Project 3

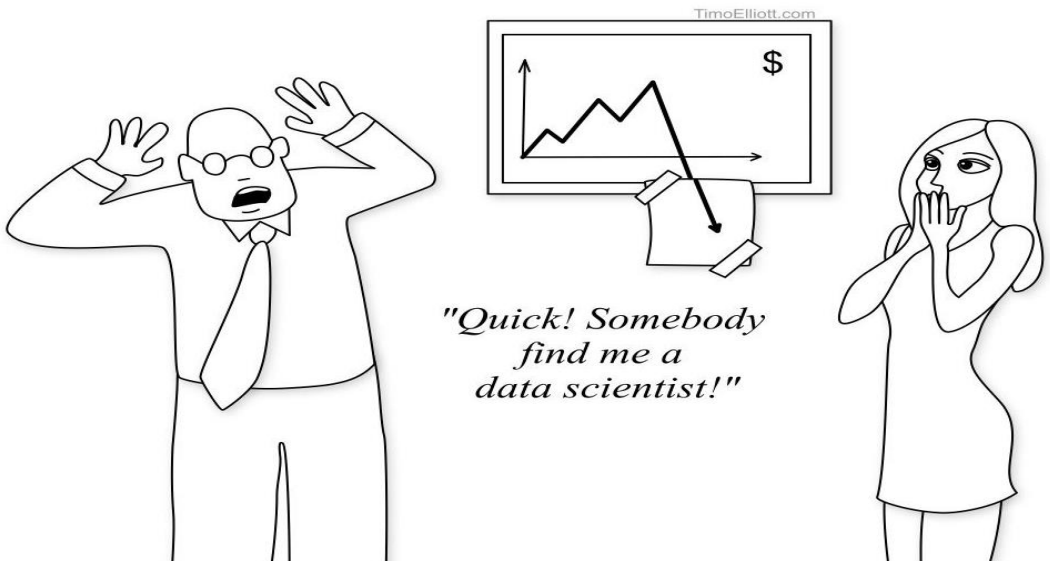
Subreddit Classification

r/datascience
r/dataengineering

Problem Statement

For someone considering a career switch to the data industry,

- Would one be able to decide which path to transit into?
- What are the key differences between the two?



Workflow

1. Data Collection
2. Exploratory Data Analysis
3. Data Cleaning / Pre-processing
4. Model Testing / Evaluation
5. Conclusions and Recommendations

Data Collection

Data collected using Pushshift's API from Reddit's

- r/datascience (1,000 posts)
- r/dataengineering (1,000 posts)

Data collected from to 21st Dec 2021 to 13th Jan 2022

	all_awardings	allow_live_comments	author	author_flair_css_class	author_flair_richtext	author_flair_text	author_flair_type	author_fullname	author_is_
0	[]	False	mercury0100	NaN	[]	NaN	text	t2_i33mpady	
1	[]	False	Fail-Wooden	NaN	[]	NaN	text	t2_8sgme8om	
2	[]	False	pykit_org	NaN	[]	NaN	text	t2_9060ijmj	
3	[]	False	MonteSS_454	NaN	[]	NaN	text	t2_ubz9o	
4	[]	False	1Minnee	NaN	[]	NaN	text	t2_3evxli8x	

	all_awardings	allow_live_comments	author	author_flair_background_color	author_flair_css_class	author_flair_richtext	author_flair_template_id	au
0	[]	False	urs123	transparent	NaN	[]	fbc7d3e6-ac9c-11eb-adda-0e0b12e4a59b	
1	[]	False	manish_ks	NaN	NaN	[]	NaN	
2	[]	False	Junior_Abies_2213	NaN	NaN	[]	NaN	
3	[]	False	LowProgram6449	NaN	NaN	[]	NaN	
4	[]	False	barnyard9	NaN	NaN	[]	NaN	

