PROJECT 3 : WEB API & CLASSIFICATION

MAK KWOK FAI

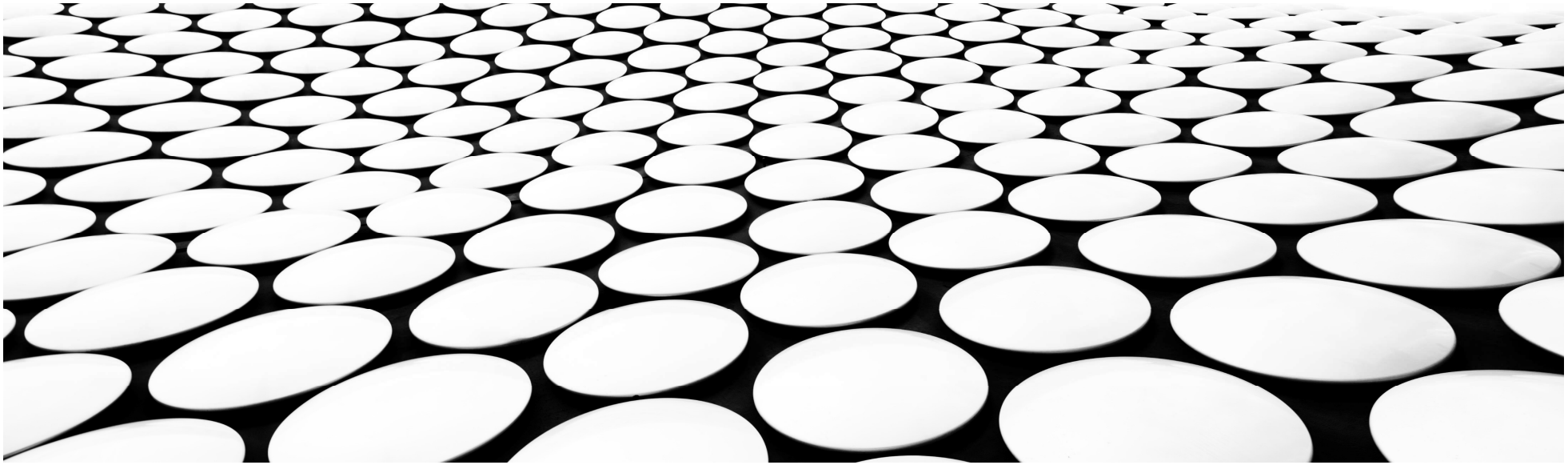
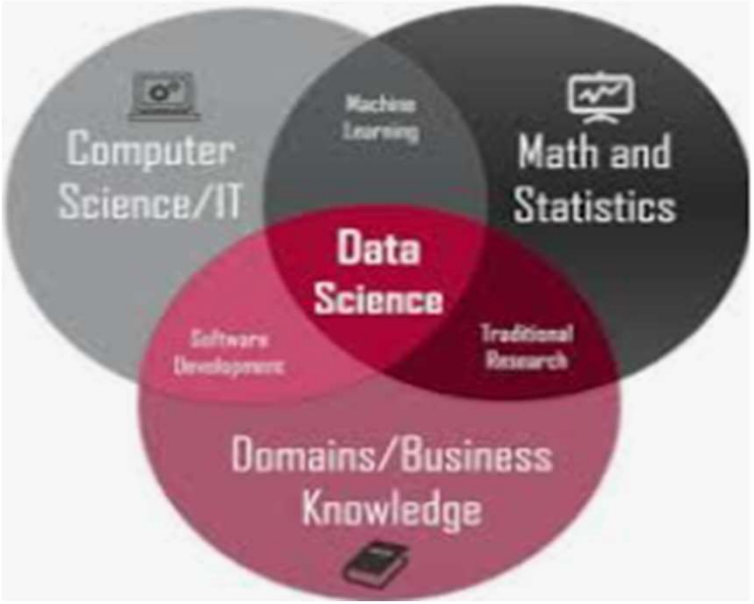


TABLE OF CONTENT

PROBLEM STATEMENT

- Using Reddit's API, I will collect posts from 2 subreddits.
- Train a classifier on which subreddit a given post came from.

