GA PROJECT 3

NLP CLASSIFICATION FOR AD TARGETING

R/CODINGBOOTCAMP VS R/CSMAJOR

MARY-ANNE, RIFQI, SHAWN, TING WEI, WEI ZHE

CONTENTS

BACKGROUND

There is an increased competition in the space for coding bootcamps.

BACKGROUND

There is an increased competition in the space for coding bootcamps.

HACK REACTOR le wagon Vertical Institute ROCKET ACADEMY

If no action is taken, General Assembly may face...

DECLINE IN MARKET SHARE

LOWER MARKETING ROI

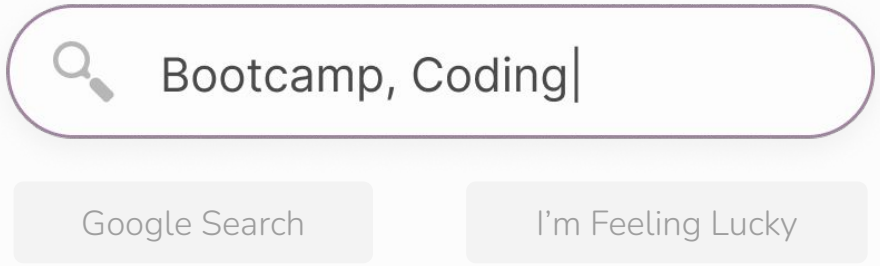
POORER LEAD GENERATION

BACKGROUND

GENERAL ASSEMBLY

MARKETING TEAM

- **Better identify the online presence** of a **bootcamp seeker** as opposed to that of a computer science major to aid in targeted advertising.
- Considering the two topics have quite a bit in common, efforts to further segregate the two could yield **better advertising ROI**.



Current classifying model using straightforward **keywords** such as 'bootcamp' and 'coding' yields around **79% accuracy**.

https://www.keywordsrock.com

Keywords allow for targeted strategies at all levels of the marketing funnel

Keywords guide marketing teams on the sort of advertising content that is required.

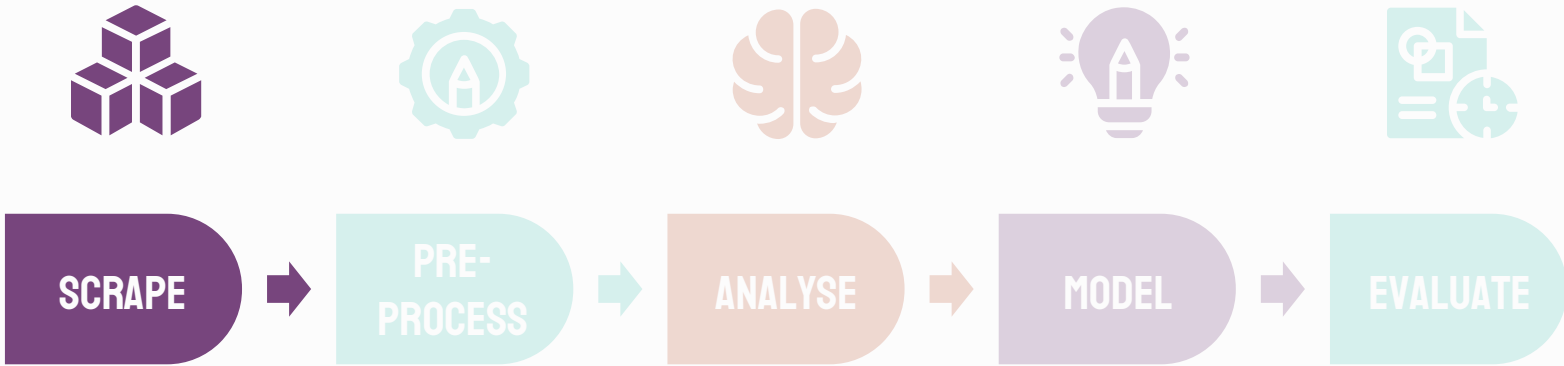
E.g. Google ads, one of the most effective platforms for generating leads and sales works well due to its ability to target users with high buying intent based on the keywords they use.

SEO Keywords · Google Ad · General Assembly · Coding Bootcamps

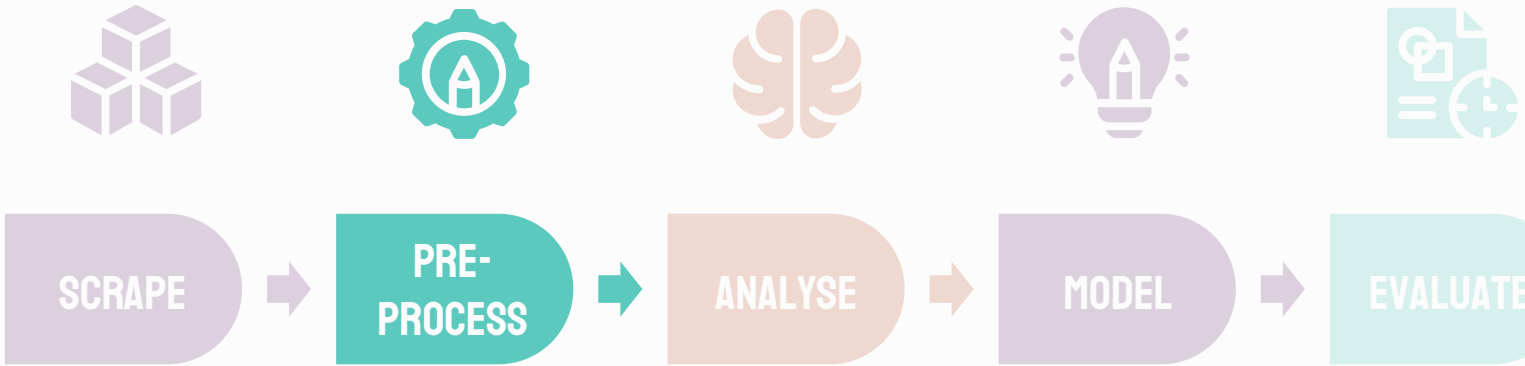
PROBLEM STATEMENT

Build a model with **>90% accuracy** that helps to identify between **those who are looking for bootcamp style learning** vs computer science majors/prospective students **based on the words they use** online.

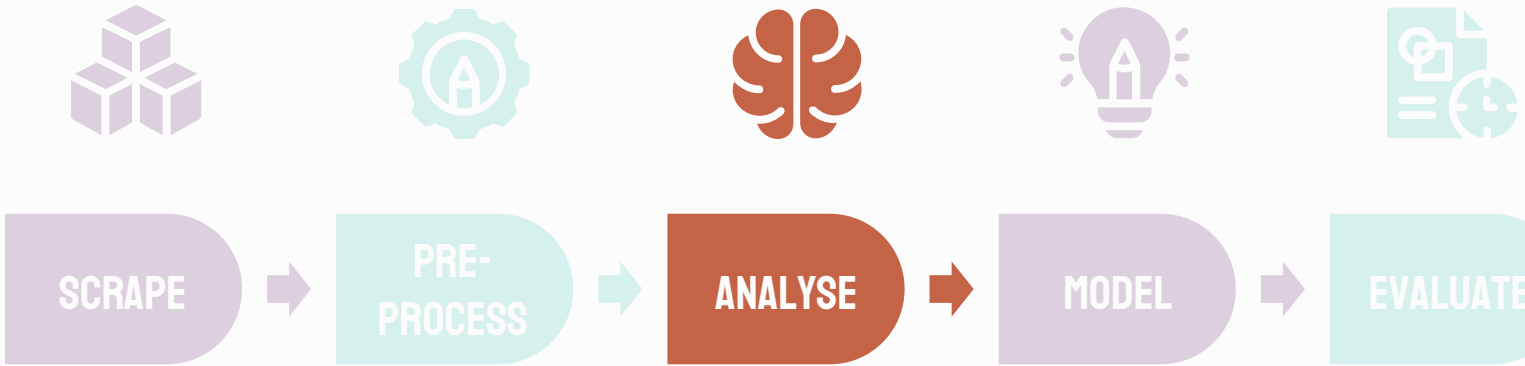
WORKFLOW



WORKFLOW

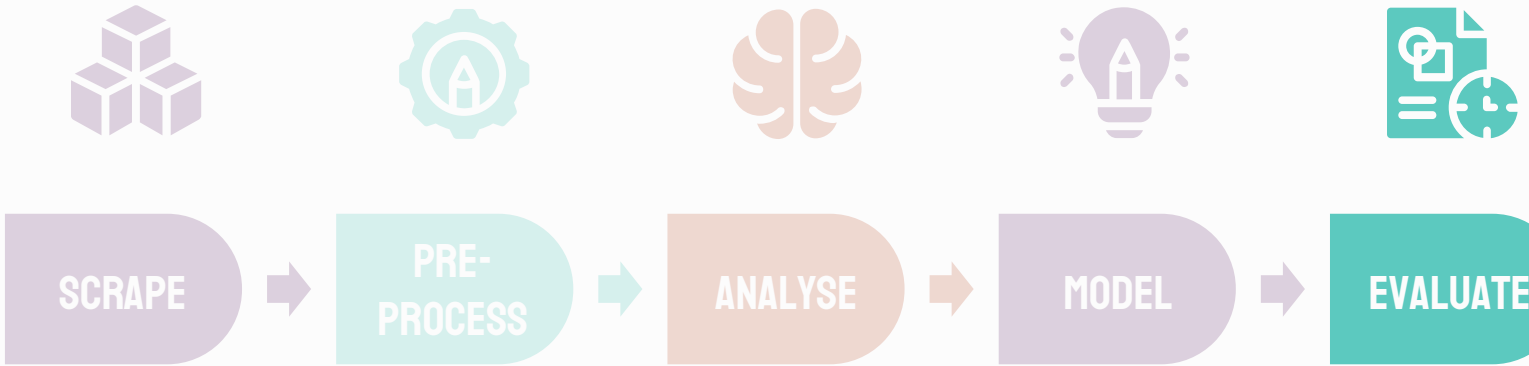


WORKFLOW



WORKFLOW

WORKFLOW



WHAT METHODS ARE WE USING TO CLEAN?

