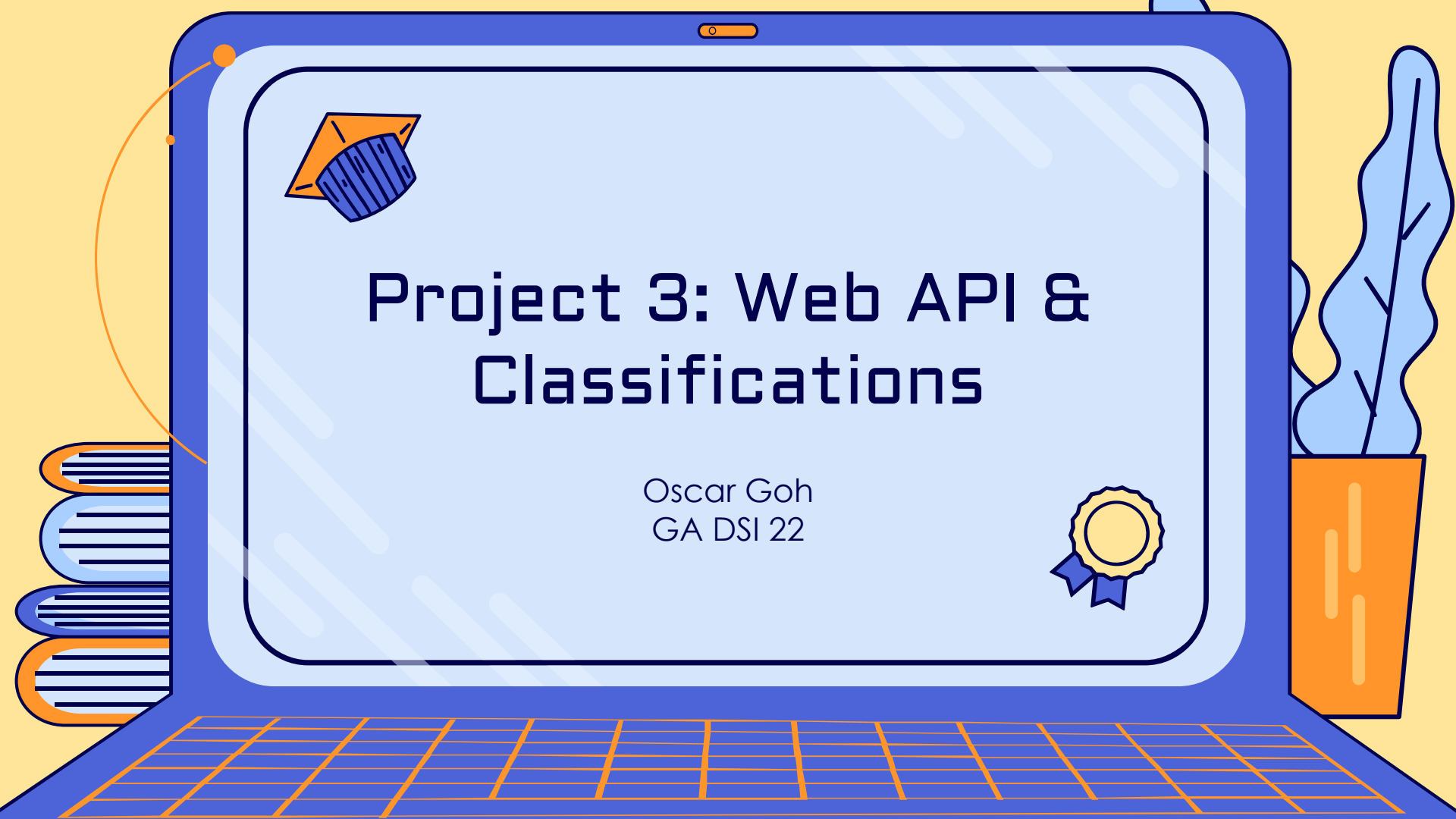

Project 3: Web API & Classifications
Oscar Goh
GA DSI 22

TABLE OF CONTENTS

01

Introduction

Background & objective

02

Data

Collection, cleaning & preprocessing

03

Modeling

Logistic regression & Naïve bayes

04

Conclusion

Analysis & recommendation

01

Introduction

Background & objective

reddit

Background

- Social news discussion website
- Contents posted by registered members
- Upvote & downvote
- More upvotes appear towards the top