DATA COLLECTION



COMPUTING DATA SCIENCE

DATA COLLECTION

- /r/computing/
- 24.1k members
- This subreddit is for links to articles related to computing and help posts.
- Created in 2011
- 8,600+ submissions

- /r/datascience/
- 341k members
- A place for data science practitioners and professional to discuss and debate data science career questions.
- Created in 2009
- 45,000+ submission

COMPUTING

DATA SCIENCE

DATA CLEANING

Computing

Subreddit	Selftext	Author_fullname	Title	. . .	Num_crossposts	Media	subreddit-_subscribers	Is-video
Computing	My PC started …	T2_1gvl4g2	PC Not Turni…	. . .	0	None	24098	False
Computing		T2_8bysr4r1	Join us every…	. . .	0	None	24098	False
Computing	If AX contain 100…	T2_616lursp	Can you help…	. . .	0	None	24098	False
Computing	Who's not getting …	T2_92a3odo1	Secure Data …	. . .	0	None	24098	False

Data Science

Subreddit	Selftext	Author_fullnam e	Title	. . .	Num_crossposts	Media	subreddit-_subscribers	Is-video
Data Science	Welcome to …	T2_414cxw07	Weekly entering …	. . .	0	None	340541	False
Data Science		T2_886htd2x	Are there any …	. . .	0	None	340541	False
Data Science	My company …	T2_140c97	Data Science is …	. . .	0	None	340541	False
Data Science	I'm an undergraduate…	T2_opauw	Undergraduate …	. . .	0	None	340541	False

DATA CLEANING

Computing

Subreddit	Selftext	Title
Computing	My PC started acting weird about a week ago, I ...	PC Not Turning On
Computing		Join us every Saturdays for a free awesome ...
Computing	If AX contain 1000101000110001 then what is the ...	Can you help me answer this?
Computing	Who's not getting flooded with 'personalized' ...	Secure Data Collaboration.

Data Science

Subreddit	Selftext	Title
Data Science	Welcome to this week's entering; transiting ...	Weekly Entering; Transitioning Thread
Data Science		Are there any people who started off with data ...
Data Science	My company hired a three man team of data scientist ...	Data Science is not an easy and quick job ...
Data Science	I'm an undergraduate junior applied for Data Science internship ...	Undergraduate Data Science internships

DATA CLEANING

Concatenate 2 Dataframe

Subreddit	Selftext	Title
Computing	My PC started acting weird about a week ago, I ...	PC Not Turning On
Computing	If AX contain 1000101000110001 then what is the ...	Can you help me answer this?
Computing	Who's not getting flooded with 'personalized' ...	Secure Data Collaboration.
Data Science	Welcome to this week's entering; transiting ...	Weekly Entering; Transitioning Thread
Data Science	My company hired a three man team of data scientist ...	Data Science is not an easy and quick job ...
Data Science	I'm an undergraduate junior applied for Data Science internship ...	Undergraduate Data Science internships

DATA CLEANING

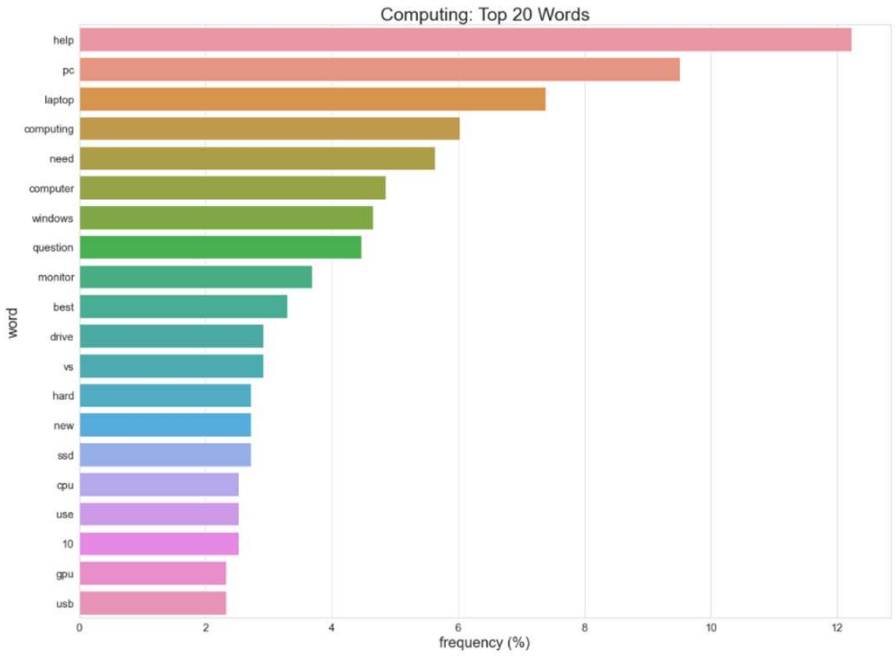
- Step 1 : Change subreddit titles to 0 and 1
- Step 2 : Remove HTML
- Step 3 : Remove non-letters
- Step 4 : Convert words to lower cases
- Step 5 : List stopwords
- Step 6 : Remove stopwords
- Step 7 : Combine “Selftext” & “Title” columns into a “New” columns