1. Web Scraping
2. Remove Null/Duplicate values
3. Remove punctuations

WHAT METHODS ARE WE USING TO CLEAN?

1. Tokenization
2. Remove stopwords
3. Stem / Lemmatize

Reddit API vs Pushshift API

- Easier retrieving data
- 5 times greater object limit

WHERE DO WE SCRAPE FROM?

Coding Bootcamp

r/codingbootcamp

Join

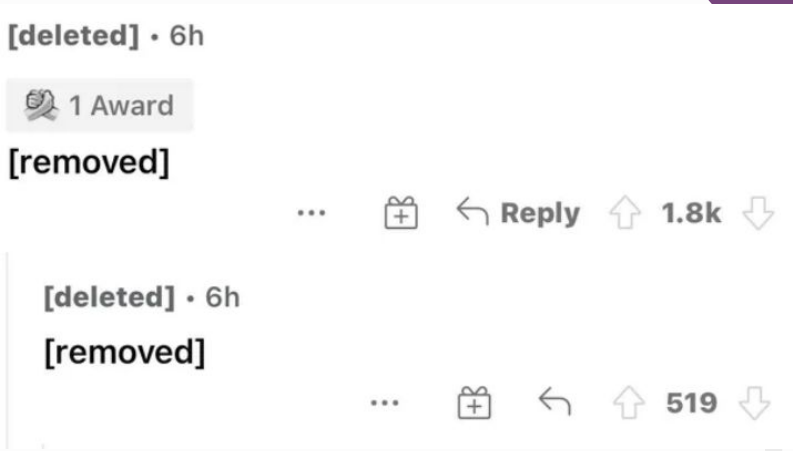
Students of Computer Science!

r/csMajors

Join

REMOVED & DELETED

-
-
- Removed & Deleted posts are not beneficial to our case



- They are replaced with an empty string

```
# Remove the words [removed] and [deleted] from selftexts
df['selftext'] = df['selftext'].replace('[removed]', '')
df['selftext'] = df['selftext'].replace('[deleted]', '')
```

REMOVE PUNCTUATION & TOKENIZATION

Removed Punctuation

Tokenization

body_text	body_text_clean	body_text_tokenized
I've been searching for the right words to tha...	Ive been searching for the right words to than...	[ive, been, searching, for, the, right, words,...
Free entry in 2 a wkly comp to win FA Cup fina...	Free entry in 2 a wkly comp to win FA Cup fina...	[free, entry, in, 2, a, wkly, comp, to, win, f...
Nah I don't think he goes to usf, he lives aro...	Nah I dont think he goes to usf he lives aroun...	[nah, i, dont, think, he, goes, to, usf, he, l...
Even my brother is not like to speak with me. ...	Even my brother is not like to speak with me T...	[even, my, brother, is, not, like, to, speak, ...
I HAVE A DATE ON SUNDAY WITH WILL!!	I HAVE A DATE ON SUNDAY WITH WILL	[i, have, a, date, on, sunday, with, will]

LEMMATIZATION

01. What?

02. How?

03. Why?

body_text_stemmed	body_text_lemmatized
[ive, search, right, word, thank, breather, pr...	[ive, searching, right, word, thank, breather,...
[free, entri, 2, wkli, comp, win, fa, cup, fin...	[free, entry, 2, wkly, comp, win, fa, cup, fin...
[nah, dont, think, goe, usf, live, around, tho...	[nah, dont, think, go, usf, life, around, though]
[even, brother, like, speak, treat, like, aid,...	[even, brother, like, speak, treat, like, aid,...
[date, sunday]	[date, sunday]

EDA

EDA : REMOVING ADDITIONAL STOPWORDS

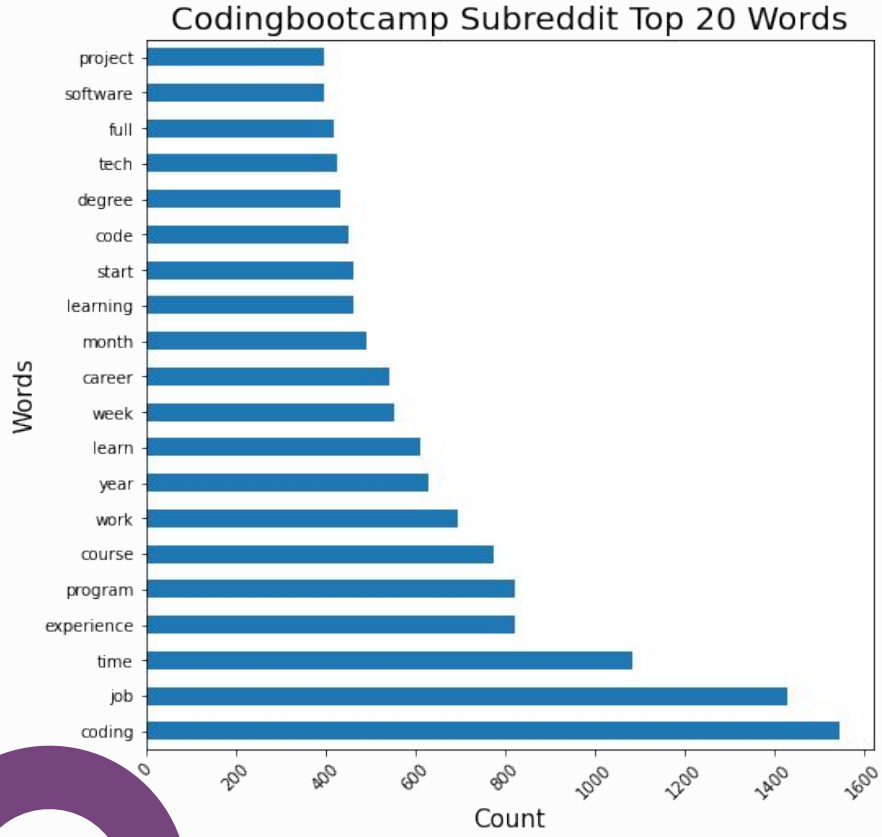
Coding Bootcamp top 50 words

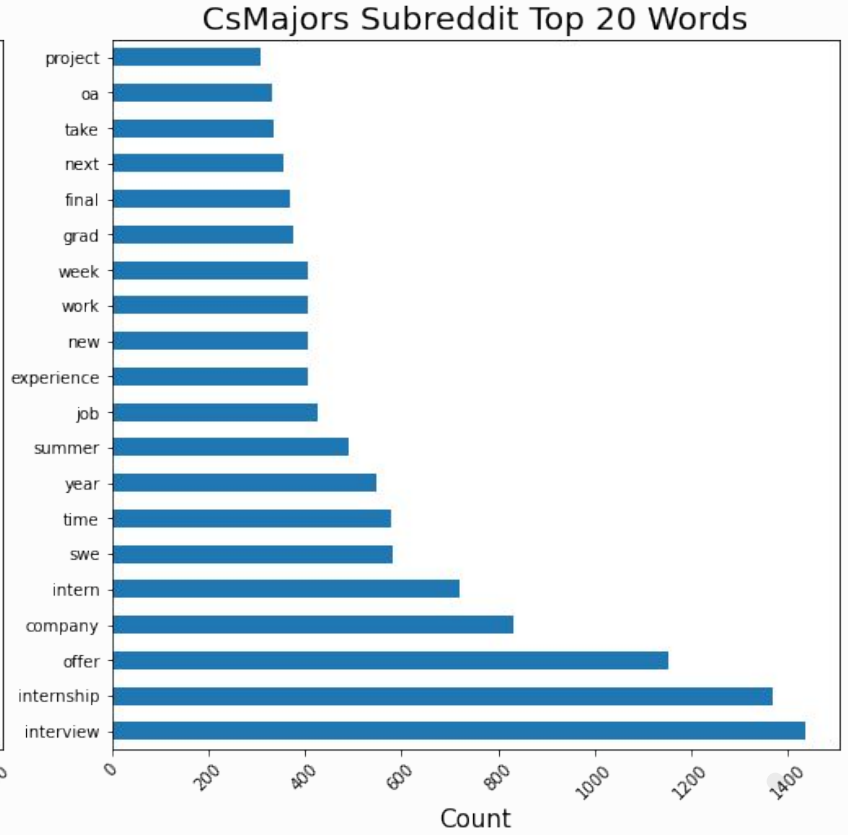
✖	bootcamp	2207
	coding	1544
	job	1427
✖	would	1238
✖	get	1149
✖	im	1117
✖	like	1095
	time	1083
✖	know	900
	experience	824
	program	823
✖	camp	788
	course	774
✖	want	756
✖	anyone	738
✖	one	720
✖	boot	710
	work	697
✖	looking	658
✖	bootcamps	633
	year	629
✖	good	618
	learn	610
✖	also	574
✖	people	571
✖	really	566
	week	554
✖	help	553
	career	542
	month	491
✖	ive	483
✖	need	482
	learning	463
	start	462
✖	make	458
	code	453
✖	dont	450
✖	question	437
	degree	432
✖	go	431
	tech	428
	full	418
✖	going	398
	software	397
	project	396
✖	lot	395
	academy	381
✖	much	376
✖	take	374
✖	feel	368

