Which gym is right for you?

Boris Doosey

Problem Statement

Crossfit and Orange Theory are two of the biggest fitness regimen studios in the country. Both boast huge followings in memberships and in online presence as well. While they both aim to get you in shape through exercise, their methods differ on how to achieve your fitness goals. Crossfit is known to have a bigger strength training component to its workouts, while Orange Theory emphasizes cardio more.

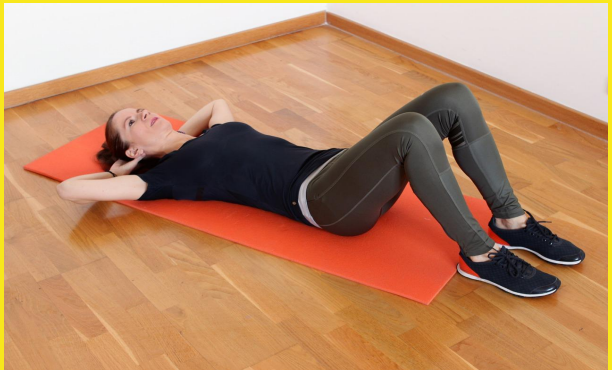
As a data scientist, the purpose of this project is to use data from the respective subreddits of each gym to build a model that will be able to differentiate between the two when given a sample of text. From there, I am hoping to be able to make a prediction on which group is happier/more satisfied with their workouts.

Crossfit

- Created Dec. 25, 2008
- 263k members
- Headquartered here in DC

Orange Theory

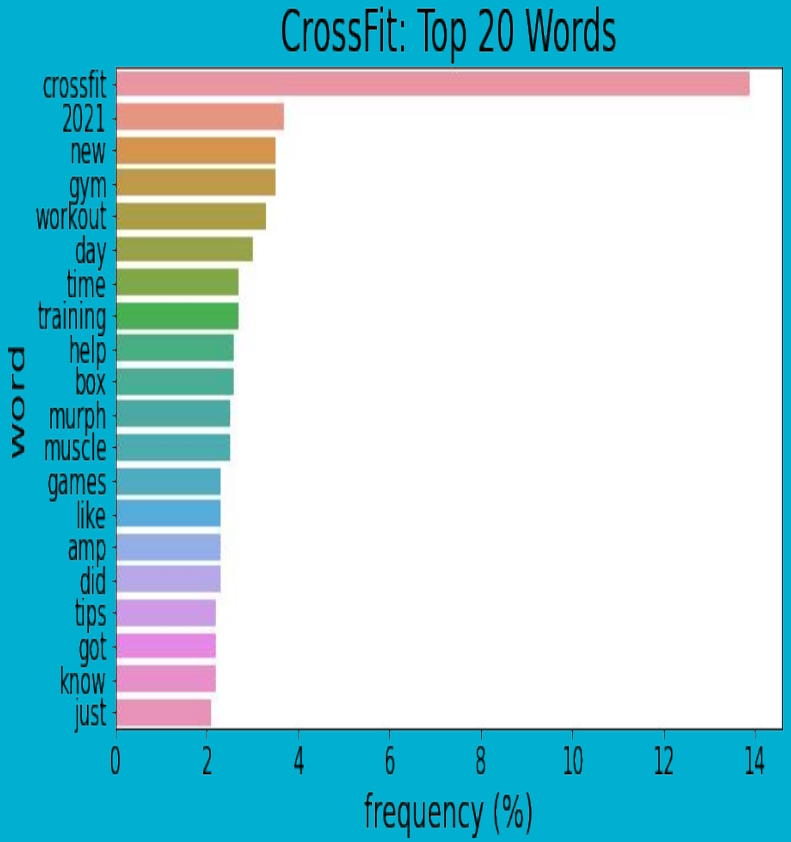
- Created Sep. 10 2015
- 137k members
- Workouts based on HIIT, and involve wearing heart rate monitors to stay within one of 5 heart rate zones.

