Reddit:

Zerowaste & environment

Problem Statement

The project is to find out the differences between these two environmental groups.

The groups are chosen due to their relative closeness of their nature in the topics.

Data

r/Environment	r/zerowaste
1351 post	2701 post
1351 titles	2701 titles
63 selftext	205 selftext

Workflow

Web Scraping With Json API

Data Cleaning

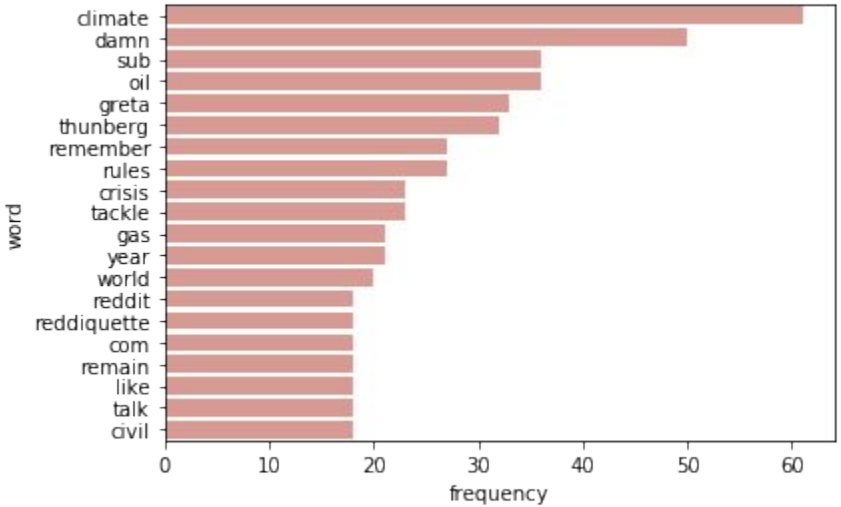
EDA

Model Validation: Logistic Regression & Multinomial Classification

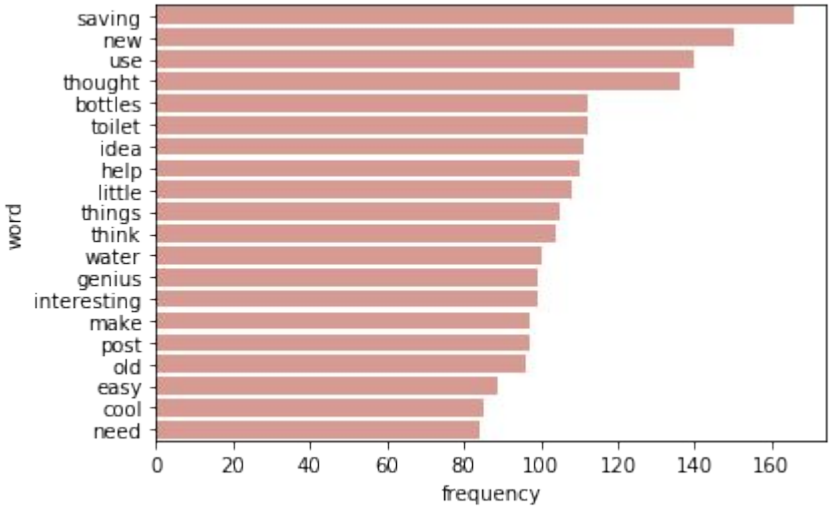
Model Fitting & Prediction

Workflow - EDA

environment



zerowaste



WORKFLOW – MODEL FITTING

Cvec params

Max features = 500

stop_words = 'english'

WORKFLOW
– MODEL EVALUATION

Confusion Matrix			
Naive Bayes		Logistic Regression	
True Negatives	212	True Negatives	214
False Positives	2	False Positives	0
False Negatives	2	False Negatives	0
True Positives	56	True Positives	58

Conclusion

It seems that although they are similar in topics but their word frequency used are much different.

Thus a simple logistic regression works very well