DSI Project 3

HIKING OR GARDENING?

CLASSIFYING TEXTS FROM TWO SUBREDDITS

Report By Aurora Victoria David

PRESENTATION OUTLINE

01 PROBLEM

02 DATA COLLECTION

03 DATA CLEANING & EDA

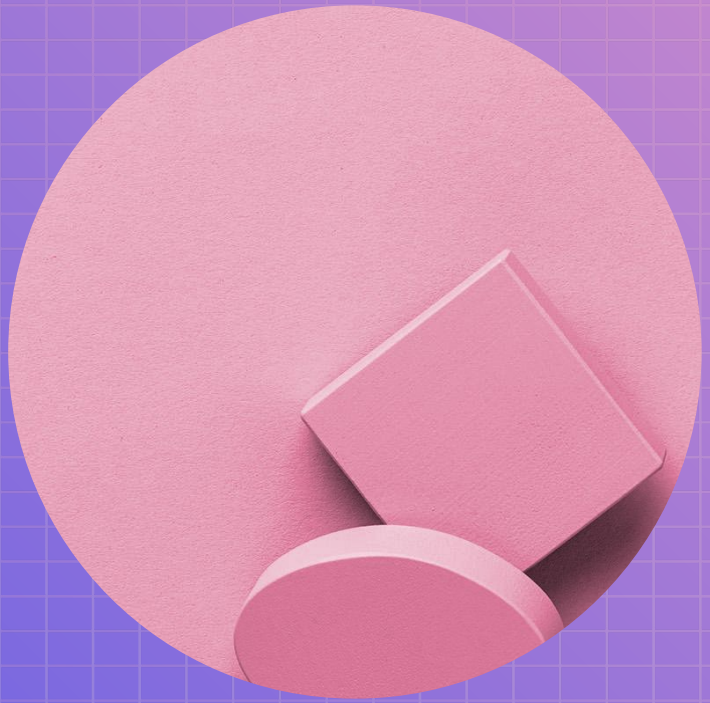
04 PREPROCESSING & MODELING

05 MODEL EVALUATION

06 CONCLUSION

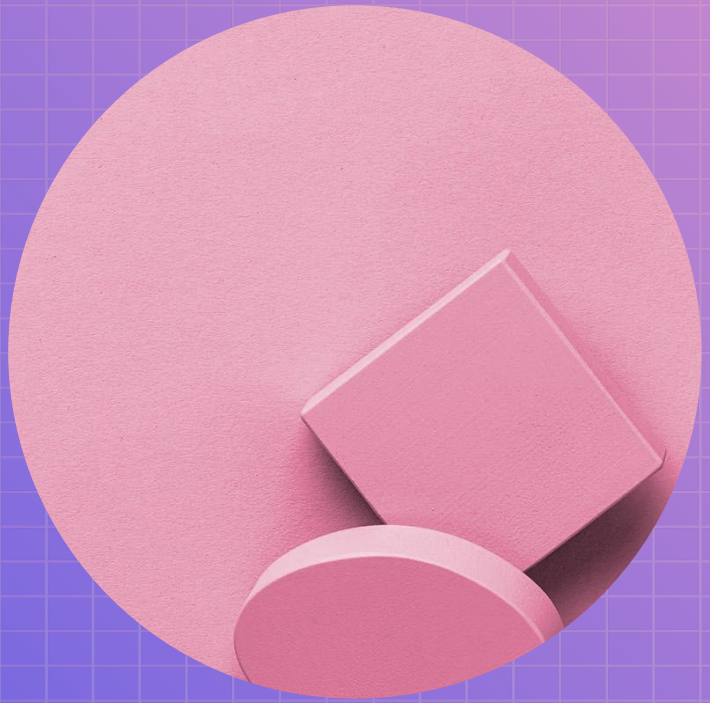
"What model best predicts the topic of a subreddit post?"