CS Majors top 50 words

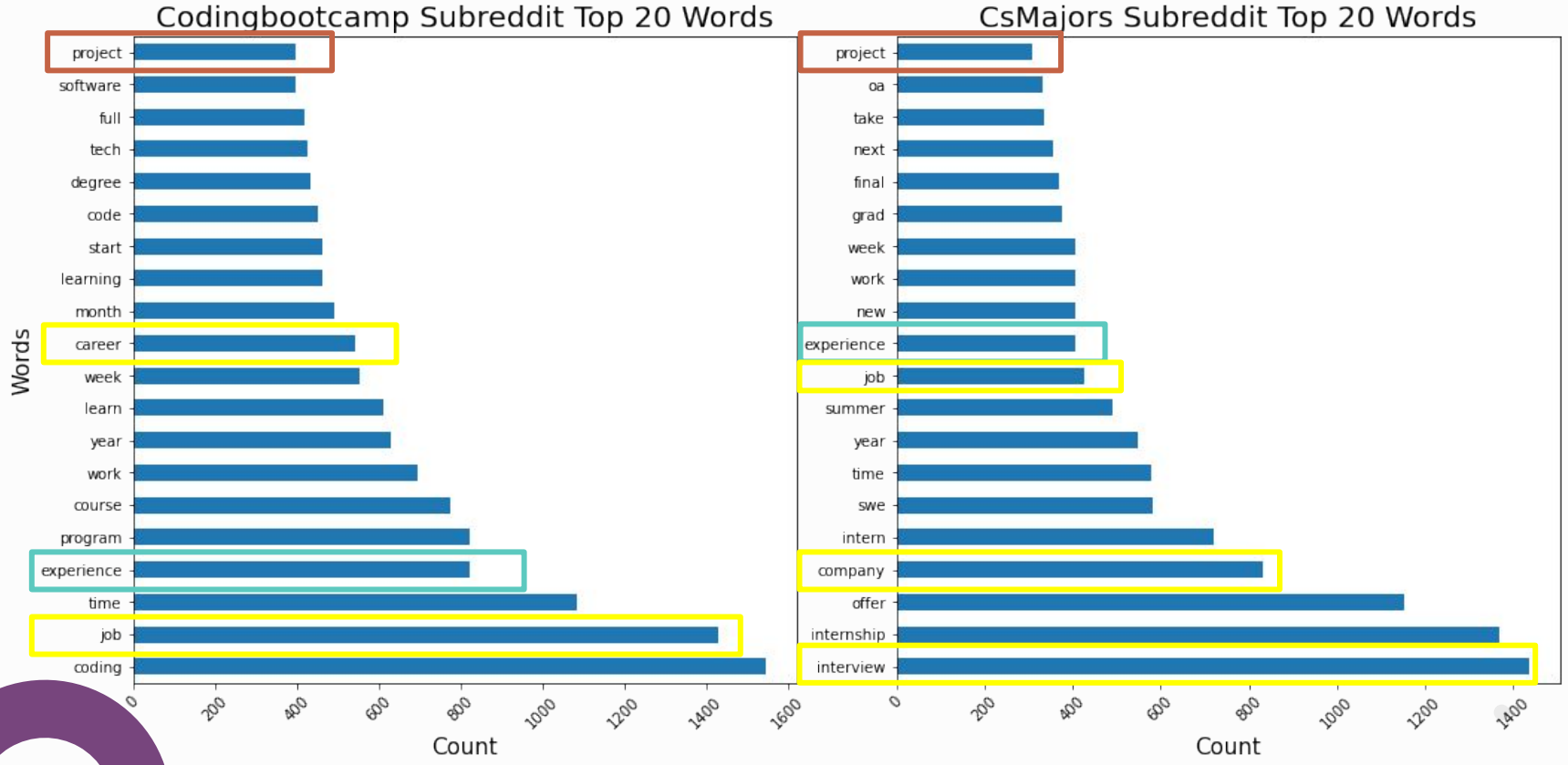
	interview	1435
	internship	1368
	offer	1152
	company	832
✖	would	831
✖	anyone	812
✖	im	808
✖	get	758
✖	like	736
	intern	720
✖	know	650
✖	got	635
✖	question	614
✖	one	590
	swe	582
	time	578
	year	547
	summer	492
✖	also	434
	job	427
	round	425
	experience	407
	work	406
	new	406
	week	405
✖	really	404
	view	398
✖	back	390
✖	want	383
	grad	375
	final	369
✖	good	368
	next	356
	take	336
	oa	331
✖	people	321
✖	dont	314
	first	313
	recruiter	307
	project	307
✖	think	299
✖	still	292
	tech	292
	resume	291
	school	284
✖	even	283
✖	getting	278
✖	much	278
✖	feel	275
	class	275

