Subreddit Classification

Using machine learning to make classification predictions

What are we doing?

- Can we achieve at least 80% accuracy using machine learning and NLP to predict which subreddit a post originated from?
- Can we make any informed inferences about each subreddit from our data?
- How can we use this tool?

Which Subreddits?

r/Ecology

Definition: The scientific biological study of the relation between organisms to one another and their surroundings. (Oxford Languages - Google)

r/Environmental Science

Definition: An interdisciplinary academic field that draws on ecology, geology, meteorology, biology, chemistry, engineering, and physics to study environmental problems and human impacts on the environment. (britannica.com)

Which Subreddits? cont.

r/Ecology

- Created Nov 13, 2008
- 59.1k members
- 7,594 posts scraped
- 2,889 non-null posts

r/Environmental Science

- Created Dec 30, 2010
- 38.9k members
- 5,099 posts scraped
- 2,970 non-null posts

Original Data Features:

1) Subreddit
2) Post title
3) Body text
4) UTC

Engineered Data Features:

1) Body text word count
2) Body text character count
3) Title word count
4) Title character count

Data Cleaning

1) Drop extraneous column
2) Drop any index with nothing in the body

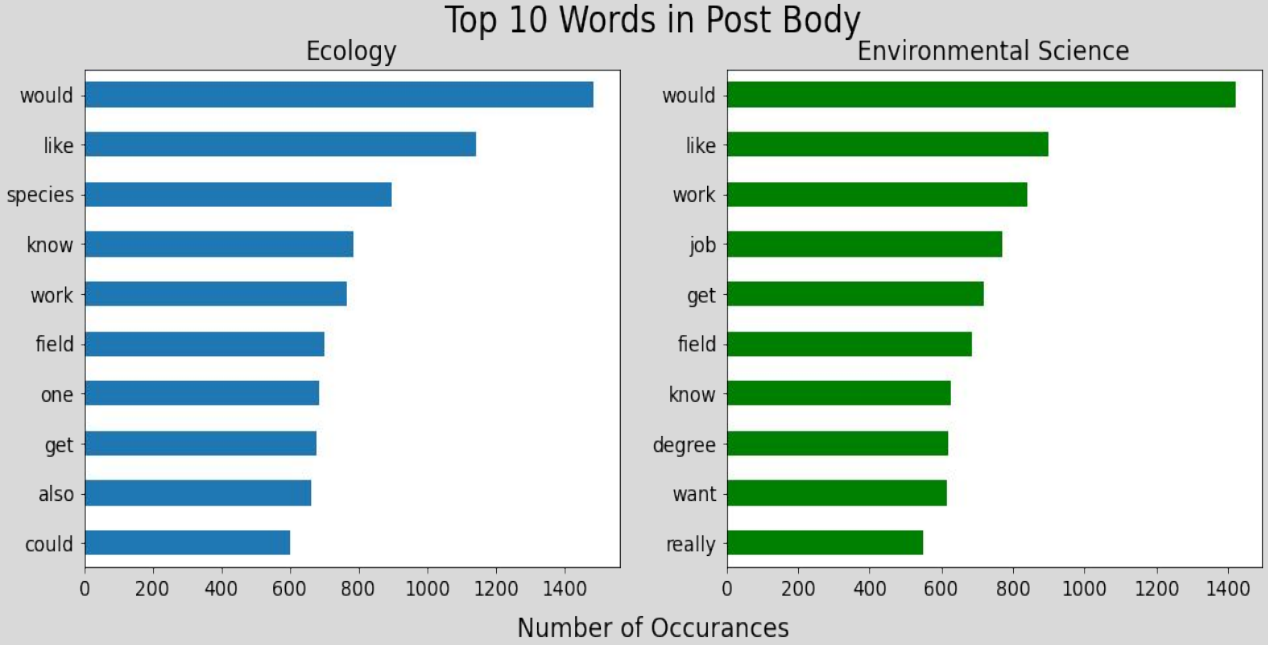
Stop Words

- English
- Additional words: subreddit name, https, www, com.

EDA - Single Words

Words in common:

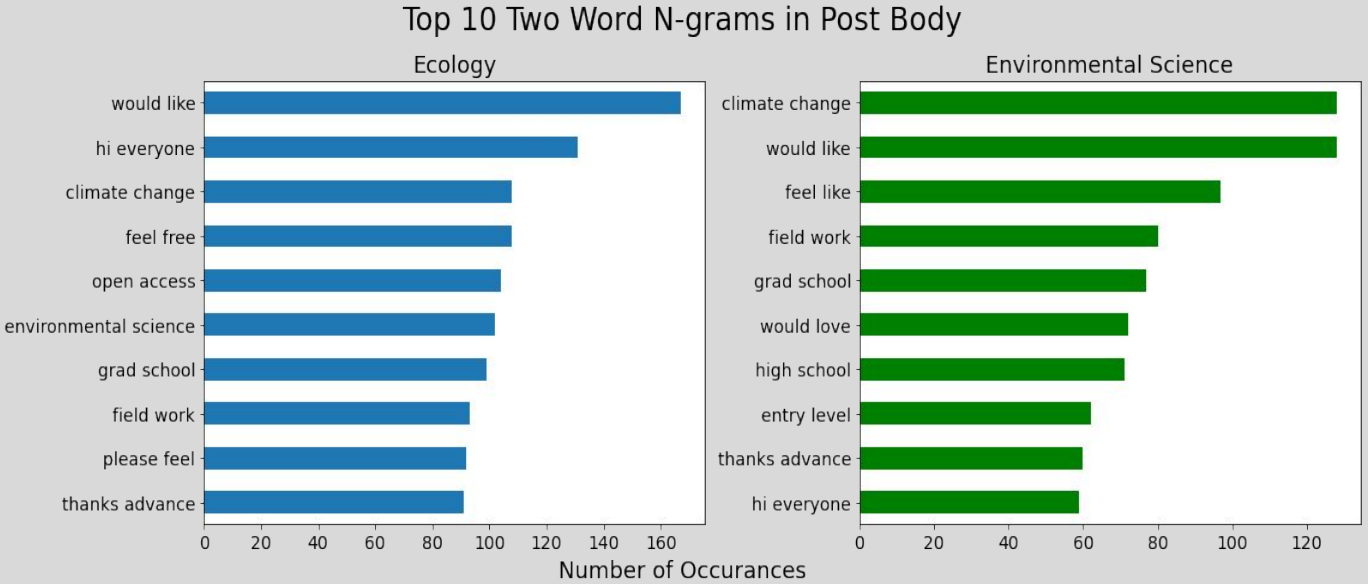
1. Would
2. Like
3. Work
4. Field
5. Get
6. Know



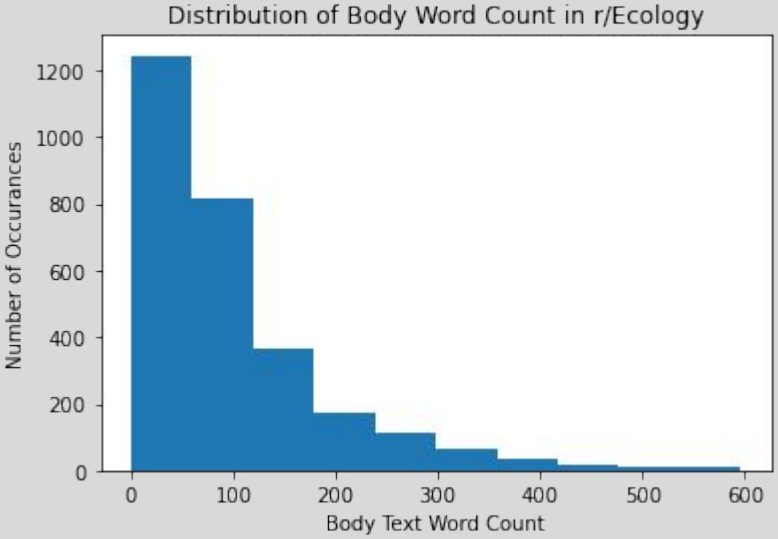
EDA cont. - N-grams

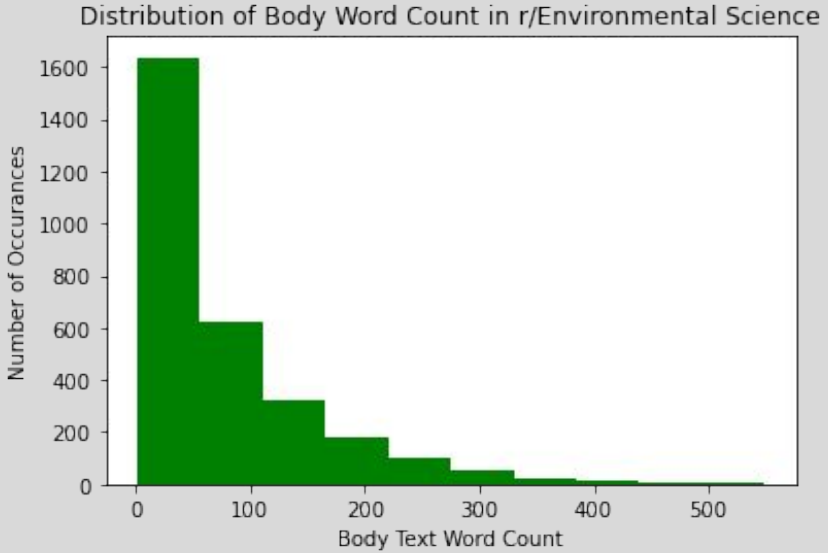
N-grams in common:

1. Would like
2. Grad school
3. Thanks advance
4. Field work
5. Climate change

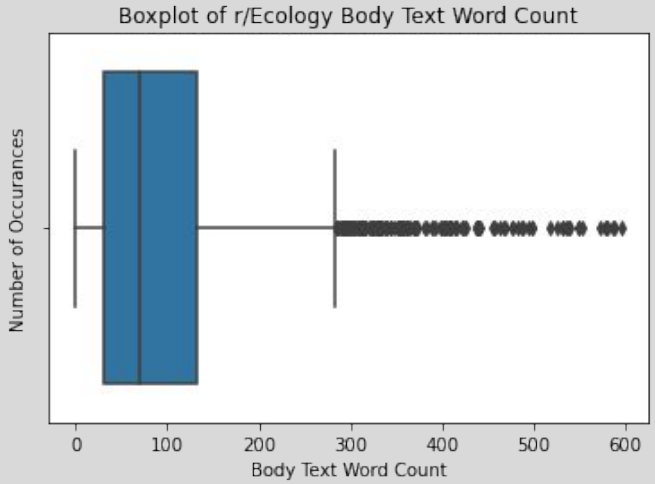


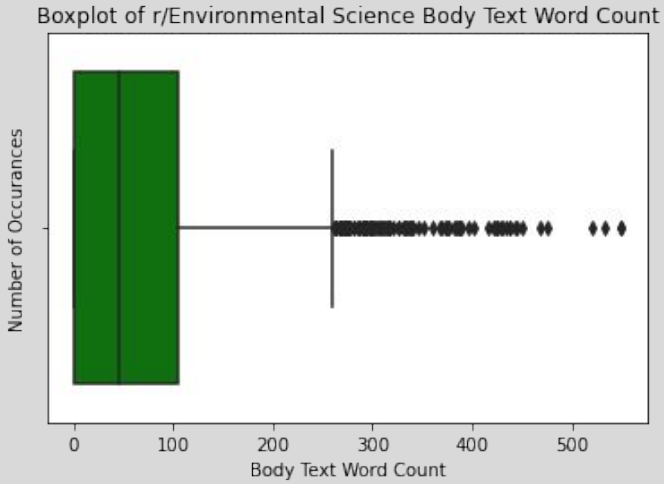
EDA cont. - Word Count Distribution





EDA - Word Count cont.





r/Ecology mean: 96

r/Env Sci mean: 69

Modeling

Models used:

1. Multinomial Naïve Bayes
2. Random Forest Classifier
3. C-Support Vector Classification

NLP used:

1. Count vectorization
2. TFID vectorization

Model Performance

Top Performer: Random Forest with count vectorization

Model Type	NLP	Score
Naïve Bayes	Count vectorizer	77.5
Naïve Bayes	TF-IDF	79.7
Random Forest	Count vectorizer	80.4
Random Forest	TF-IDF	80.1
SVM	Count vectorizer	78.9

Best Params

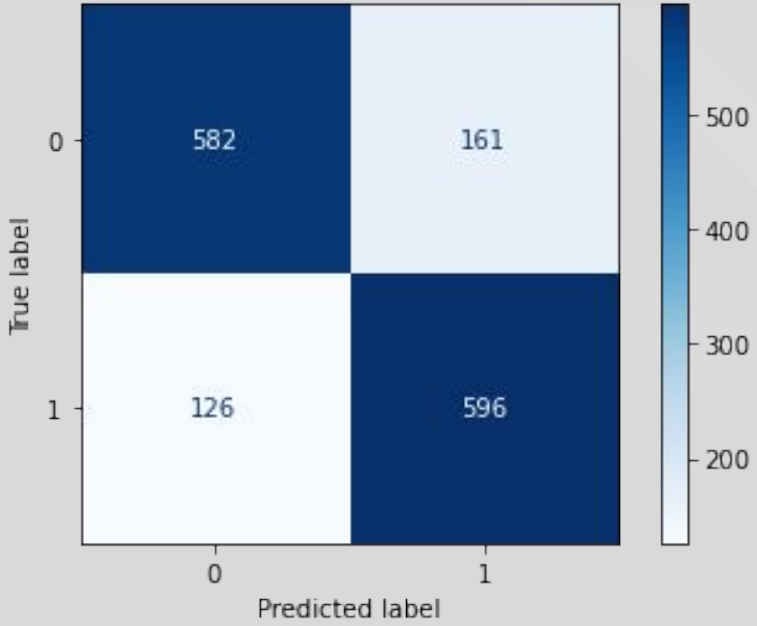
Random forest/Cvec:

Max depth: None
Max features: .5
N-estimators: 350

Scoring

Accuracy: 80.4%

Recall: 82.5%



Conclusion and Inference

Conclusion: Success!

- Surpassed our null hypothesis accuracy of 50.7% with 80.4%

Inferences:

- r/Ecology seems to have more posts pertaining to articles and real world happenings with ecology.
- r/Environmental Science has more posts with questions about job advice and guidance for higher education.

Potential Use:

1) Non-profit fundraising

2) Community involvement and engagement

End

Thank you for your time and attention!