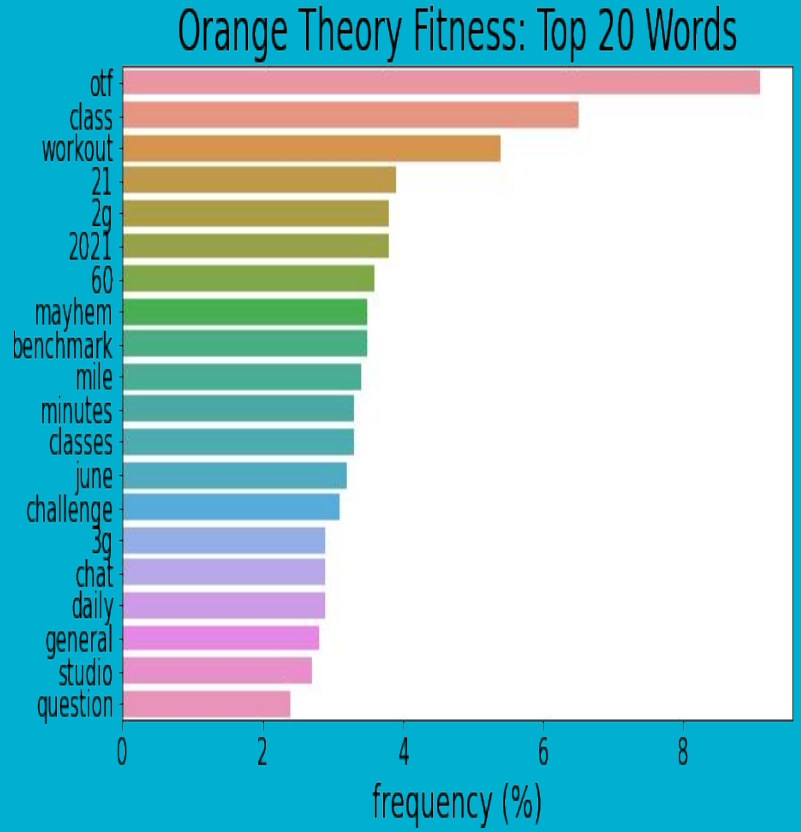


Data Gathering and Processing

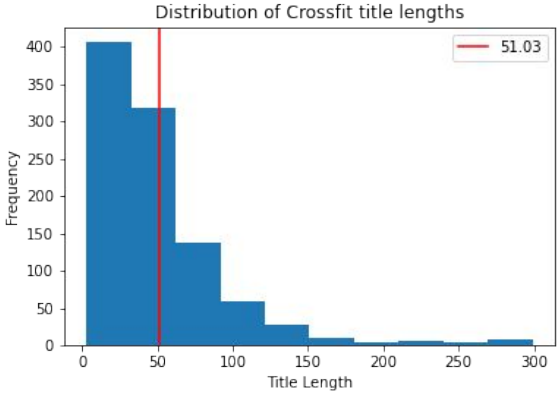
- 1000 posts from each subreddit using pushshift api calls
- Custom function to clean posts
- Tokenize posts and filter stopwords

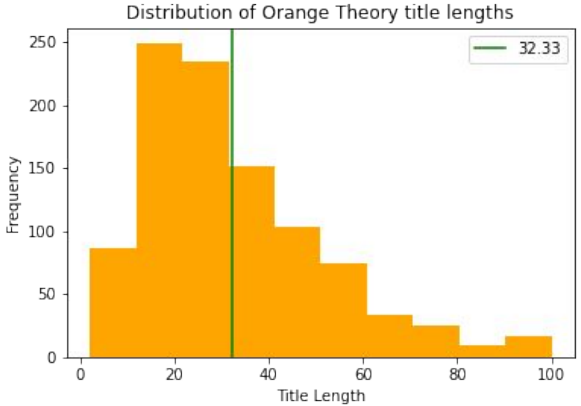
Top words

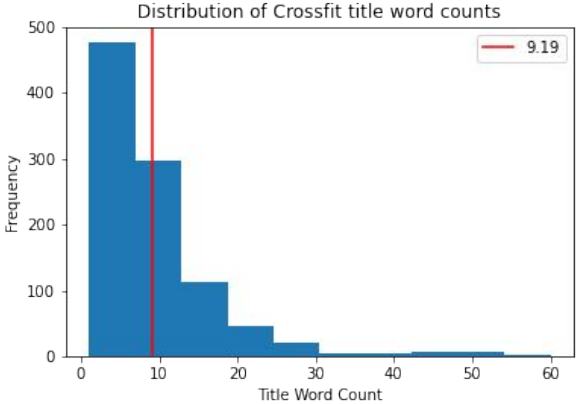

CrossFit: Top 20 Words
word
crossfit
2021
new
gym
workout
day
time
training
help
box
murph
muscle
games
like
amp
did
tips
got
know
just
0
2
4
6
8
10
12
14
frequency (%)