EDA : TOP 20 WORDS

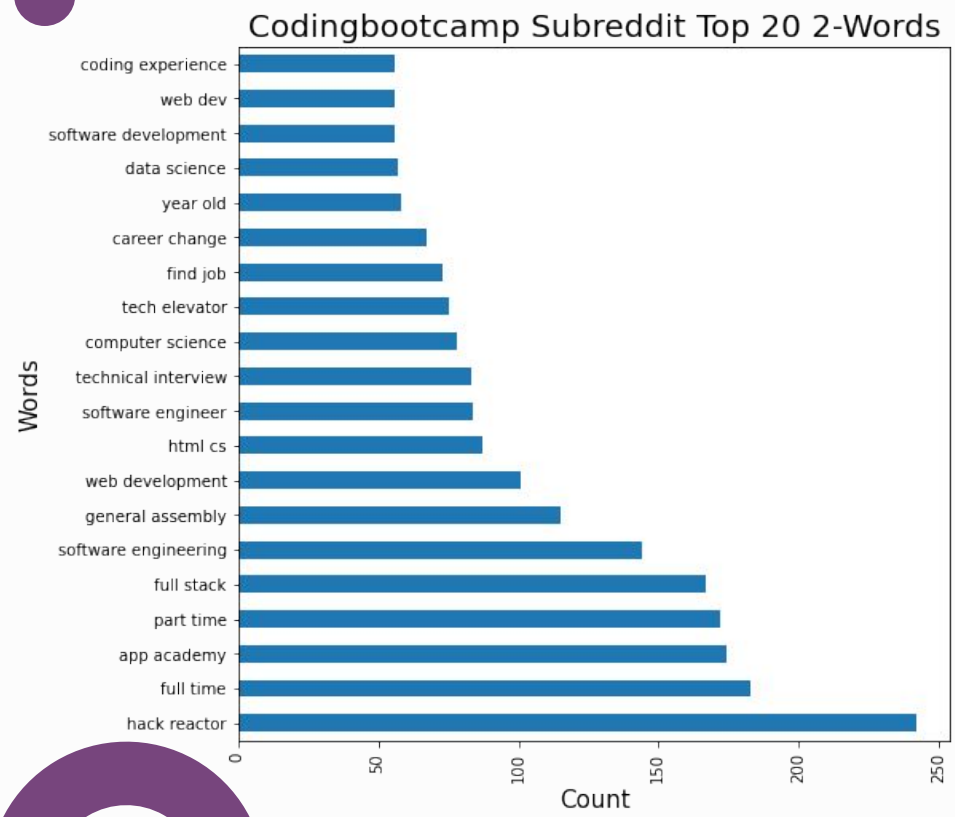


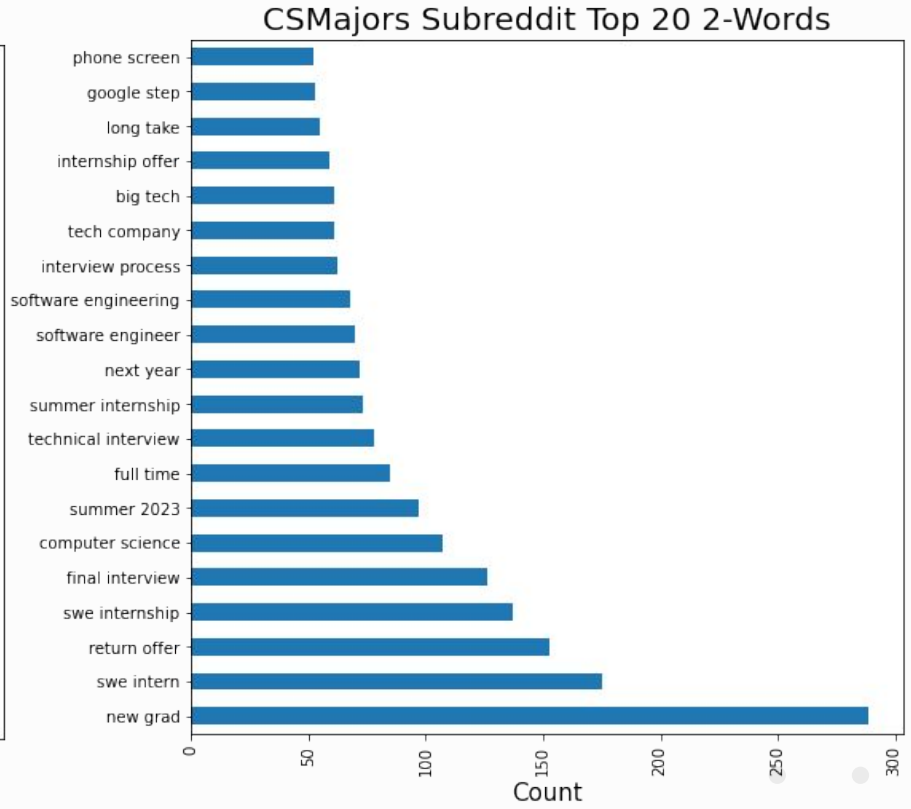


EDA : TOP 20 WORDS - SIMILARITIES

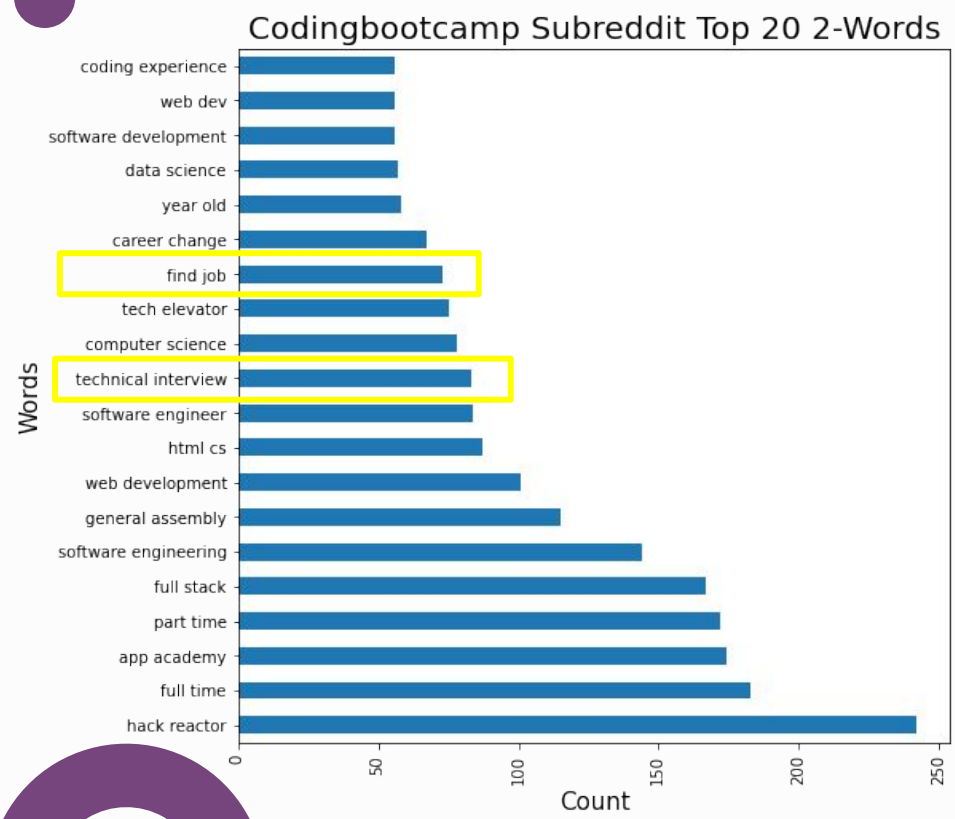


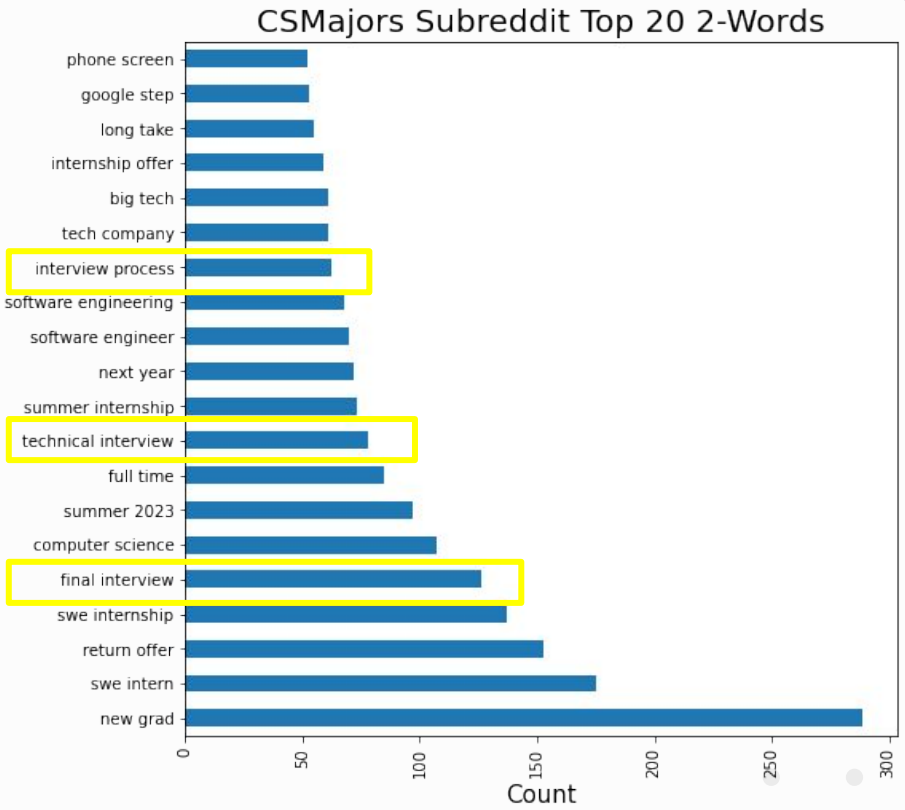
EDA : TOP 20 2-WORDS



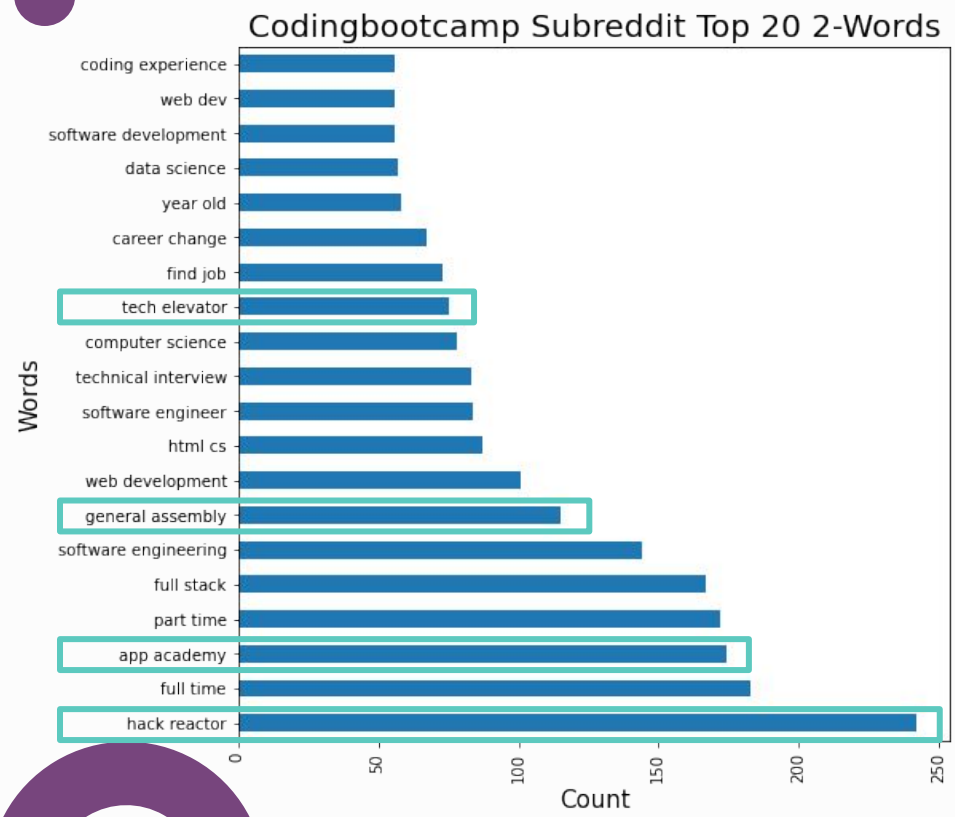


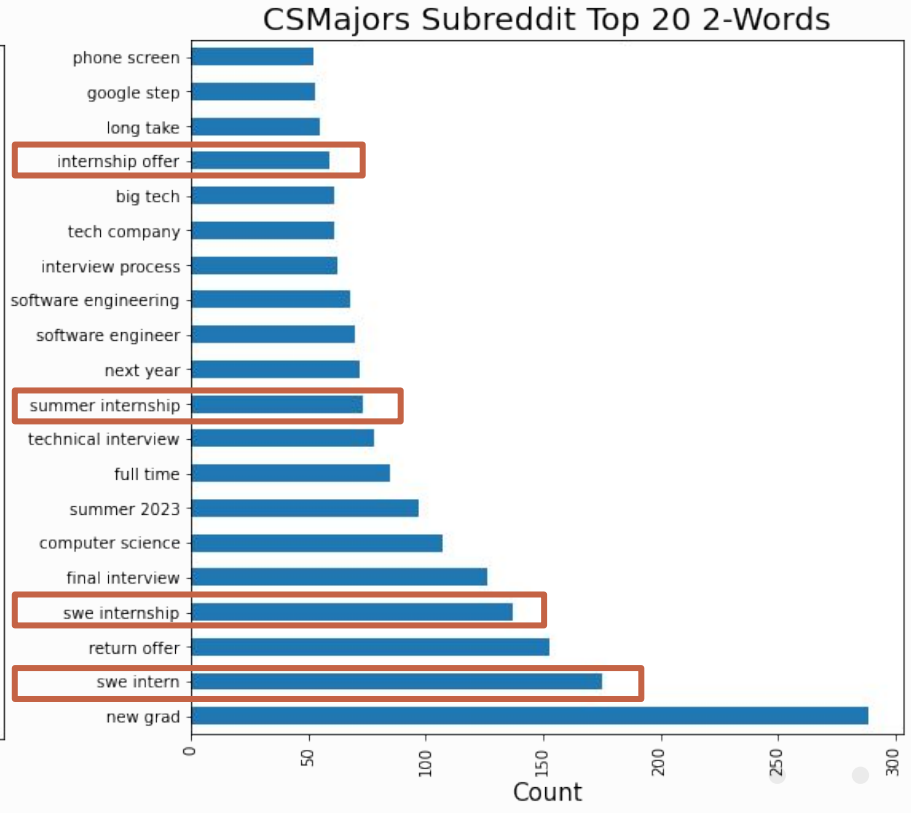
EDA : TOP 20 2-WORDS - SIMILARITIES: CAREER



