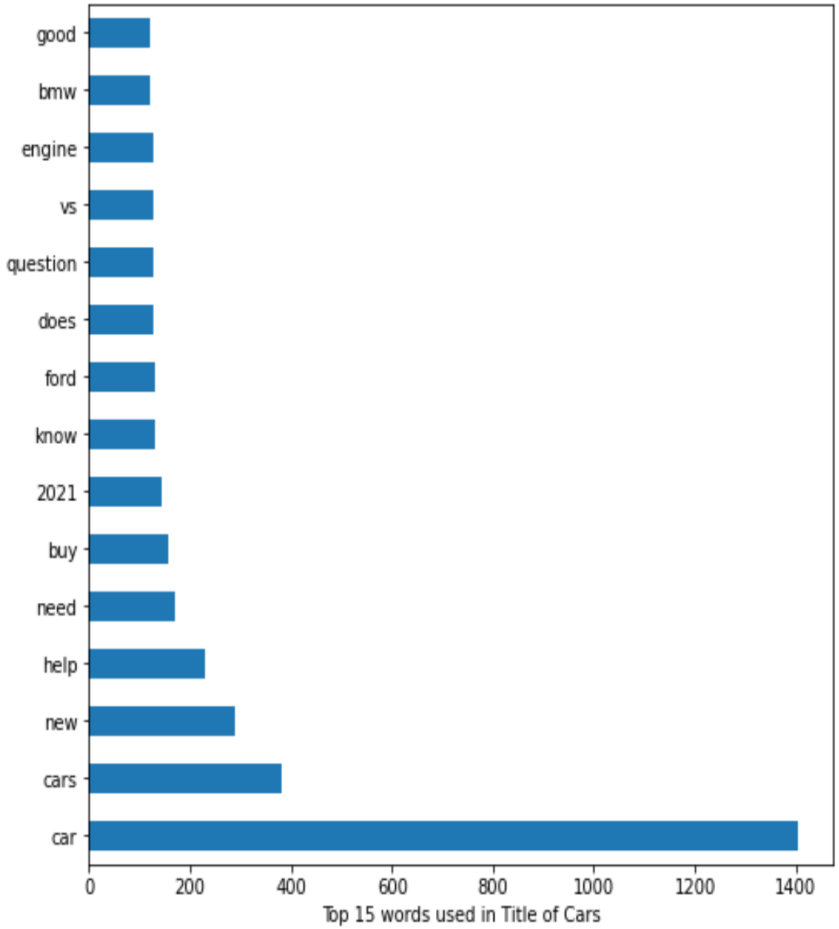
- **Data cleaning:**
  - Drop duplicate
  - Convert subreddit to 0 (cars) and 1 (motorcycles)



- Preprocessing :
  1) Removal of special characters
  2) Tokenizing
  3) Lemmatizing
  4) Stopwords
- **Modeling:**
  1) Random Forest
  2) Logistic Regression

Top 15 words in Title for both Cars & Motorbike





Model Evaluation

OR

1) Random Forest

Train Accuracy:0.82608

Test Accuracy: 0.76876

2) Logistic Regression

Train Accuracy:0.78270

Test Accuracy: 0.77284

Most Popular word •**Autopilot**

Conclusions

- **Random Forest Model gave slightly better accuracy with the training data.**
- **The accuracy classification is fine but not that good.**
- **The main reason is similarity of subreddits.**
- **Many characteristics of cars and motorcycles are same that makes similarity inn subreddits.**

Recommendations

This code is used to classify any two subreddits by changing the names while retrieving data.

To get a better result, min_df should be ignored.

The performance of models can change depending on selection of subreddits, number of posts, and way of processing the data.

Thank You!