Objective

Collection

Collect posts from 2 subreddits using Reddit's API

Prediction

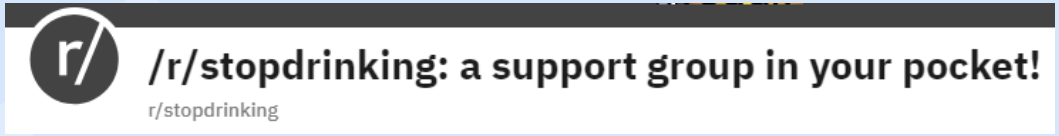
Use NLP to train a classifier on which subreddit a given post came from

02

Data

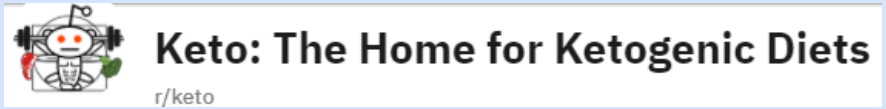
Collection, cleaning & preprocessing

Data collection



Stopdrinking

A place to motivate each other to control or stop drinking



Keto

A place to share experiences around eating within a Ketogenic lifestyle

1989

Total number of posts scraped

1737

Total number of unique posts

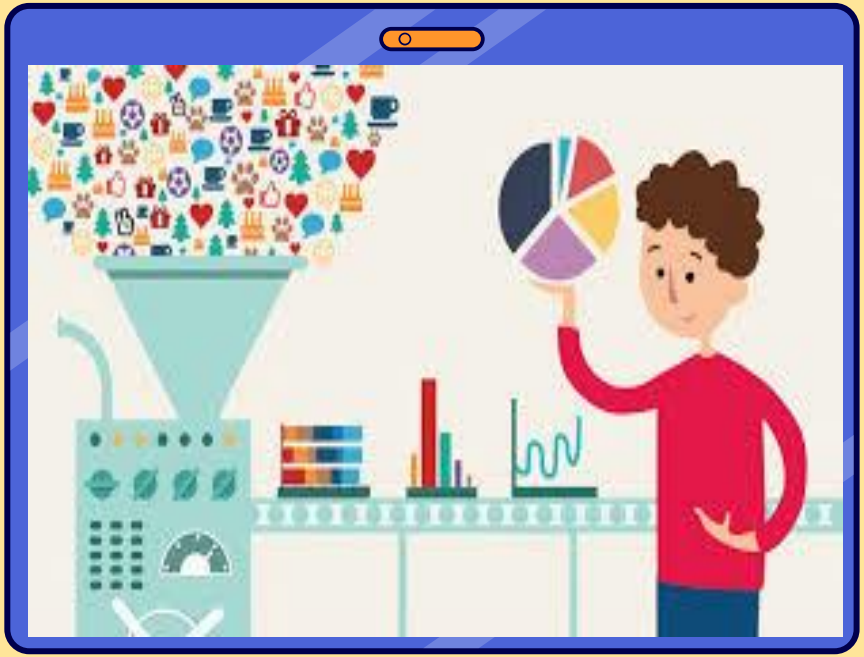
742

From Keto

995

From Stopdrinking

Cleaning & preprocessing



- Remove Url, non-letters
- Split into individual lower case words
- Remove stopwords
- Remove subreddit and words closely related

03

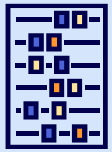
Modeling

Logistic regression & Naïve bayes

Models & Transformers

Log reg with count vec

Train score	Test score
0.9946	0.9609



Naïve with count vec

Train score	Test score
0.9800	0.9770

Log reg with Tfidf vec

Train score	Test score
1.0	0.9586

Naïve with Tfidf vec

Train score	Test score
0.9777	0.9494

04

Conclusion

Analysis & recommendation

Collection

1989 posts out of an intended 2000 and getting 1737 unique posts for analysis

Prediction

2 models are successfully fitted, tuned and tested. Best model: Naïve bayes with count vectorizer at a test score of 0.977.

THANKS!

Do you have any questions?
oscargsh@yahoo.com.sg
GA DSI 22