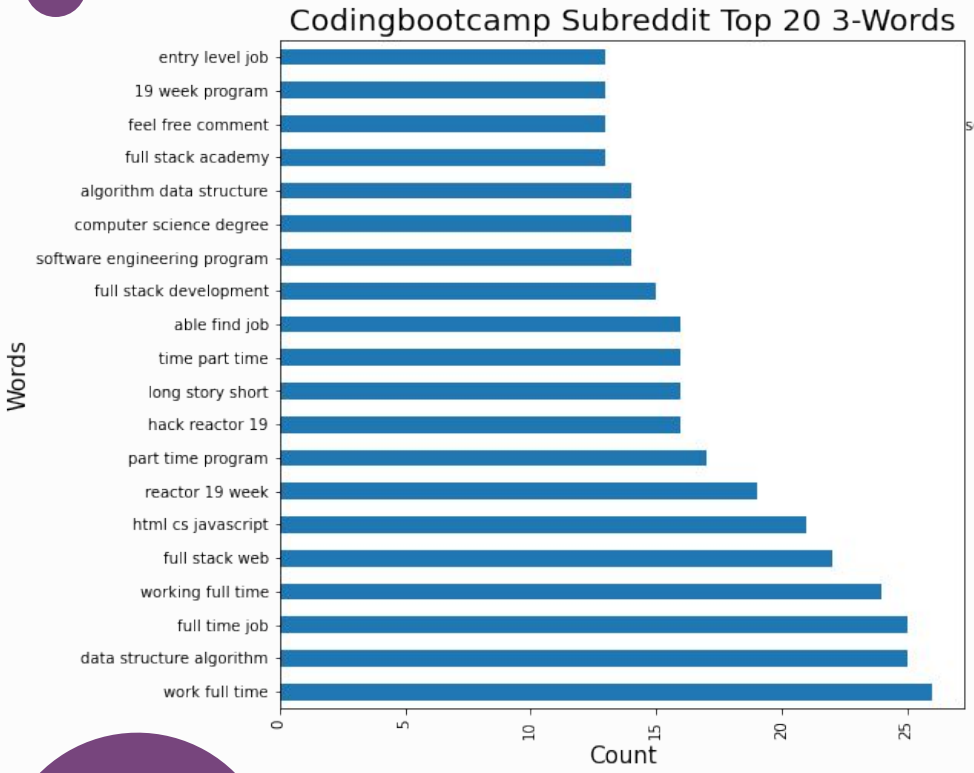


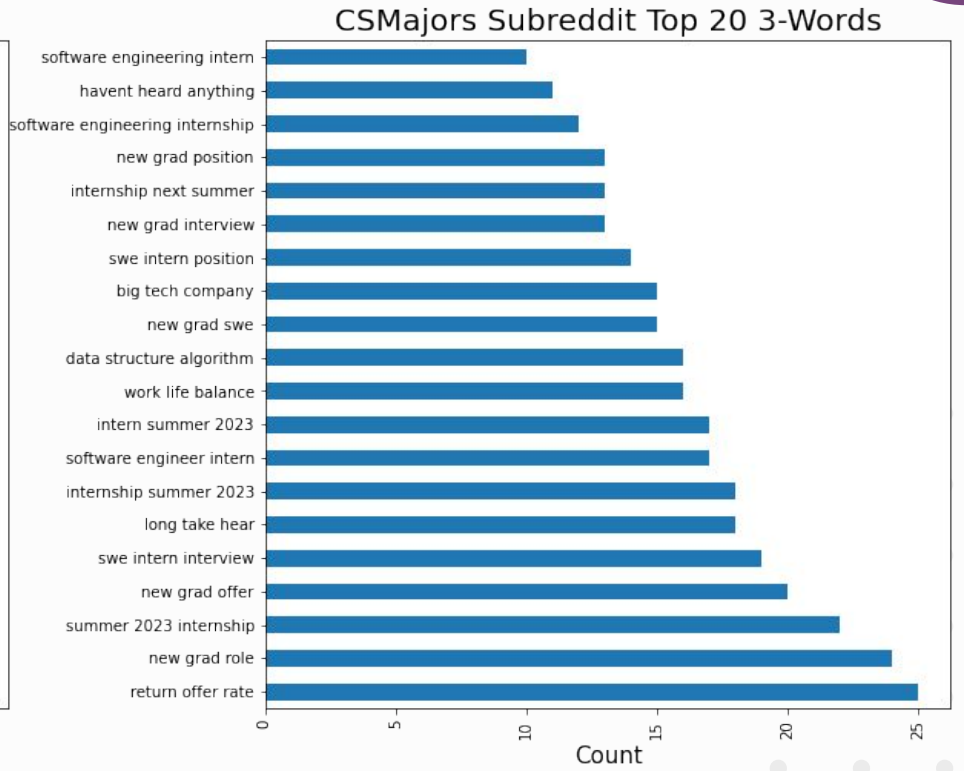
EDA : TOP 20 2-WORDS - DIFFERENCES: SCHOOLS VS INTERNSHIPS



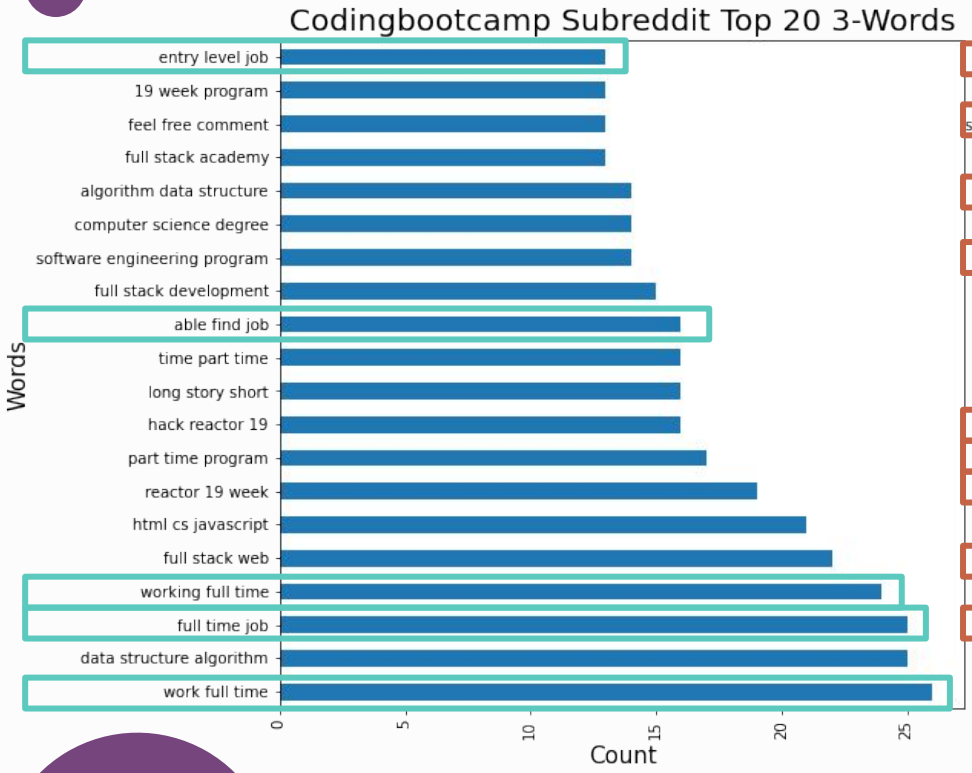


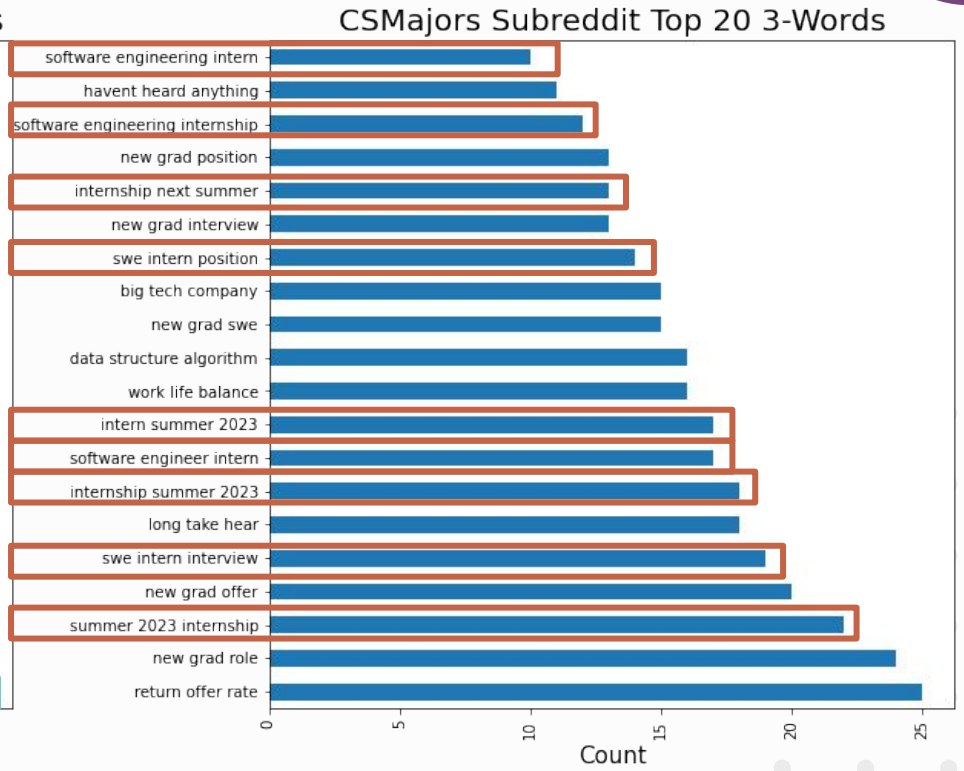
EDA : TOP 20 3-WORDS



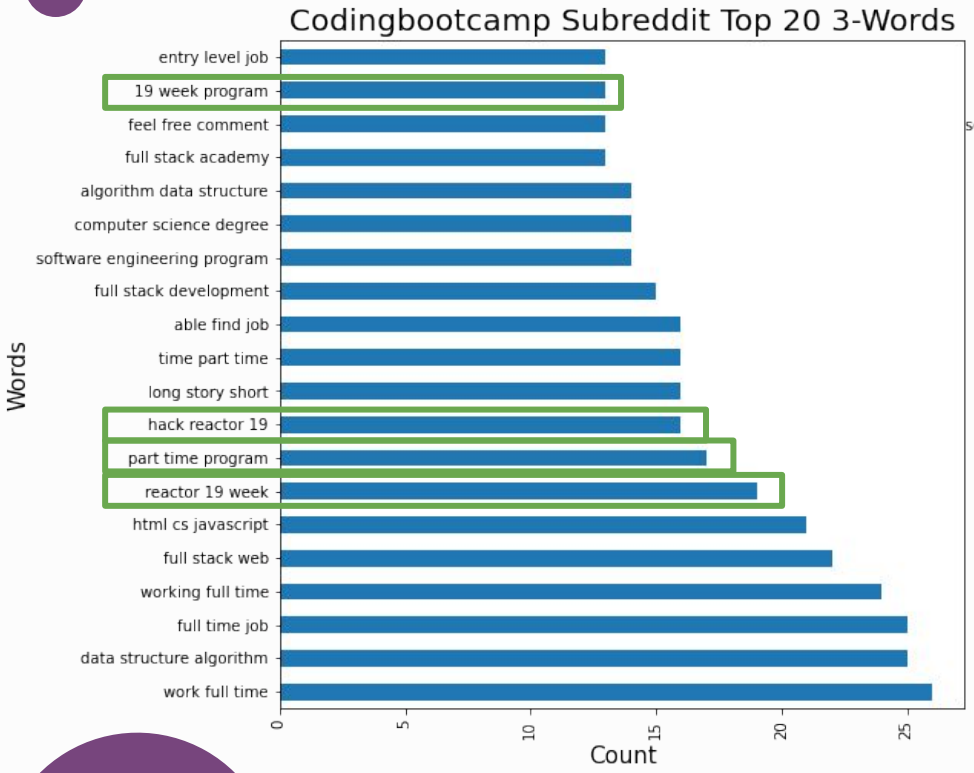


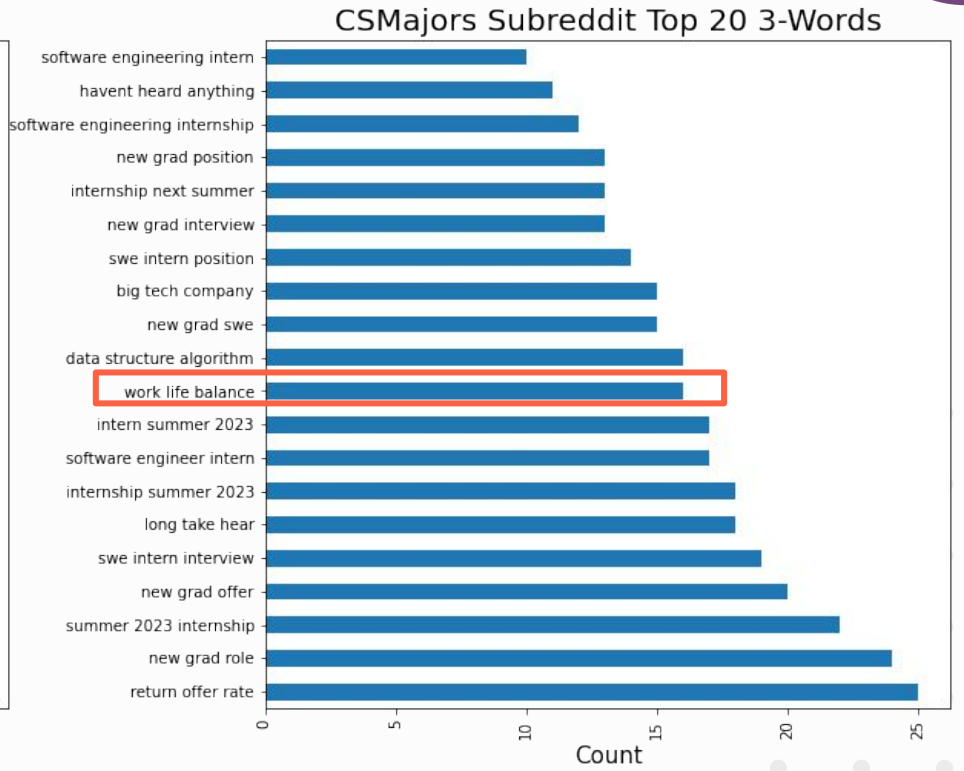
EDA : TOP 20 3-WORDS - DIFFERENCES: FULL TIME JOB VS INTERNSHIP





EDA : TOP 20 3-WORDS - DIFFERENCES: TIME VS BALANCE

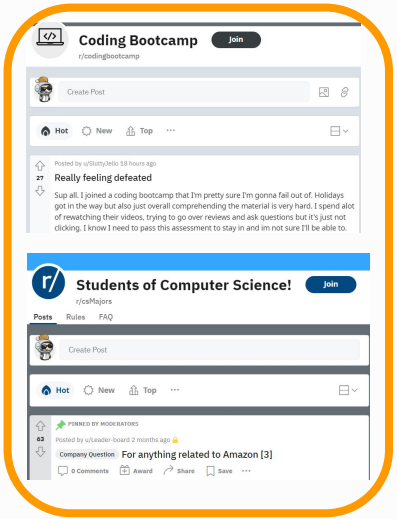




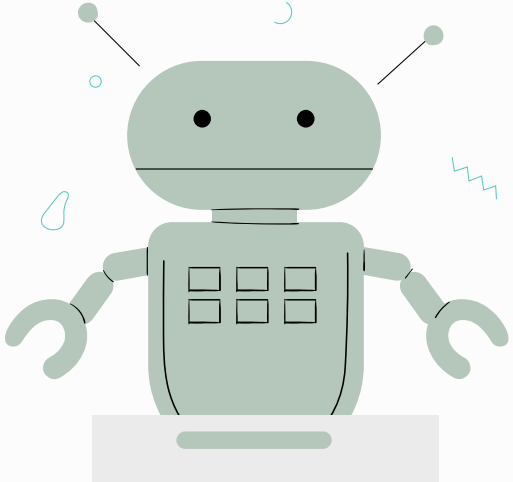
MODELLING

PURPOSE OF MODEL

Reddit Posts



Model



>90% accuracy

Bootcamp Style

4 years uni course

BUILDING A CLASSIFICATION MODEL

Converting text to numerical representation

Classification model selection

Improving model accuracy

Methods used:

- Countvectorizer
- N-grams
- TF-IDF (Term Frequency-Inverse Document Frequency)

Models used:

- Bernoulli Naive Bayes
- Multinomial Naive Bayes
- Logistic Regression

Optimization:

- Hyperparameter tuning

MODEL SELECTION

VECTORIZATION TYPE	CLASSIFICATION MODEL	TRAIN ACCURACY SCORE	TEST ACCURACY SCORE
Baseline		0.78039	0.78646
Countvectorizer	Bernoulli Naive Bayes	0.84611	0.84215
Countvectorizer	Multinomial Naive Bayes	0.93214	0.93149
Countvectorizer	Logistic Regression	0.98678	0.93429