Exploratory Data Analysis

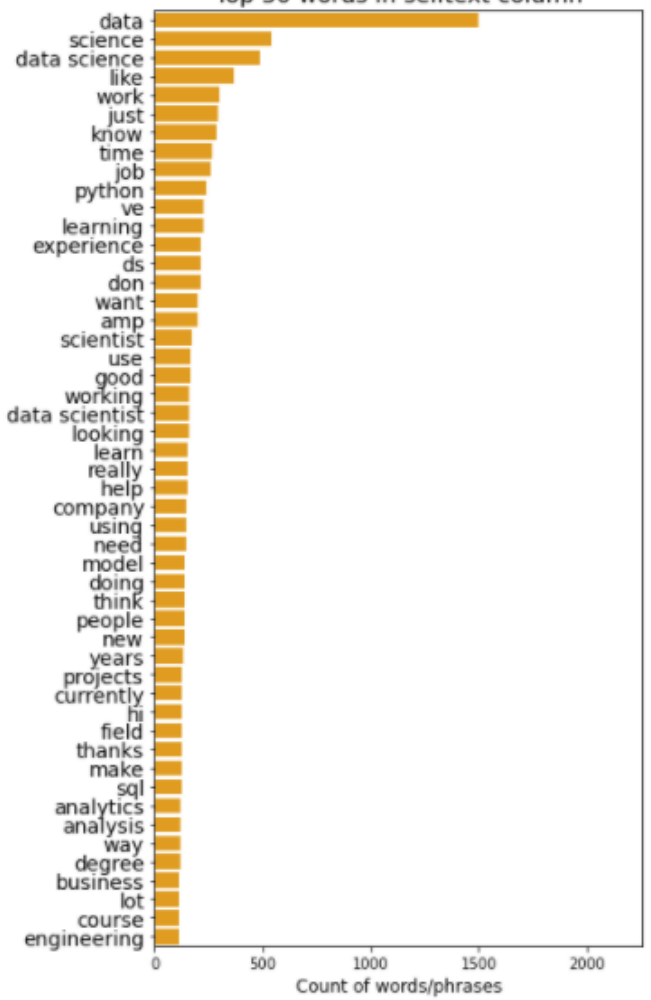
Top 50 words using unigrams and bigrams

- Heavily skewed towards words existing in subreddit topics
- Contains words from Reddit specific markdown formatting
- Filler words (‘hi’, ‘like’, ‘want’)