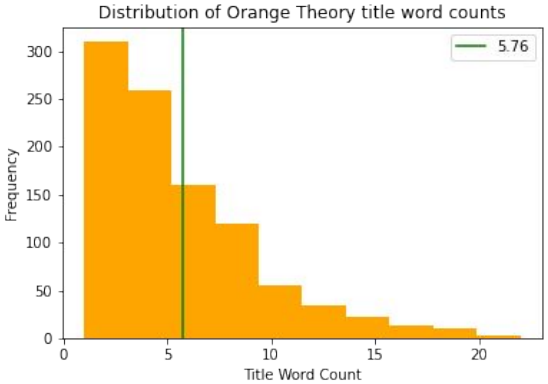

Orange Theory Fitness: Top 20 Words
otf
class
workout
21
2g
2021
60
mayhem
benchmark
mile
minutes
classes
june
challenge
3g
chat
daily
general
studio
question
0
2
4
6
8
frequency (%)

Distributions

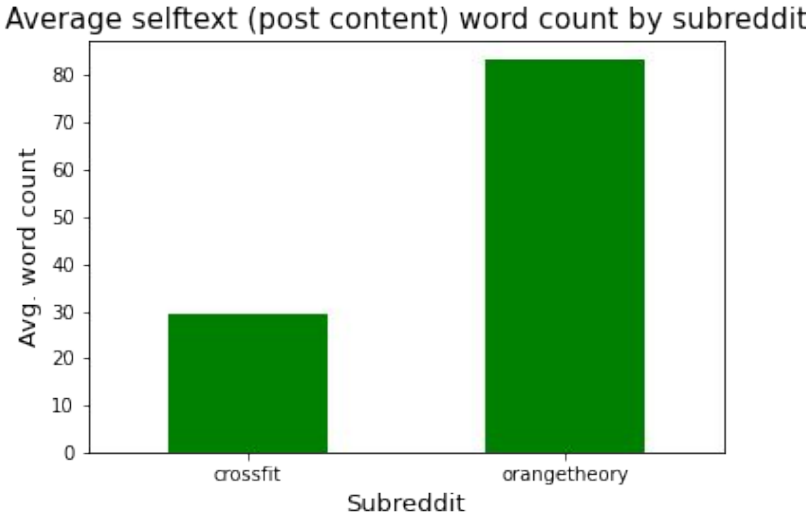


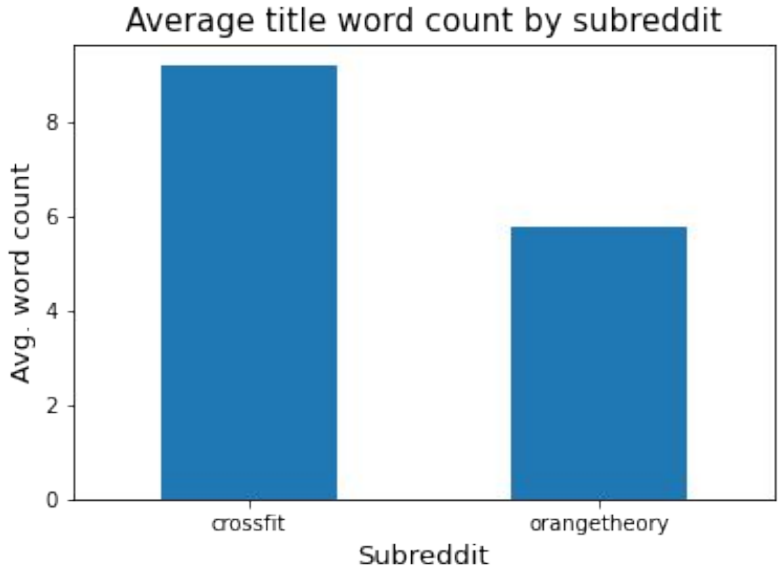




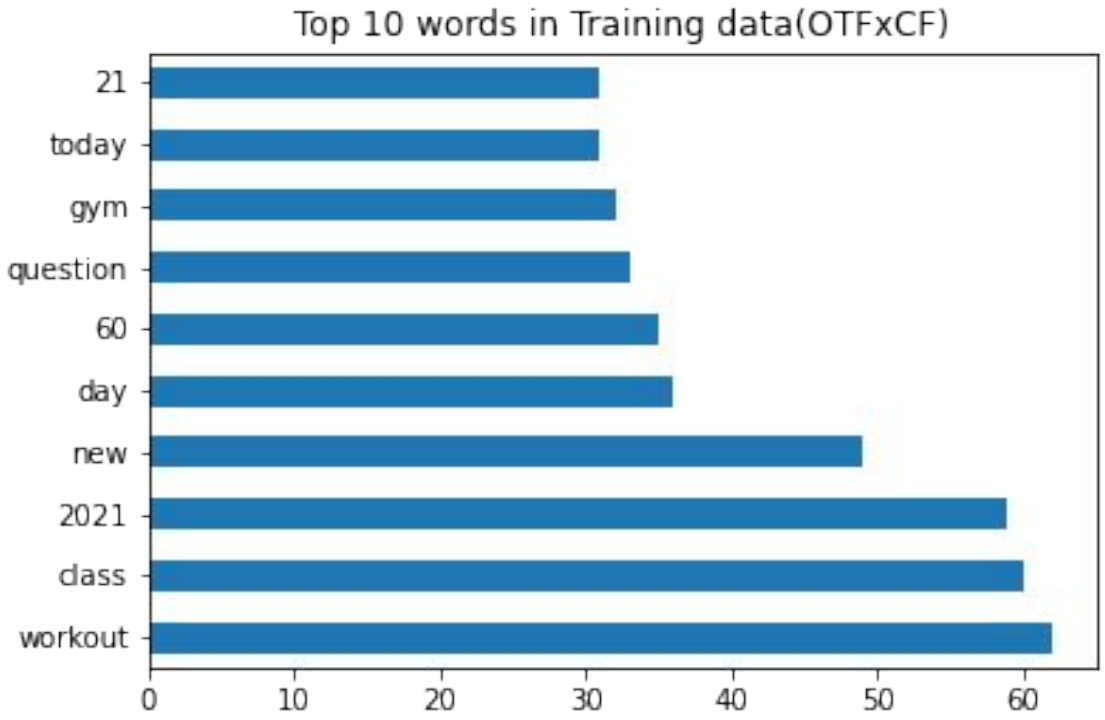


Distributions cont..



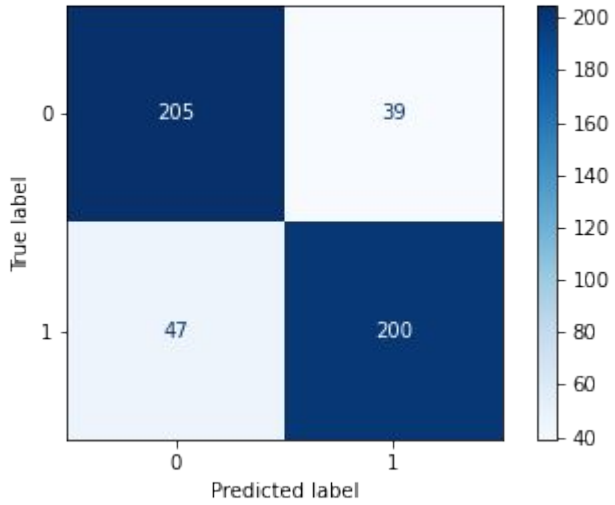


Top Words of combined subreddits (after processing)



Modeling

Model	Multinomial Naive Bayes
Vectorizer	TFIDF
Training Score	95.72%
Testing Score	81.87%
Mean Cross-Val Score:	82.89%



Specificity : 82.78%

81.87%

Achieved on testing data of best model, compared to 50.26% of baseline accuracy, in which Orange Theory was the greater class

Sentiment Analysis

	neg	neu	pos	compound
count	975.000000	975.000000	975.000000	975.000000
mean	0.056265	0.837358	0.106376	0.093832
std	0.136578	0.211725	0.178028	0.320572
min	0.000000	0.000000	0.000000	-0.836000
25%	0.000000	0.686000	0.000000	0.000000
50%	0.000000	1.000000	0.000000	0.000000
75%	0.000000	1.000000	0.204000	0.310800
max	1.000000	1.000000	1.000000	0.948900