N-Gram*	Bernoulli Naive Bayes	0.90435	0.86057
N-Gram*	Multinomial Naive Bayes	0.98464	0.90545
N-Gram*	Logistic Regression	0.94416	0.875
TF-IDF	Bernoulli Naive Bayes	0.95698	0.92548
TF-IDF	Multinomial Naive Bayes	0.95431	0.92748
TF-IDF	Logistic Regression	0.96193	0.94231

*Only the best train-test result between Bi-Gram & Tri-Gram for the model is shown.

Higher accuracy score Better Model Performance

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

- Vectorization method that penalizes terms that occur multiple times across different documents.

Text 1	i love natural language processing but i hate python
Text 2	i like image processing
Text 3	i like signal processing and image processing

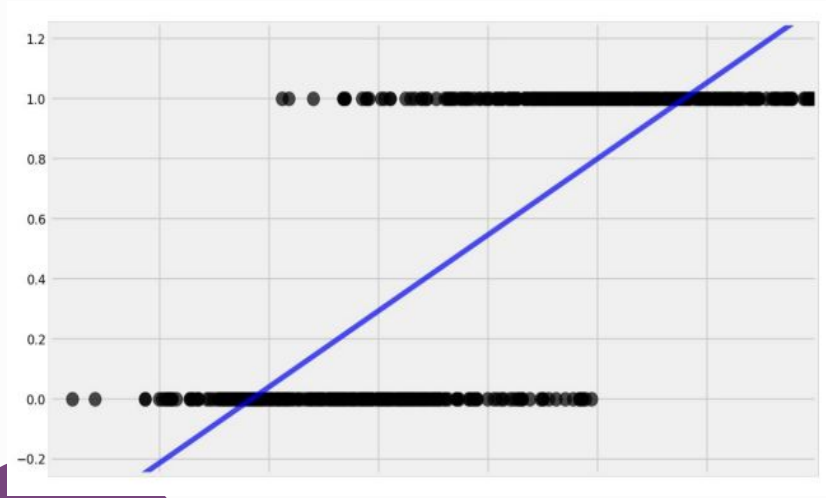
	and	but	hate	i	image	language	like	love	natural	processing	python	signal
Text 1	0	1	1	2	0	1	0	1	1	1	1	0
Text 2	0	0	0	1	1	0	1	0	0	1	0	0
Text 3	1	0	0	1	1	0	1	0	0	2	0	1

Term	and	but	hate	i	image	language	like	love	natural	processing	python	signal
IDF	0.47712	0.47712	0.4771	0	0.1760913	0.477121	0.1760913	0.477121	0.47712125	0	0.477121	0.477121

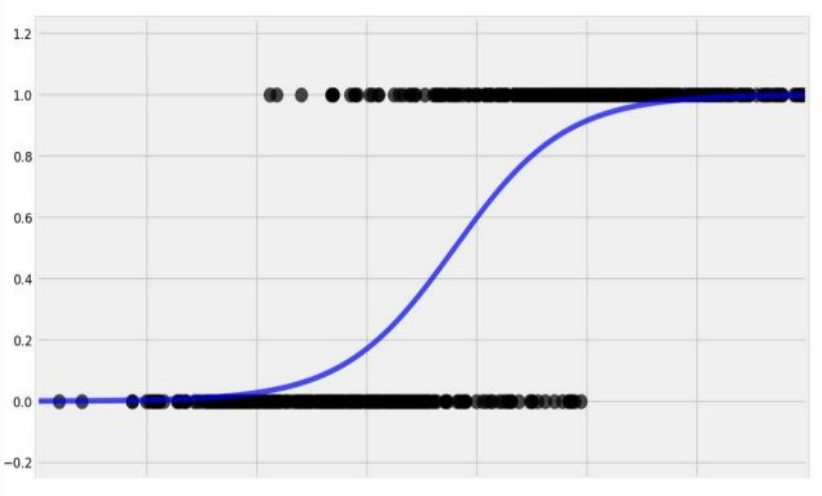
LOGISTIC REGRESSION MODEL

- Logistic regression “bends” our best fit line, to match the range or set of values.
- Useful in predicting binary outcomes.