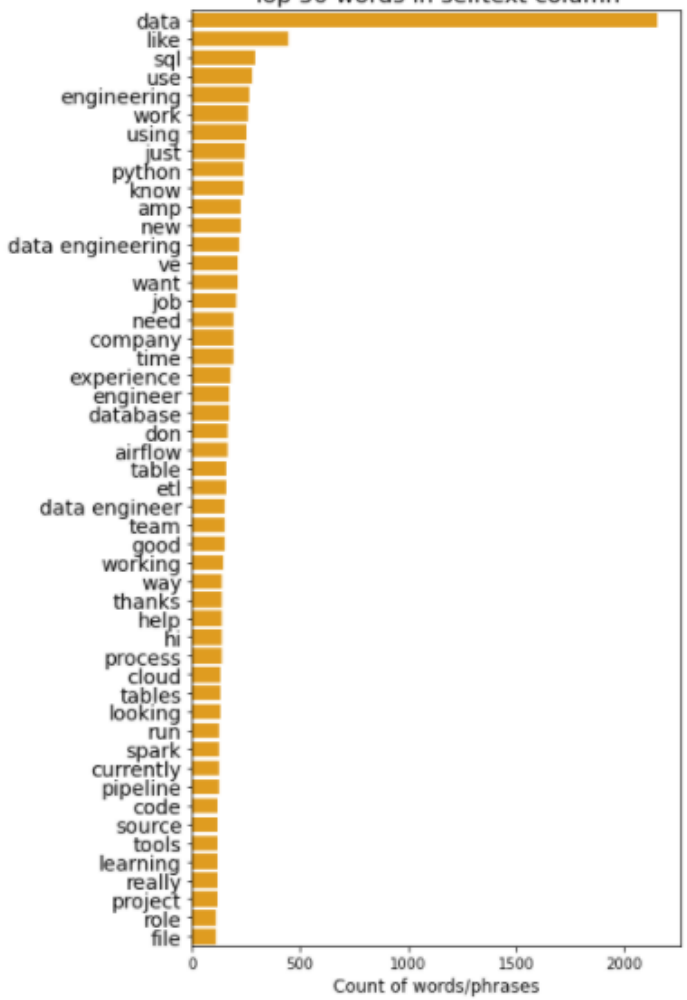
Text Cleaning

- Converted to lowercase
- Removed HTML links, special characters, whitespaces, Reddit specific markdowns
- Lemmatization

