Web APIs & Classification

Kevin Roesch
Data Scientist | General Assembly
kevincroesch@gmail.com

Goal

Using Reddit's API, we'll collect posts from two subreddits of our choosing.

We'll then use NLP to train a classifier on which subreddit a given post came from..

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

Define the Problem

Can we train a classifier to accurately predict which subreddit a given post came from?

- Subreddit 1 = CrossFit
- Subreddit 2 = Good Mythical Morning

CrossFit

Define the Problem

Why did I pick this subreddit?

- I’m familiar with CrossFit
- Very strong and very vocal community

Good Mythical Morning

Why did I pick this subreddit?

Define the Problem

- YouTube show with 15+ Million Subscribers
- Vast and loyal community – referred to as “Mythical Beasts”

Obtain the Data

https://www.reddit.com/r/crossfit.json

https://www.reddit.com/r/goodmythicalmorning.json

GENERAL ASSEMBLY

Obtain the Data

951 Posts

979 Posts

Obtain the Data

All 950 were unique posts

All 979 were unique Posts

Explore the Data

- Located the text from each post
- Created a data frame with all posts
- Created a new column to identify each post as a CrossFit post
- Cleaned up each post by removing line breaks and extra spaces
- Removed any post that was “ ‘ ’ “

Explore the Data

	Post	from_cf
1	Hi all, I started building me a home gym and I...	1
2	I live pretty close to Central but traveling a...	1
4	Hey all, I have just been curious lately about...	1
5	Can someone teach how to record biometrics dur...	1
6	So, maybe I should be turning to my therapist ...	1
7	I have gone to 2 CrossFit classes and each one...	1
8	While getting back from a knee injury by resti...	1
10	Hello! I was wondering if anyone could recomm...	1
12	As the header says I'm wondering how bad this ...	1
14	The Start/Stop has water damage to it and I ju...	1