Linear Regression



Logistic Regression



BUILDING A CLASSIFICATION MODEL

DATA PREPARATION

Converting text to numerical representation

Methods used:

- Countvectorizer
- N-grams
- TF-IDF (Term Frequency-Inverse Document Frequency)

Countvectorizer

Sentence: The Three Musketeers

	The	Three	Musketeers
Sentence	1	1	1

BUILDING A CLASSIFICATION MODEL

DATA PREPARATION

Converting text to numerical representation

Methods used:

- Countvectorizer
- N-grams
- TF-IDF (Term Frequency-Inverse Document Frequency)

Bi-gram

Sentence: The Three Musketeers

	The Three	Three Musketeers
Sentence	1	1

Tri-gram

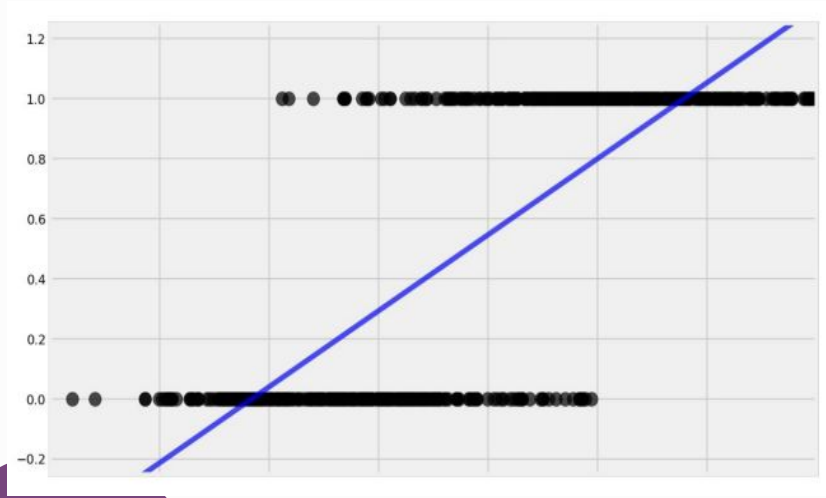
Sentence: The Three Musketeers

	The Three Musketeers
Sentence	1

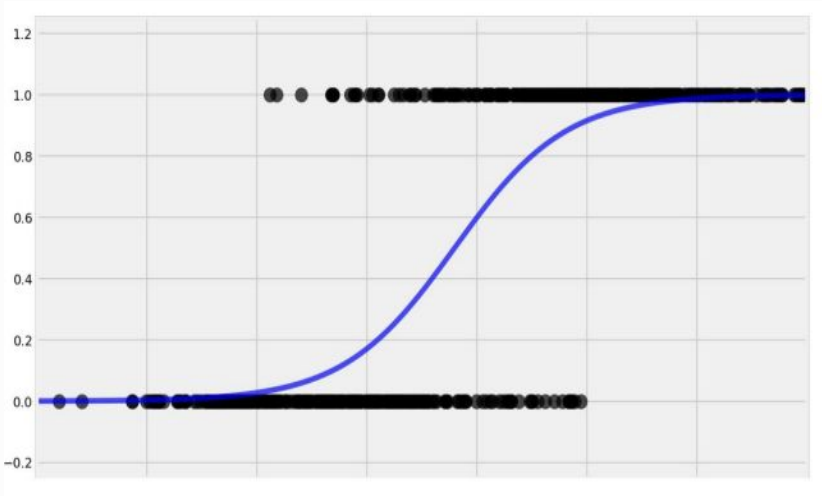
BERNOULLI/ MODEL

- Logistic regression “bends” our best fit line, to match the range or set of values.
- Useful in predicting binary outcomes.

Linear Regression



Logistic Regression

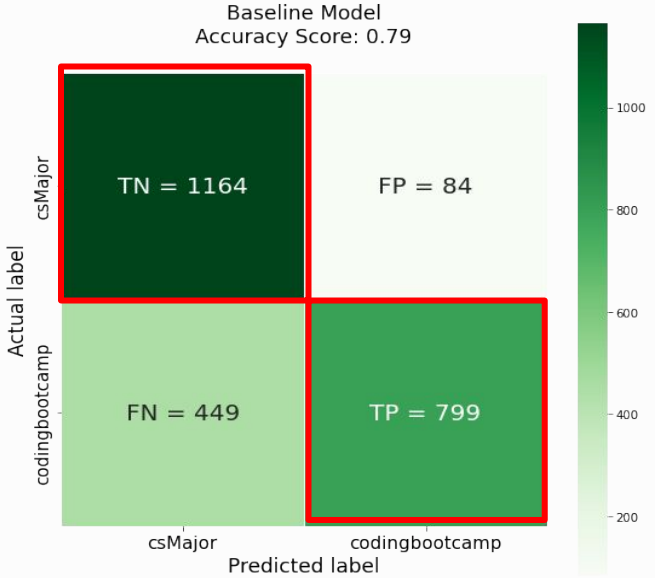


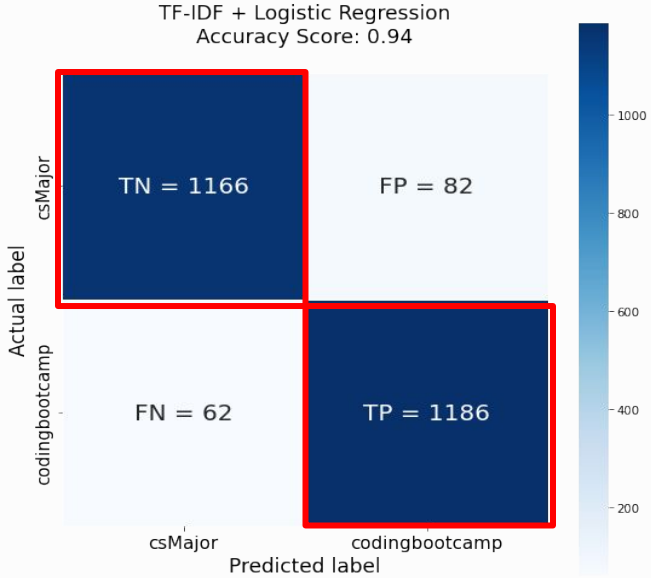
MODEL OPTIMIZATION

VECTORIZATION + MODEL TYPE	PARAMETERS OPTIMIZED	IMPROVEMENT
TF-IDF + Logistic Regression	max features min_df max_df lg_solver	~0.04%

MODEL EVALUATION

CONFUSION MATRIX - HIGHER ACCURACY FOR MODEL





- TN: True Negative, TP: True Positive → Predictions are correct, for either classes
- FN: False Negative, FP: False Positive → Predictions are wrong, for either classes
- Positive class: codingbootcamp, Negative class: csMajor.
- Accuracy = True Predictions / Total Predictions.

CLASSIFICATION REPORT - HIGHER F1-SCORE

BASELINE

Baseline

	precision	recall	f1-score
csMajor	0.72	0.93	0.81
codingbootcamp	0.90	0.64	0.75
accuracy			0.79
macro avg	0.81	0.79	0.78
weighted avg	0.81	0.79	0.78

TF-IDF + LOGISTIC REGRESSION

TF-IDF + Logistic Regression

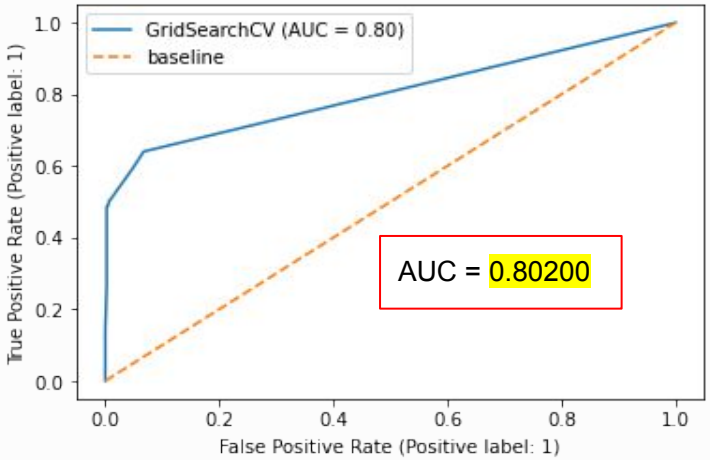
	precision	recall	f1-score
csMajor	0.95	0.93	0.94
codingbootcamp	0.94	0.95	0.94
accuracy			0.94
macro avg	0.94	0.94	0.94
weighted avg	0.94	0.94	0.94

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-Score = Weighted Average of Precision and Recall
  - Offers a better overall measure of performance