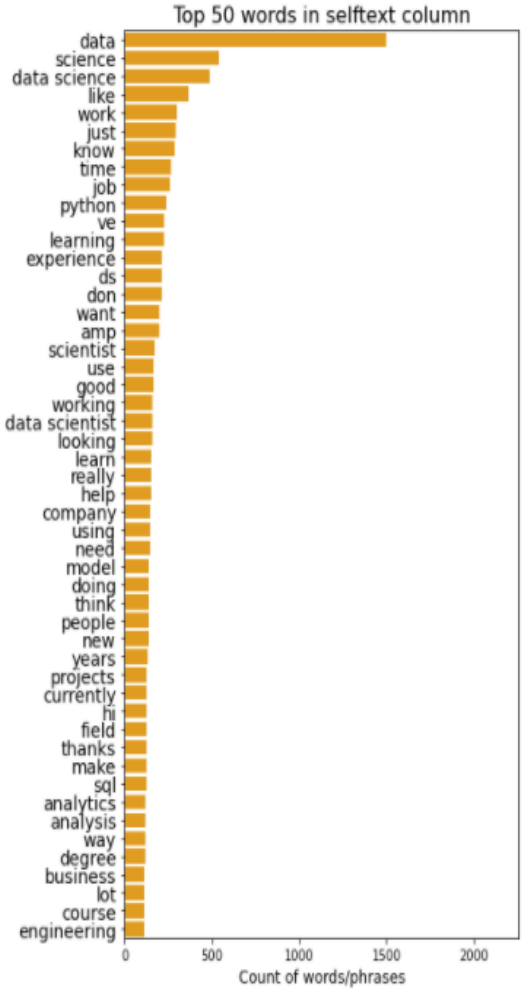
Data Science

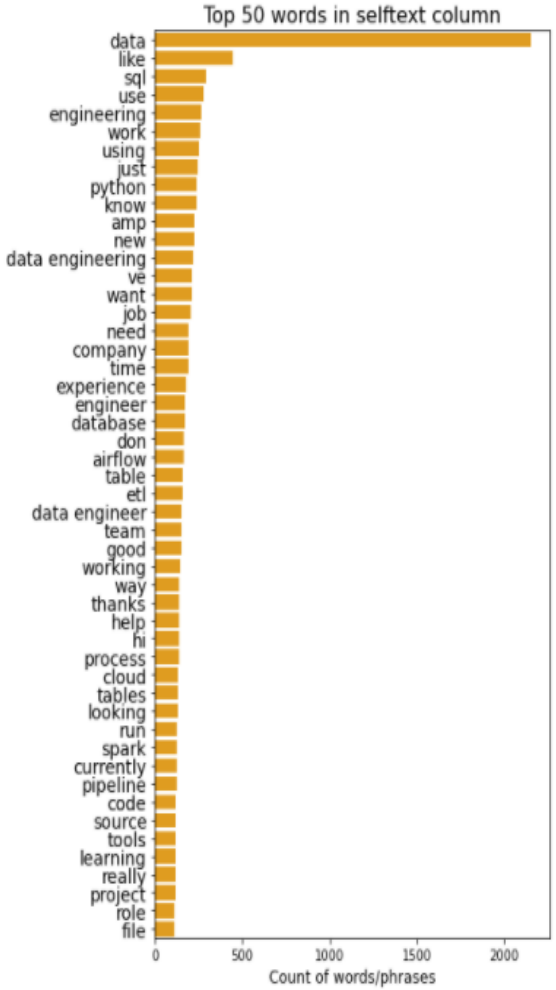


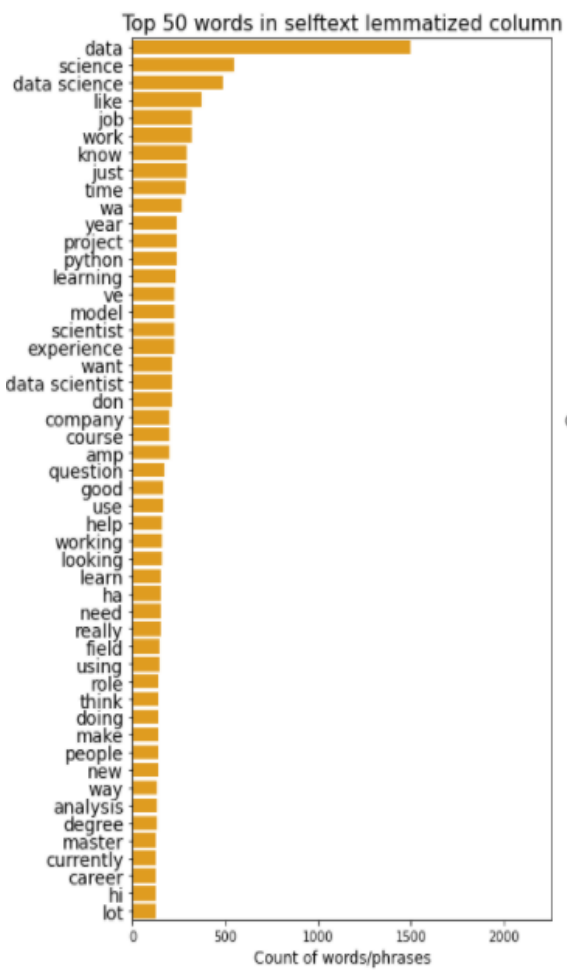
Data Engineering

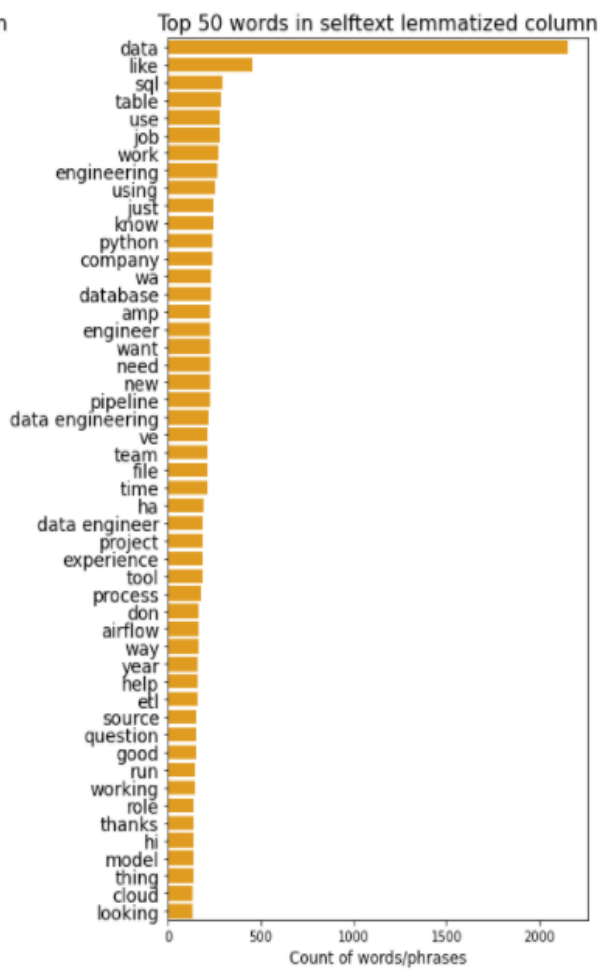


Text Cleaning









Pre-processing

1. Dropped removed and deleted posts
2. Text cleaning
3. Fill null values in selftext columns
4. Lemmatization
5. Merging title and selftext columns
6. Final number of rows = 1901 rows

lemma_comb
edge regression with edge feature in graph neu...
aspiring to be data analyst can someone sugges...
how to build your own chatbot in python using ...
d bi graduate certificate v master of d c degr...