DATA CLEANING

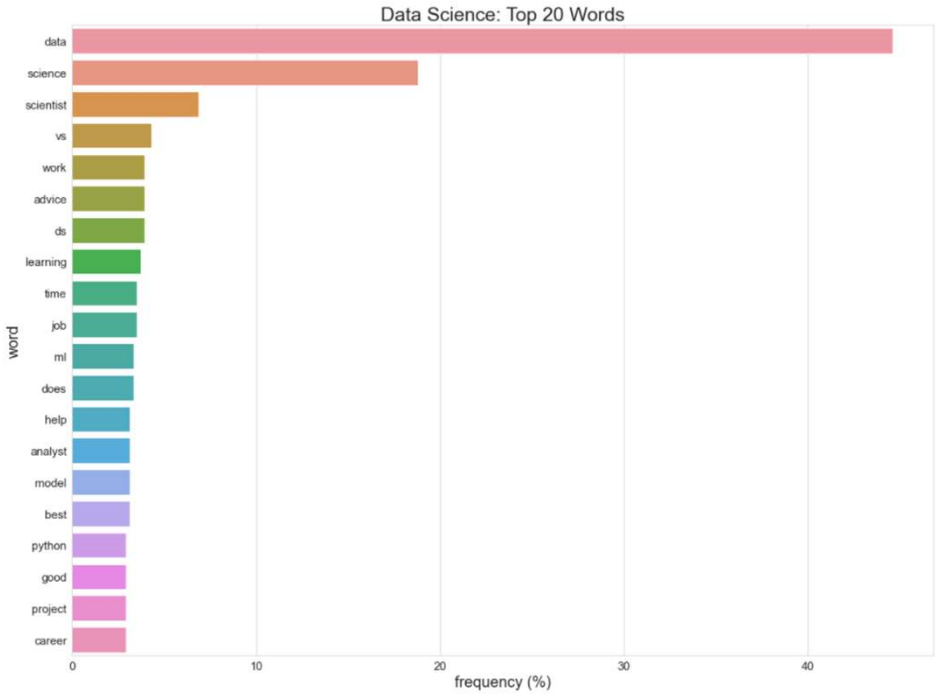
After applying the function ...

Subreddit	New
0	pc turning pc started acting weird week ago would ...
0	help answer ax contain value found shift ...
0	secure collaboration exclusive sneak peak getting ...
1	question possible project this would like starting ...
1	Percentage non traditional path scientist ...
1	Permutation importance sklearn highly correlated ...