Crossfit: 'title' SIA - avg. compound = 9.3%

	neg	neu	pos	compound
count	987.000000	987.000000	987.000000	987.000000
mean	0.054322	0.864228	0.081451	0.025959
std	0.154718	0.229225	0.182385	0.239759
min	0.000000	0.000000	0.000000	-0.812600
25%	0.000000	0.738500	0.000000	0.000000
50%	0.000000	1.000000	0.000000	0.000000
75%	0.000000	1.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	0.884200

OTF : 'title' SIA - avg. compound = 2.5%

	neg	neu	pos	compound
count	341.000000	341.000000	341.000000	341.000000
mean	0.047935	0.833774	0.118287	0.385459
std	0.066213	0.102247	0.092725	0.571745
min	0.000000	0.514000	0.000000	-0.965400
25%	0.000000	0.778000	0.056000	0.000000
50%	0.028000	0.829000	0.106000	0.593100
75%	0.070000	0.904000	0.182000	0.877700
max	0.482000	1.000000	0.475000	0.993700

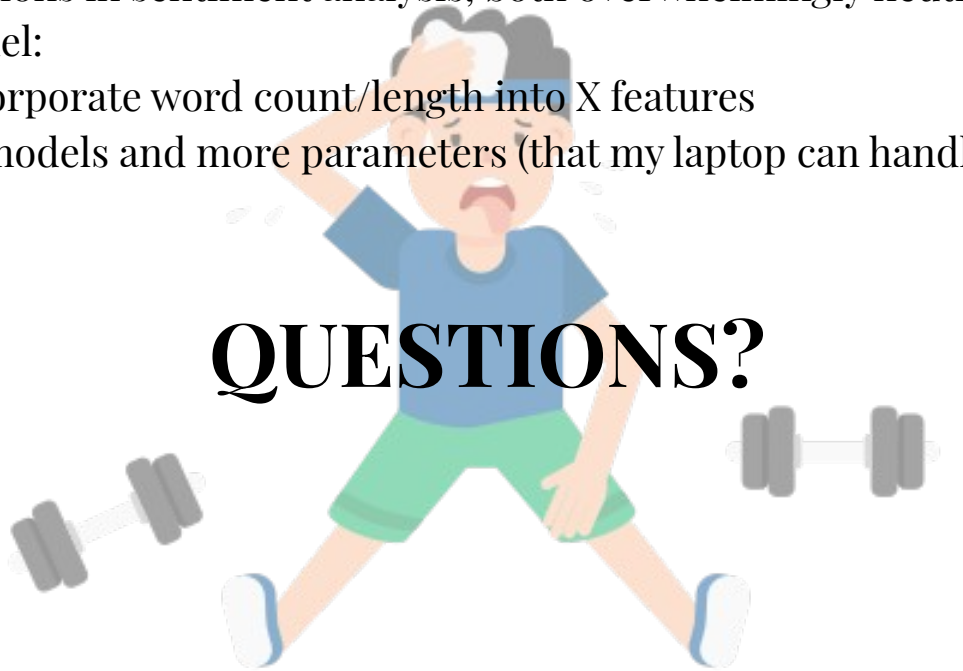
Crossfit: 'selftext' SIA -
avg. compound = 38.5%

	neg	neu	pos	compound
count	844.000000	844.000000	844.000000	844.000000
mean	0.054960	0.831831	0.113216	0.285472
std	0.060359	0.098168	0.090958	0.628662
min	0.000000	0.000000	0.000000	-0.997200
25%	0.000000	0.769750	0.050750	-0.190925
50%	0.042000	0.839000	0.103000	0.501650
75%	0.088000	0.890000	0.161000	0.827350
max	0.438000	1.000000	1.000000	0.997200

OTF : 'selftext' SIA -
avg. compound = 28.5%

Conclusions & Recomendations

- No clear distinctions in sentiment analysis, both overwhelmingly neutral
- To improve model:
  - Maybe incorporate word count/length into X features
  - Try more models and more parameters (that my laptop can handle)



QUESTIONS?