target salary for fresh phd grad in stats hi a...
hey high school student in mexico and into con...
how much of your workload is assigned to you v...
an approachable introduction to the bayesian o...
how much of your workload is assigned to you v...
merge sort part
finding part time d work hey guy doe anyone kn...

Model Testing

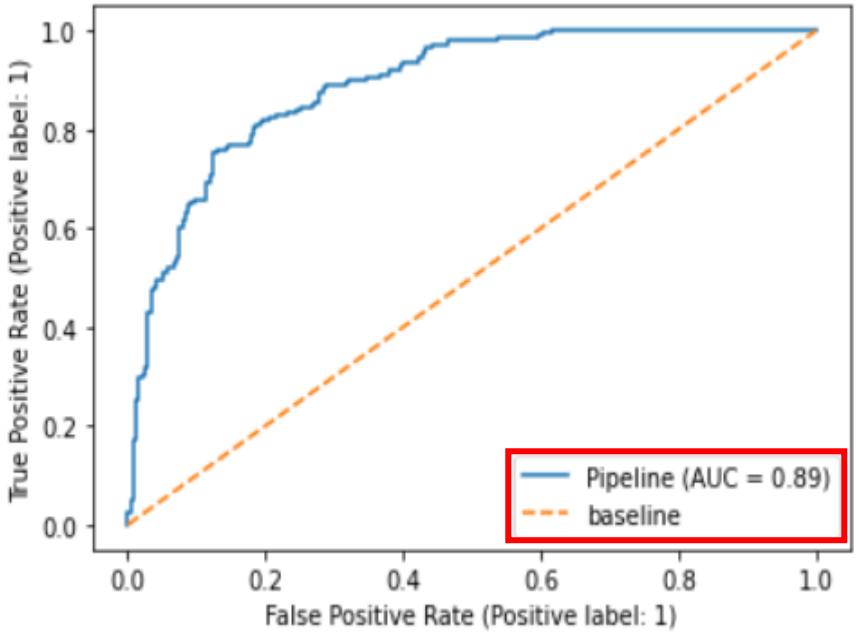
Models tested

- Multinomial Naïve Bayes (CountVectorizer)
- Multinomial Naïve Bayes (TfidfVectorizer)
- Logistic Regression (CountVectorizer)
- Logistic Regression (TfidfVectorizer)
- Random Forest (CountVectorizer)
- Random Forest (TfidfVectorizer)