EDA - MOST FREQUENT WORDS

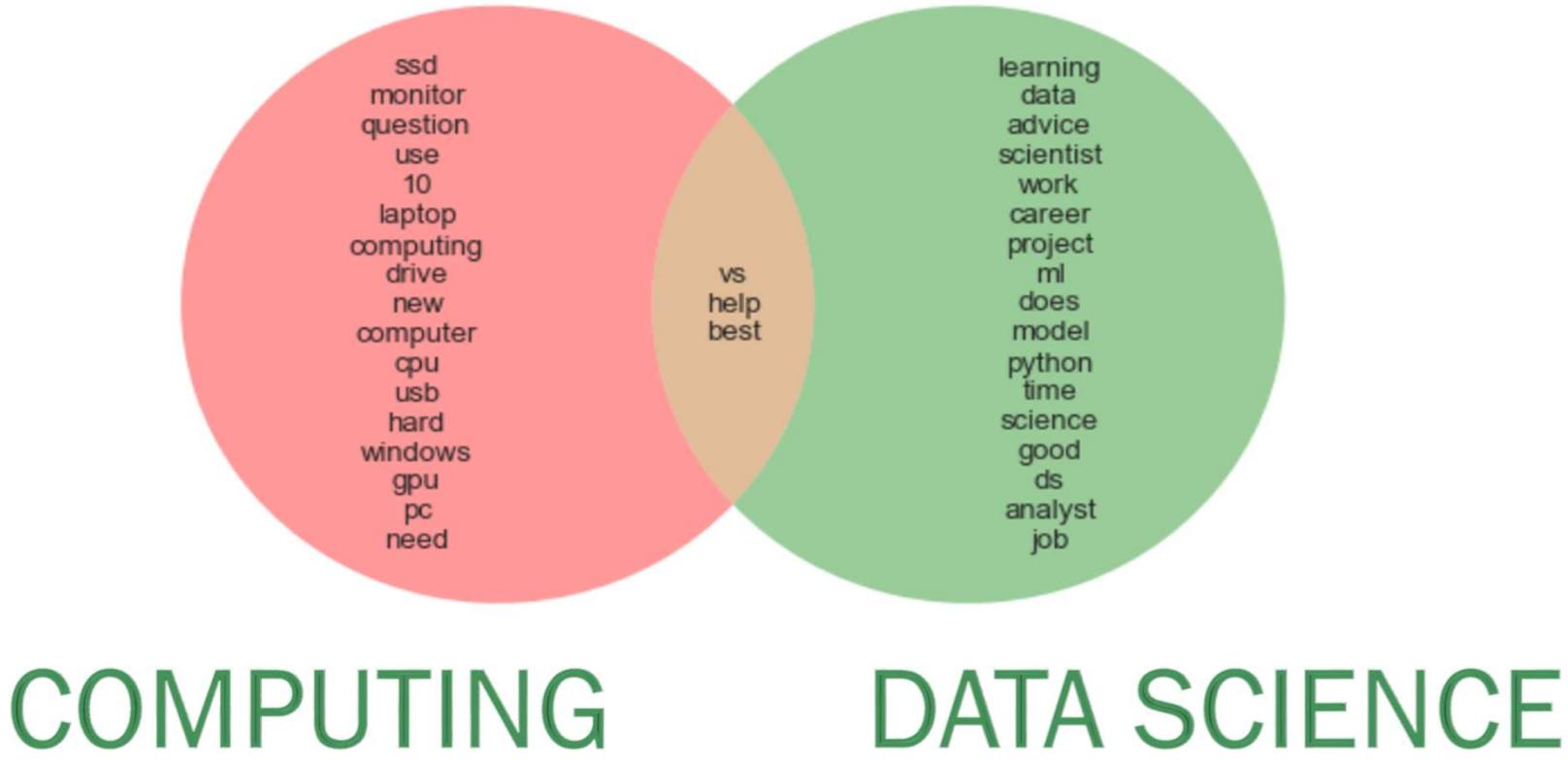


COMPUTING



DATA SCIENCE

EDA – WORD COMPARISON



CREATE FEATURE MATRIX AND TARGET

- X = df ['New']
- y = df ['subreddit']

- vectorizer = CountVectorizer()
- X_train , X_test , y_train , y_test = train_test_split (X , y , test_size = 0.25 , random_state = 42)

- X_train_vec = vectorizer.fit_transform (X_train)
- X_test_vec = vectorize.transform (X_test)

MODELING – NAIVE BAYES

	Pred Positive	Pred Negative
True Positive	108	19
True Negative	3	127

Training Score : 0.979
Testing Score : 0.914

MODELING – NAIVE BAYES

- Accuracy : 0.914
- Sensitivity : 0.977
- Specificity : 0.850
- Precision : 0.870

MODELING – LOGISTIC REGRESSION

	Pred Positive	Pred Negative
True Positive	114	13
True Negative	20	110

Training Score : 1
Testing Score : 0.872

MODELING – LOGISTIC REGRESSION

- Accuracy : 0.872
- Sensitivity : 0.846
- Specificity : 0.898
- Precision : 0.894

CONCLUSION & RECOMMENDATION

By comparing the result of these 2 classifiers, we found that the “Naive Bayes” classifier is able to predict better on unseen data.

With a testing score of 91.4%, it can give a good prediction on which subreddit does a new post comes from.