Topics: Hiking and Gardening



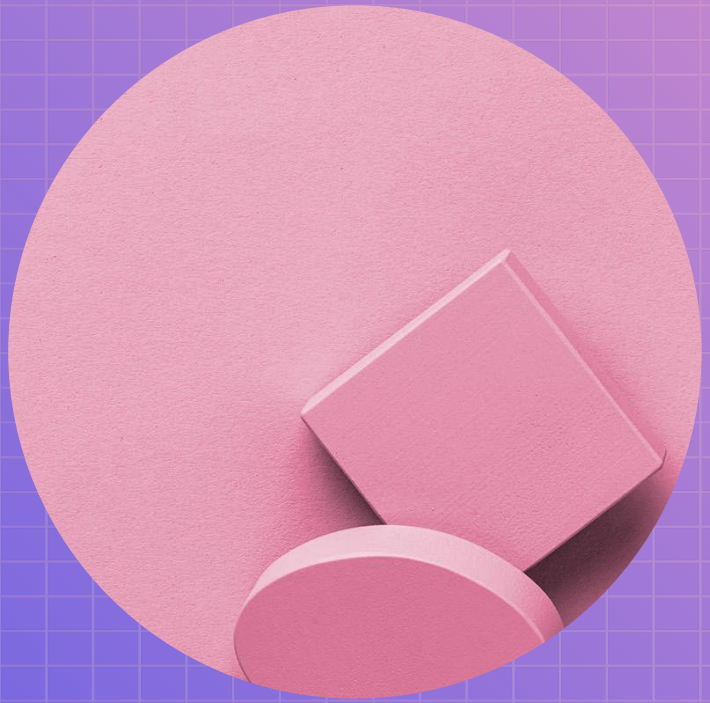
DATA COLLECTION

- Created a function
- Inputs are topic and frequency
- Merges all dataframes and saves final dataframe into csv file
- Prints length of each dataframe to confirm number of posts successfully pulled



DATA COLLECTION

- Hiking
  - 5,000 submissions
  - 2,000 comments
- Gardening
  - 3,000 submissions
  - 2,500 comments
- Total: 12,500 posts



DATA CLEANING

- Dropped “deleted” posts (left with 12,453 posts)
- Dropped duplicates
- Removed hyperlinks, emoticons, emojis, character groupings, punctuation marks
- Made text lowercase
- Left with 11,857 posts

PREPROCESSING AND MODELING

- More deep cleaning
- Lemmatization

Preprocessing

- Vectorizing
- Standard Scale

Logistic Regression

- Vectorizing

Bayes

Null model

56% accuracy rate

TWO FUNCTIONS

MODEL FIT & SCORES

- Returns accuracy and balanced accuracy scores on train and test datasets
- Returns best parameters of gridsearch

MODEL SAVE SCORES

- Appends model scores and other info into a dataframe for comparison

Estimator	Vectorizer	Data Cleaning	Lemmatized	Accuracy Score	Balanced Accuracy Score
Bayes	count	minimal	no	0.9171483622	0.9163733084
Bayes	count	deep	yes	0.9129848229	0.9114615623
Bayes	tfidf	deep	yes	0.9059021922	0.9049318448
Logistic Regression	count	deep	yes	0.9042158516	0.902283386
Bayes	tfidf	deep	no	0.9032040472	0.9014473942
Logistic Regression	tfidf	minimal	no	0.9020552344	0.8987845817
Logistic Regression	count	minimal	no	0.9020552344	0.8979094269
Logistic Regression	tfidf	deep	yes	0.8984822934	0.8965921935
Bayes	count	deep	no	0.897470489	0.8955302767
Logistic Regression	count	deep	no	0.8971332209	0.8933438013

Estimator	Vectorizer	Data Cleaning	Lemmatized	Accuracy Score	Balanced Accuracy Score
Top 1: Bayes	count	minimal	no	0.9171483622	0.9163733084
Top2: Bayes	count	deep	yes	0.9129848229	0.9114615623

Top 1:

Countvectorizer
max_features: 6700
ngram_range: (1, 1)
stop_words: english

Bayes alpha: 0.12

Top 2:

Countvectorizer
max_features: 5000
ngram_range: (1, 1)
stop_words: english

Bayes alpha: 0.07

Estimator	Vectorizer	Data Cleaning	Lemmatized	Accuracy Score	Balanced Accuracy Score
Logistic Regression	count	deep	yes	0.9042158516	0.902283386
Logistic Regression	tfidf	deep	yes	0.8984822934	0.8965921935
Logistic Regression	count	deep	no	0.8971332209	0.8933438013

Countvectorizer
max_features: 4500
ngram_range: (1, 1)
stop_words: None

Logistic Regression C: 0.003

Estimator	Vectorizer	Data Cleaning	Lemmatized	Accuracy Score	Balanced Accuracy Score
Bayes	count	deep	yes	0.9129848229	0.9114615623
Bayes	tfidf	deep	yes	0.9059021922	0.9049318448
Bayes	tfidf	deep	no	0.9032040472	0.9014473942
Bayes	count	deep	no	0.897470489	0.8955302767

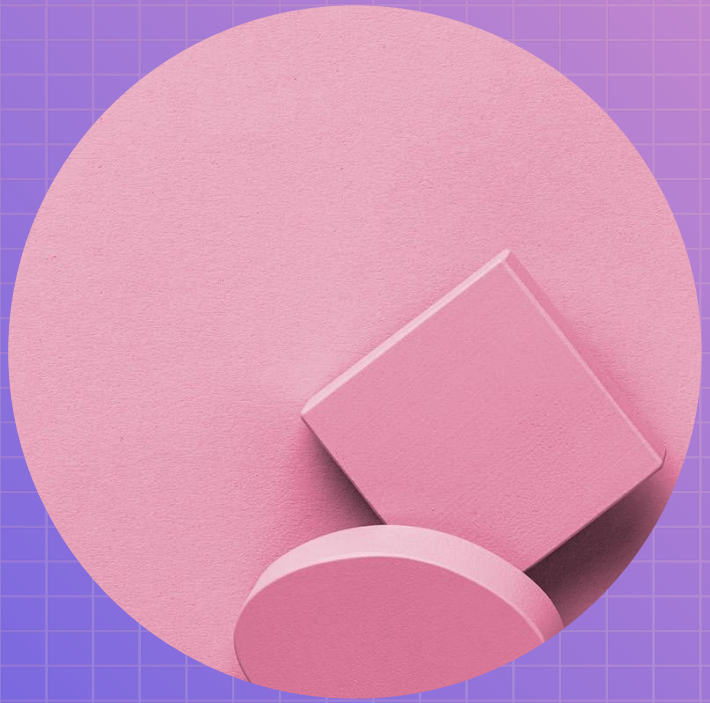
```
Countvectorizer
max_features: 5000
ngram_range: (1, 1)
stop_words: english

Bayes alpha: 0.07
```

Hiking

Top Words	Bayes	Logistic Reg
1	hike	hike
2	trail	trail
3	park	hiking
4	usa	mountain
5	mountain	park
6	national	colorado
7	day	lake
8	lake	usa
9	fall	view
10	state	boot

Gardening

Top Words	Bayes	Logistic Reg
1	plant	plant
2	grow	garden
3	like	grow
4	garden	flower
5	just	tomato
6	year	tree
7	look	bloom
8	tree	tulip
9	help	zone
10	leave	gardening



RECOMMENDATIONS

- Use Bayes model
- Be discerning in data cleaning
- Lemmatize
- Consider both countvectorizer and tfidfvectorizer

Estimator	Vectorizer	Data Cleaning	Lemmatized	Accuracy Score	Balanced Accuracy Score
Top 1: Bayes	count	minimal	no	0.9171483622	0.9163733084
Top2: Bayes	count	deep	yes	0.9129848229	0.9114615623

Top 1:

Countvectorizer
max_features: 6700
ngram_range: (1, 1)
stop_words: english

Bayes alpha: 0.12

Top 2:

Countvectorizer
max_features: 5000
ngram_range: (1, 1)
stop_words: english

Bayes alpha: 0.07

THANK YOU!

Questions?