Model Performance

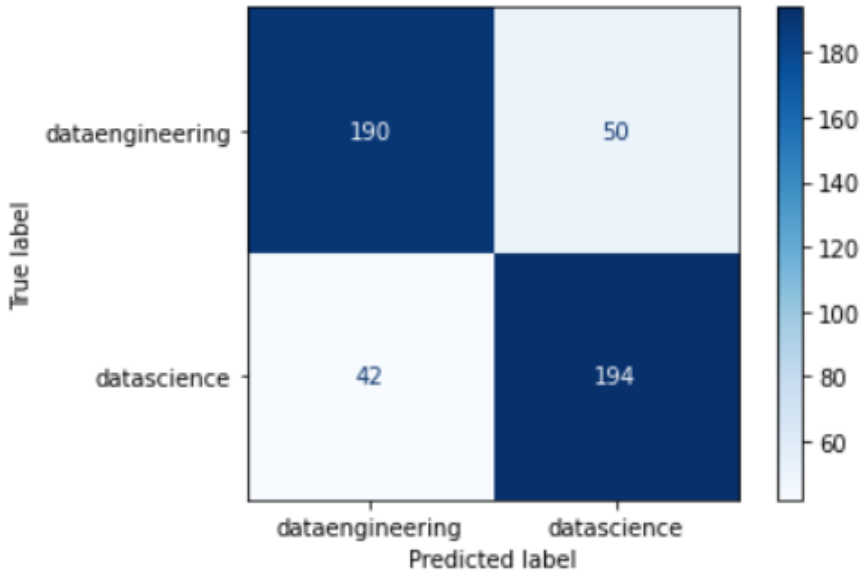
	Train Accuracy	Test Accuracy
Multinomial Naive Bayes (CountVectorization)	0.847	0.760
Multinomial Naive Bayes (TfidfVectorization)	0.905	0.798
Logistic Regression (CountVectorization)	0.982	0.756
Logistic Regression (TfidfVectorization)	0.943	0.806
Random Forest Classifier (CountVectorizer)	0.994	0.773
Random Forest Classifier (TfidfVectorizer)	0.994	0.781