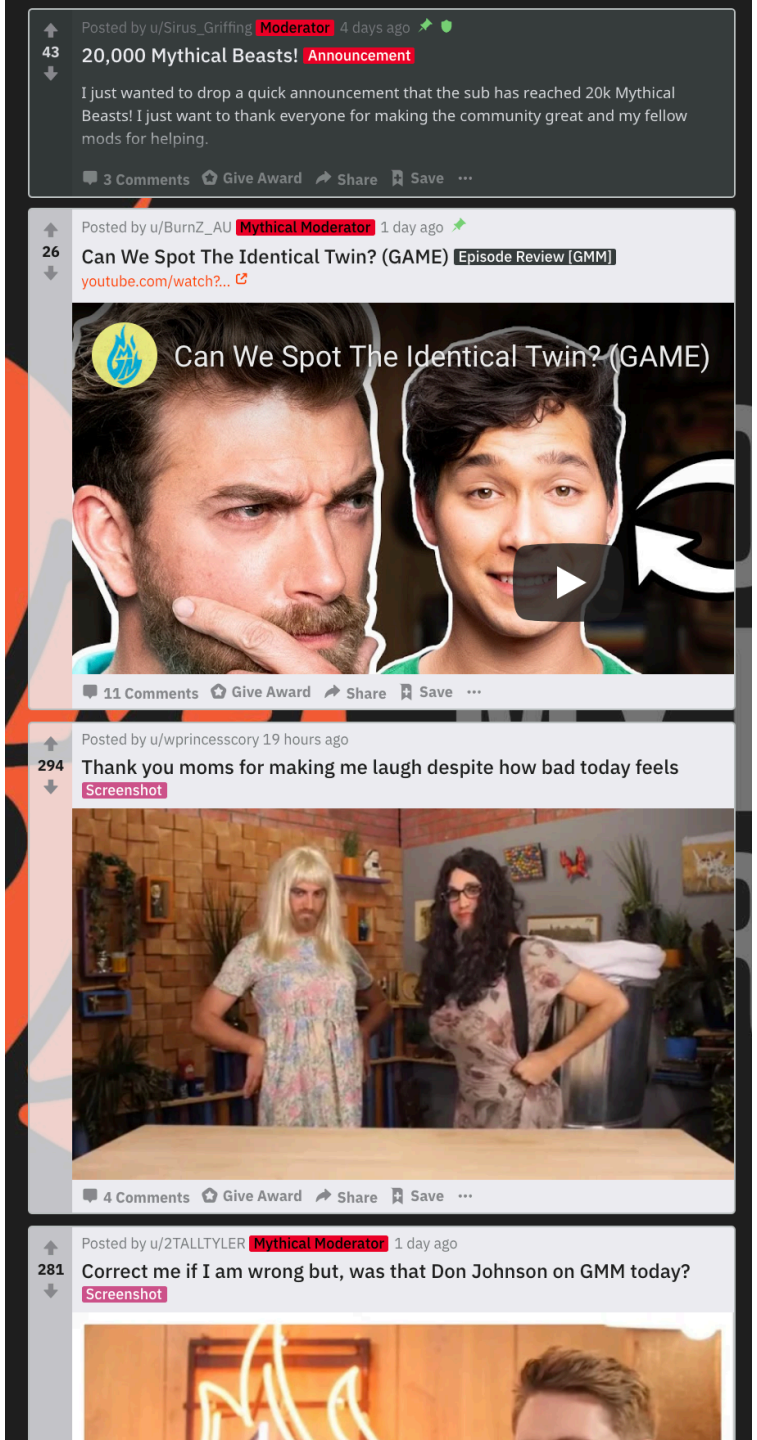
859 posts remaining

Explore the Data

- Located the text from each post
- Created a data frame with all posts
- Created a new column to identify each post as NOT a CrossFit post
- Cleaned up each post by removing line breaks and extra spaces
- Removed any post that was “ ‘ ’ “

GENERAL ASSEMBLY

Explore the Data



GENERAL ASSEMBLY

Explore the Data

Only 231 Posts remaining

GENERAL ASSEMBLY

Explore the Data

Running

https://www.reddit.com/r/running/.json

Why did I end up choosing the running subreddit?

- I enjoy running
- Also has a vast community and I thought it may be a more text based thread

GENERAL ASSEMBLY

Explore the Data

	Post	from_cf
0	***NOTE: This post was graciously stolen (w/ ...	0
1	With over 600,000 users, there are a lot of p...	0
3	Hello everyone, this is my first post in this ...	0
4	Good morning, Runnit! Another weekend of races...	0
5	How are you feeling today? Anything nagging yo...	0
6	I am making a switch over to more meatless day...	0
7	F26, although I was fairly athletic in high sc...	0
8	I have a really bad case of scoliosis. On top...	0
9	My husband wants to borrow my 235 to check his...	0
10	What would be a relatively cheap running/fitne...	0

869 posts remaining

GENERAL ASSEMBLY

Obtain the Data

CrossFit®

1728 Total posts for my dataset

Model the data

Train

Test

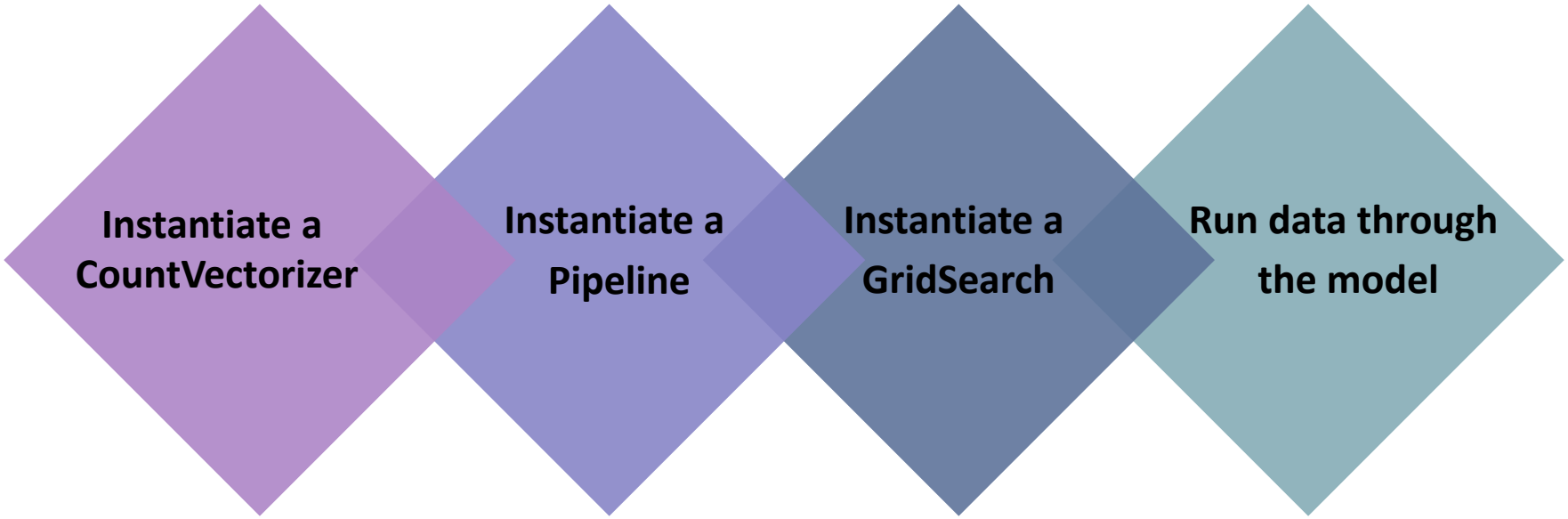
Split

Model the data

Create a loop to prepare the posts for our model

- Remove HTML
- Remove Non-Le*tters*
- Convert to lower case, split into individual words
- Convert the stop words to a set
- Remove stop words
- Join the words back into one string separated by space