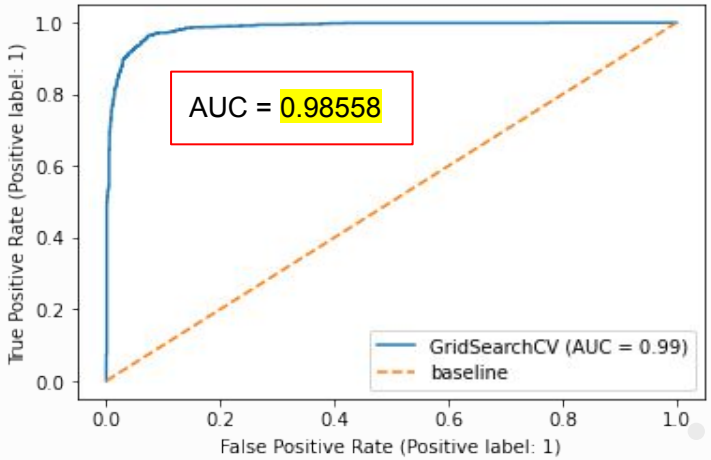
ROC CURVE - HIGHER AUC SCORE

- ROC - Receiver Operating Characteristic Curve
- AUC - Area Under the Curve

Baseline



TF-IDF + Logistic Regression



Higher AUC score

Better differentiation between categories

MOVING FORWARD

TIME & RESOURCES

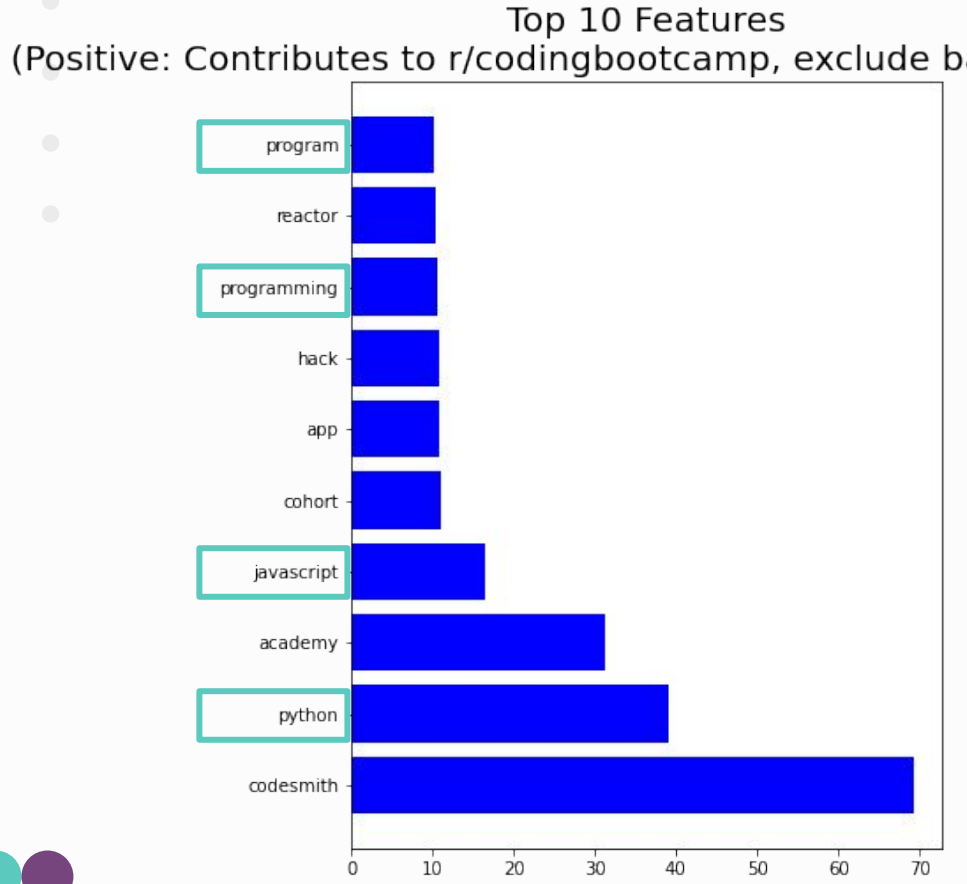
Gather more data to train the model, using information from various platforms

WEB LINGO

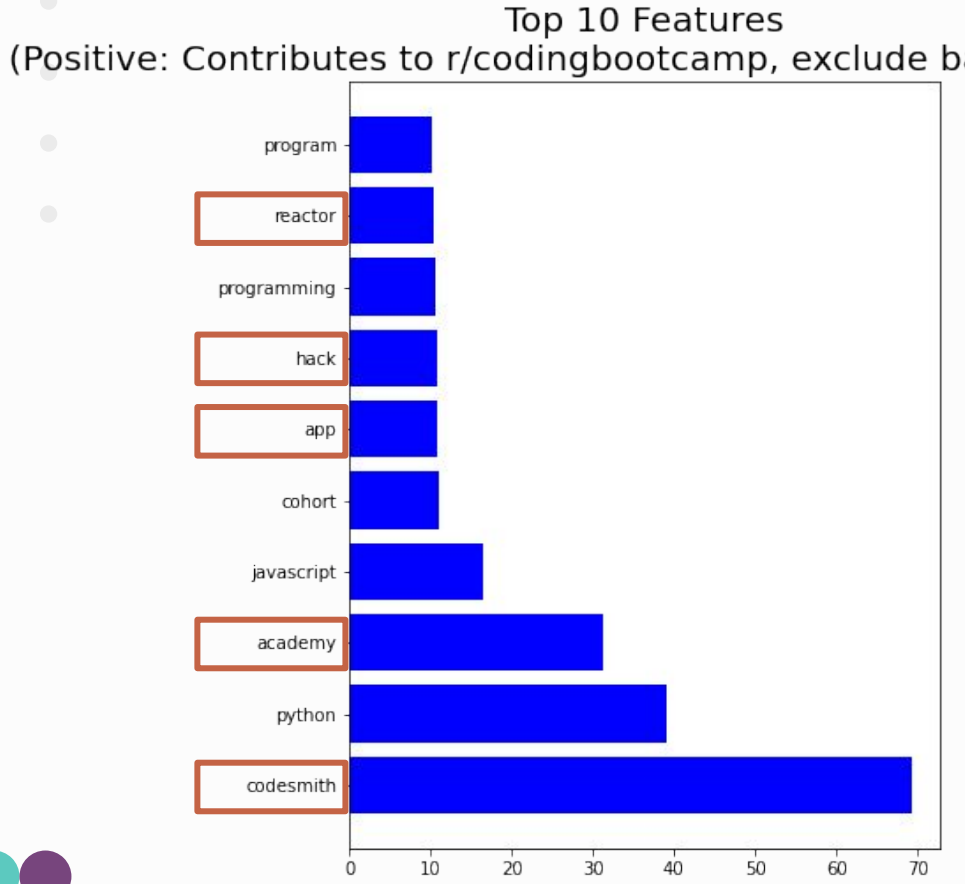
Train the model to better understand acronyms and abbreviations being used

SENTIMENT ANALYSIS

Expand the model to understand the sentiments behind the posts

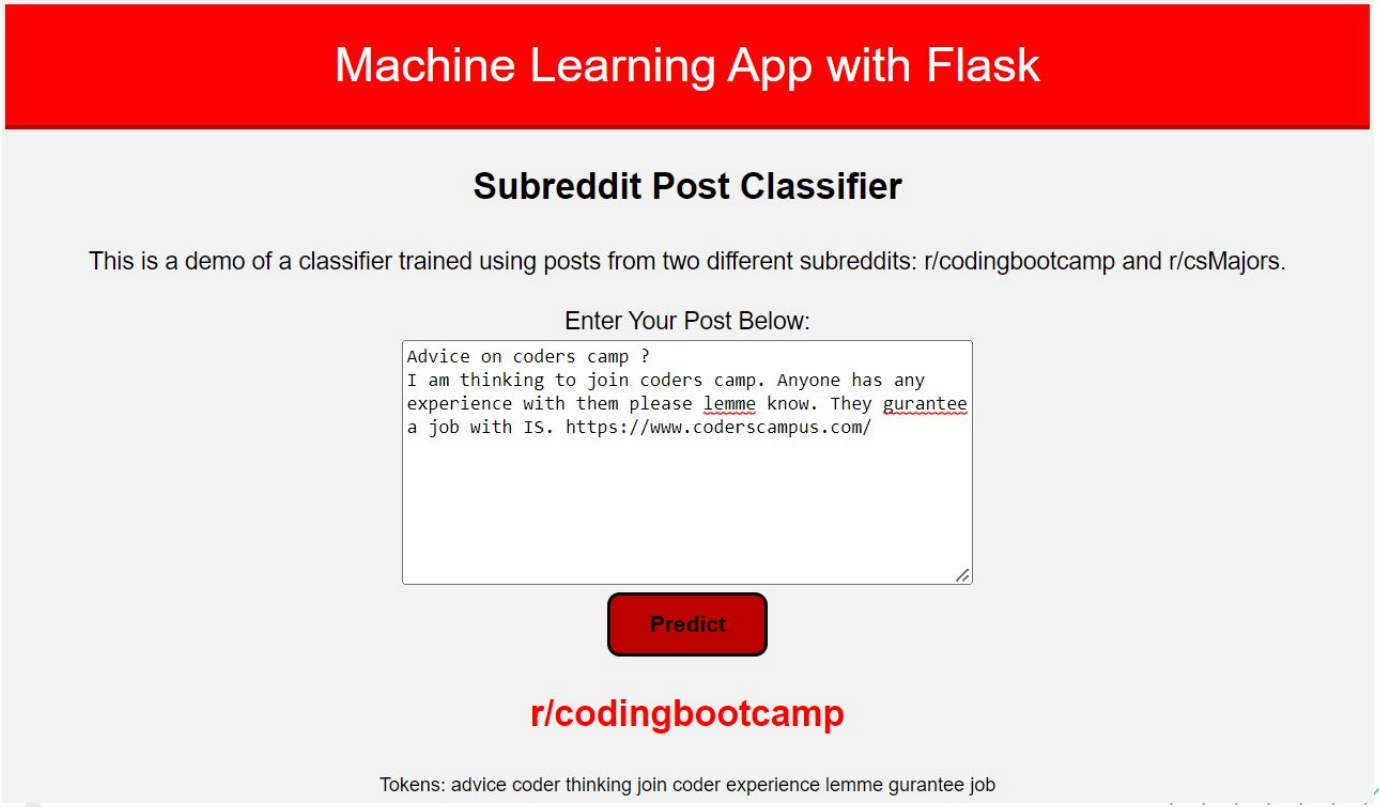


- Skill related features that we can focus on based on courses offered at General Assembly



- Competitors are mentioned more frequently on Reddit
- Creates opportunity for GA to market towards these users

SAMPLE PREDICTIONS



RECOMMENDATIONS

KEYWORDS

Features produced by our model will allow the team to better identify suitable posts to engage with.

AUTOMATION

Deployment of the model to automatically scan our social media interactions.

MARKETING

Boost marketing across channels to increase visibility compared with our competitors.

CONCLUSION

01.

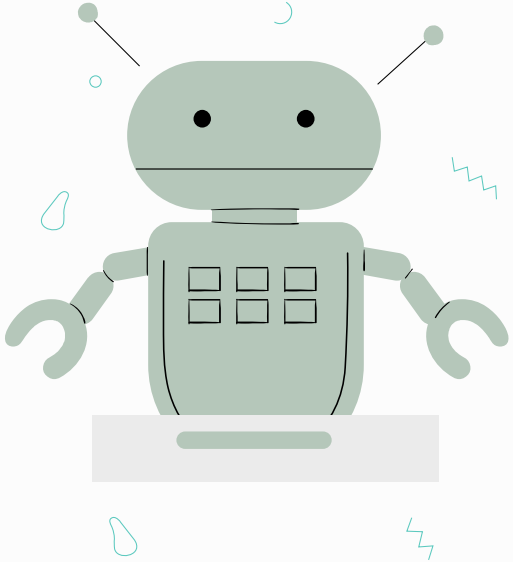
INCREASING VISIBILITY AND RESPONSE

GA needs to stand out from our competitors and speed is also essential in being able to act before our competitors.

02.

SEGMENTING AND TARGETING THE RIGHT AUDIENCE

Maximise our marketing ROI and increase our conversion rate.



THANK YOU