Model Performance

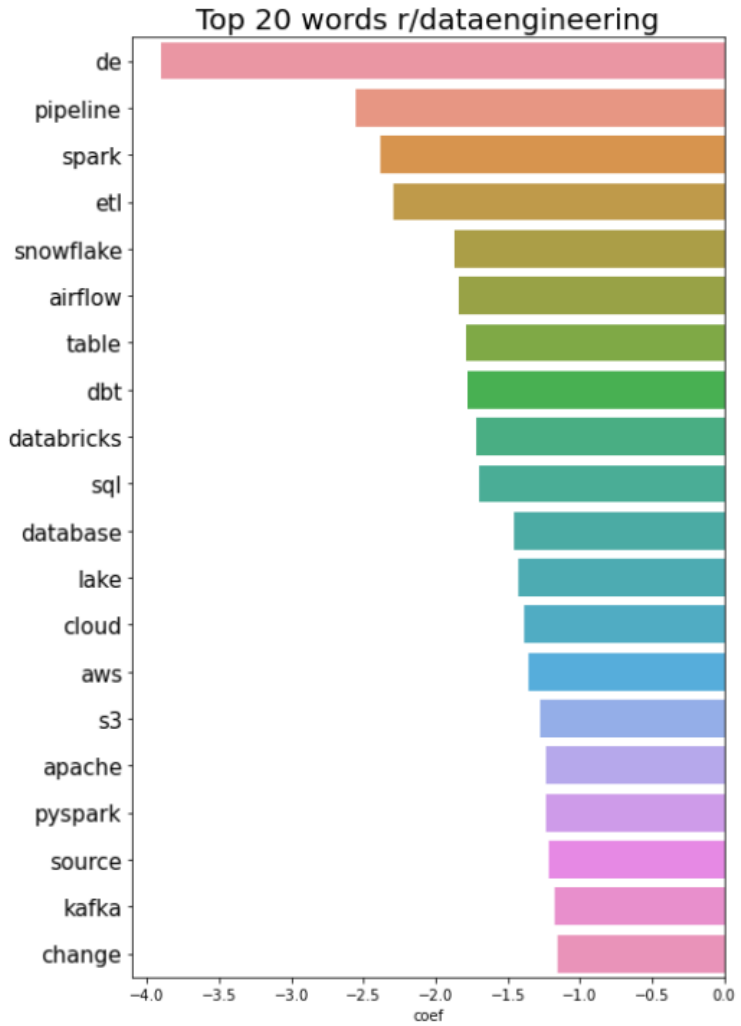


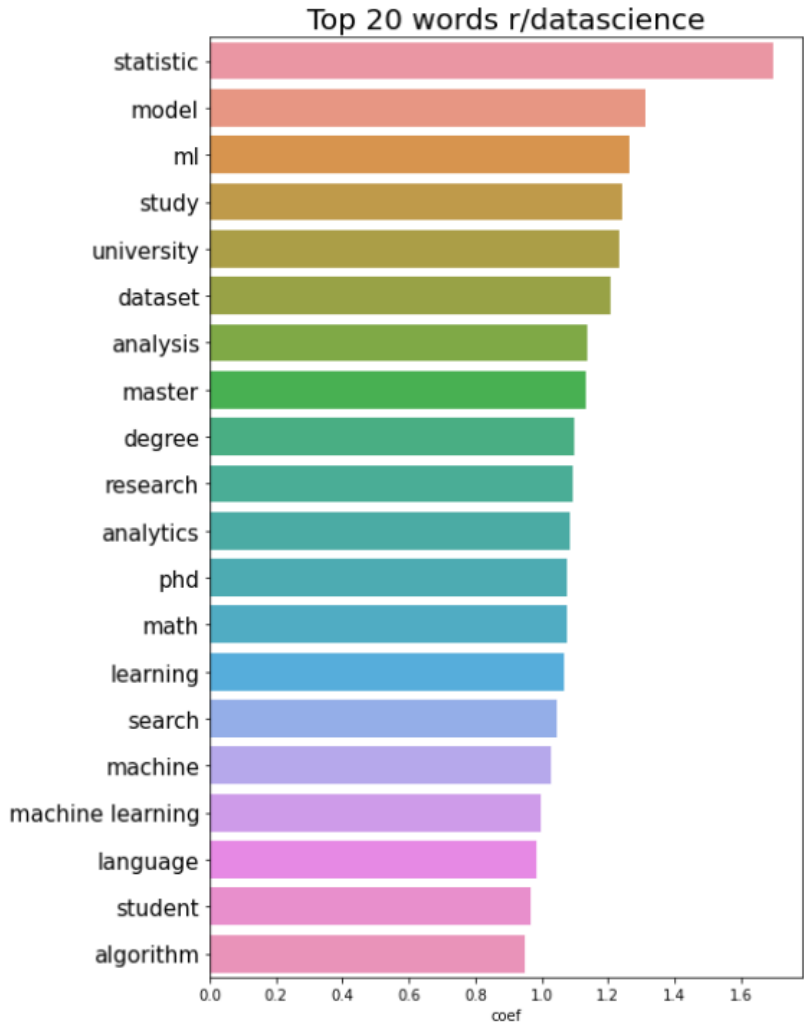
	precision	recall	f1-score	support
0	0.82	0.79	0.81	240
1	0.80	0.82	0.81	236
accuracy			0.81	476
macro avg	0.81	0.81	0.81	476
weighted avg	0.81	0.81	0.81	476

Sensitivity: 0.822
Specificity: 0.7917



Top 20 words from each subreddit





Conclusions and Recommendations

- Accuracy score of 80% is a reasonable score given the relative similarity between both subreddits.

- Data Science
  - Topics tend to tap on one's academic strength
  - High barrier of entry as suggested by the top 20 words
  - Suitable for people who likes working with datasets, machine learning models and have a good understanding of statistics

- Data Engineering
  - Focuses on the tools necessary for data engineering
  - Familiarity with tools such as apache, spark, snowflake, airflow and ETL (extract, transform and load)
  - Lower barrier of entry compared to data scientists, values experience over education
  - Suitable for people who likes to develop systems for data collection and processing

- Future steps
  - Collect more data / information
  - Increase range of models to test on, such as SVM, K Nearest Neighbour, Word2Vec and GloVE.