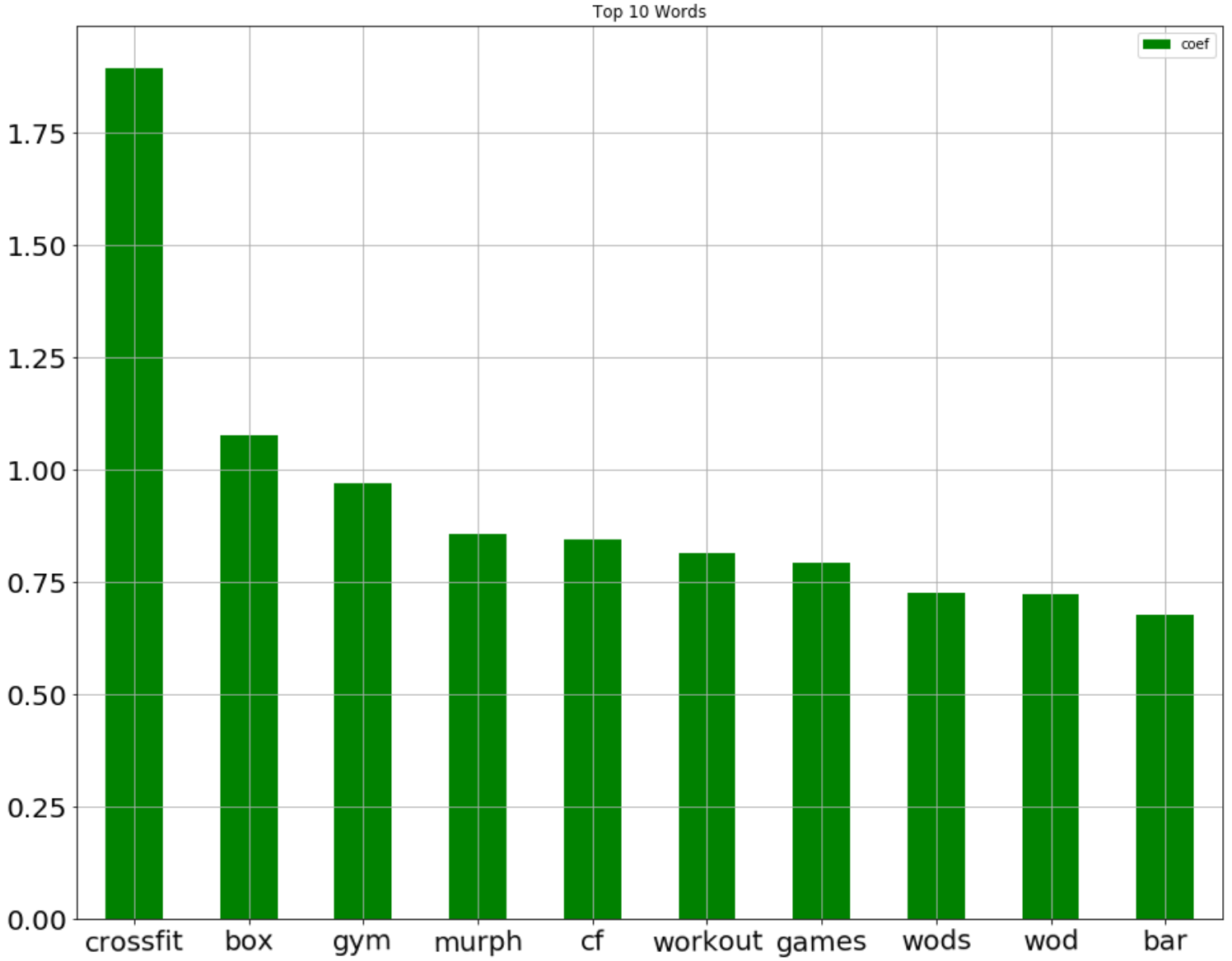
Model the data



Evaluate the Model

- Train Accuracy Score: 99.83
- Test Accuracy Score: 95.26
- Best Max Features: 10,000
- Best Model – Logistic Regression

Evaluate the Model



END

Kevin Roesch
Data Scientist | General Assembly
kevincroesch@gmail.com