

NLP: Differentiating Subreddits

Tanya Do




The subreddits

r/askscience
r/askhistorians



Sample ambiguous posts...

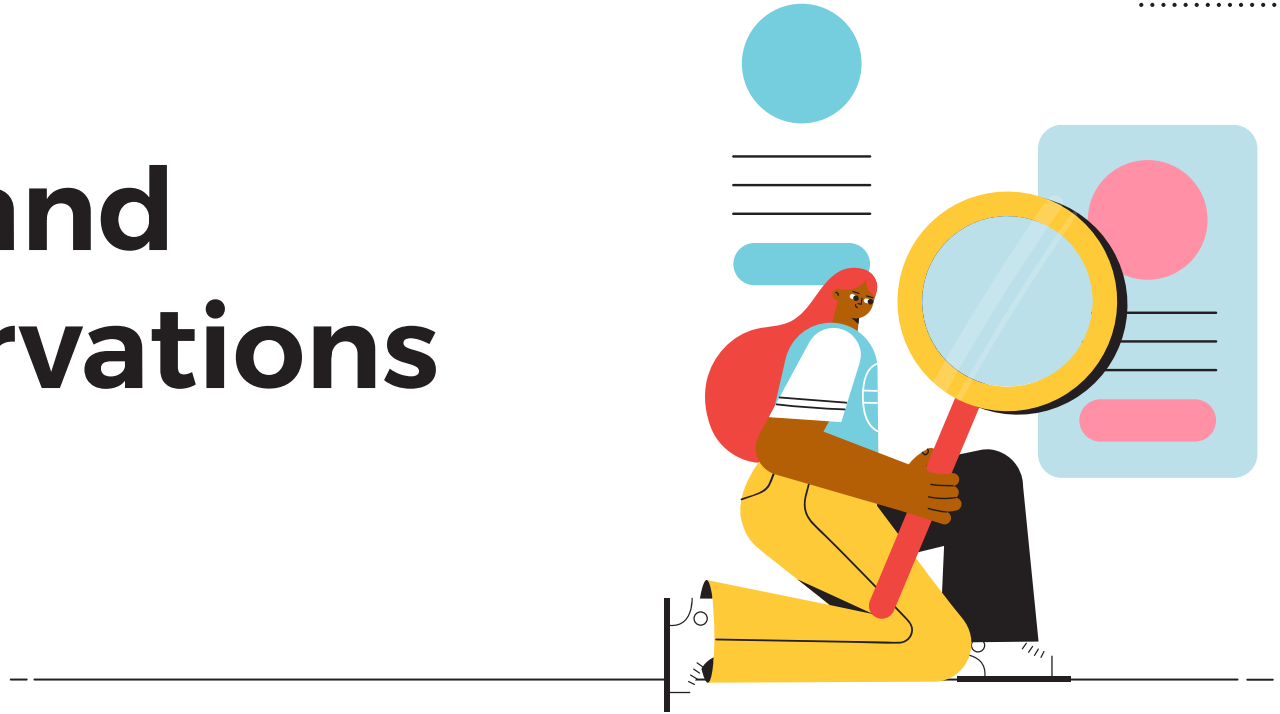


I've been exploring various tooth types, but the information in the books is presented in a confusing manner. Could someone clarify what exactly bunodont teeth are?

I'm a Portuguese into English translator and I often see the word "clínico" in the documents I translate from, but I see it "medical" instead of "clinical" in English documents. Is there a simple way to differentiate between those words?



EDA and Observations



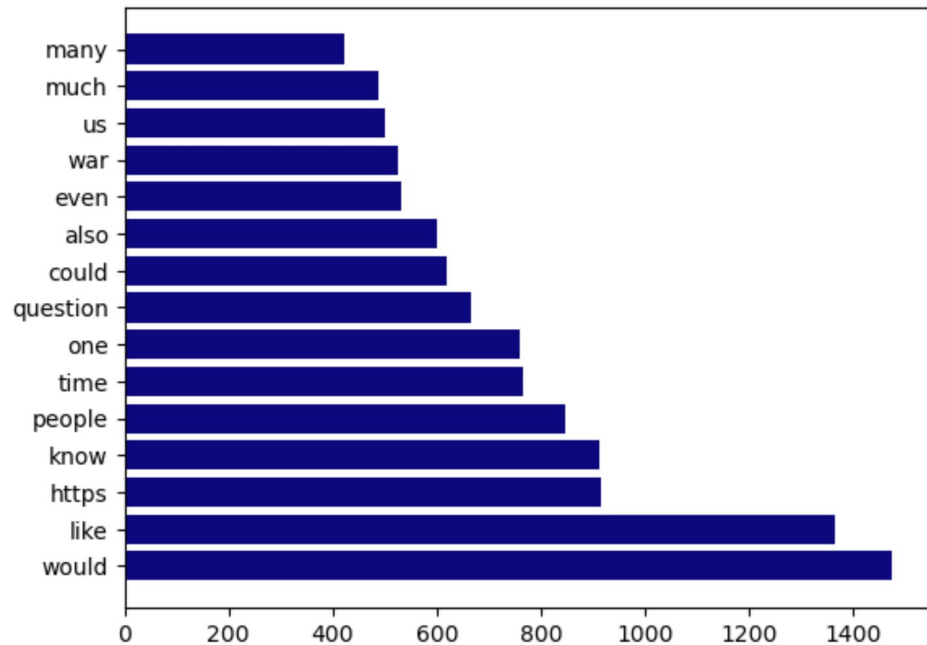
Stemming 'n Lemming

```
cvec.get_feature_names_out
```

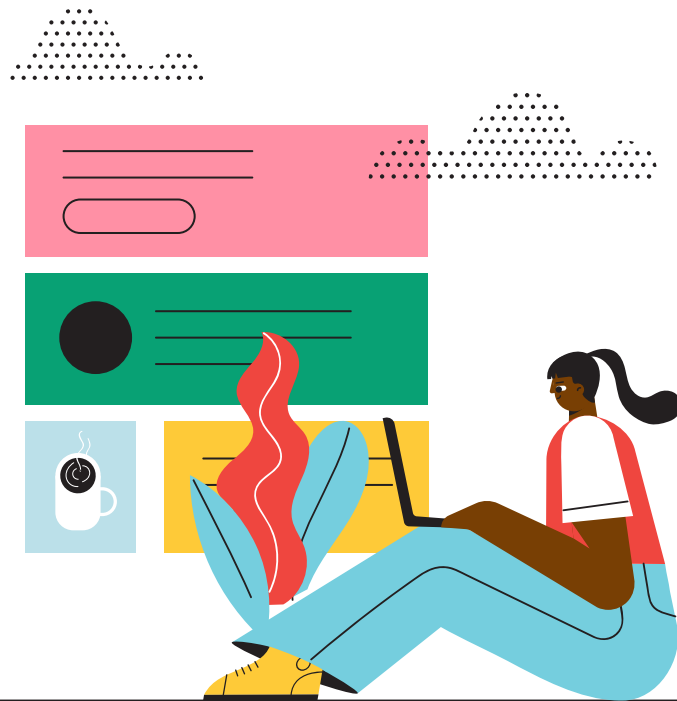
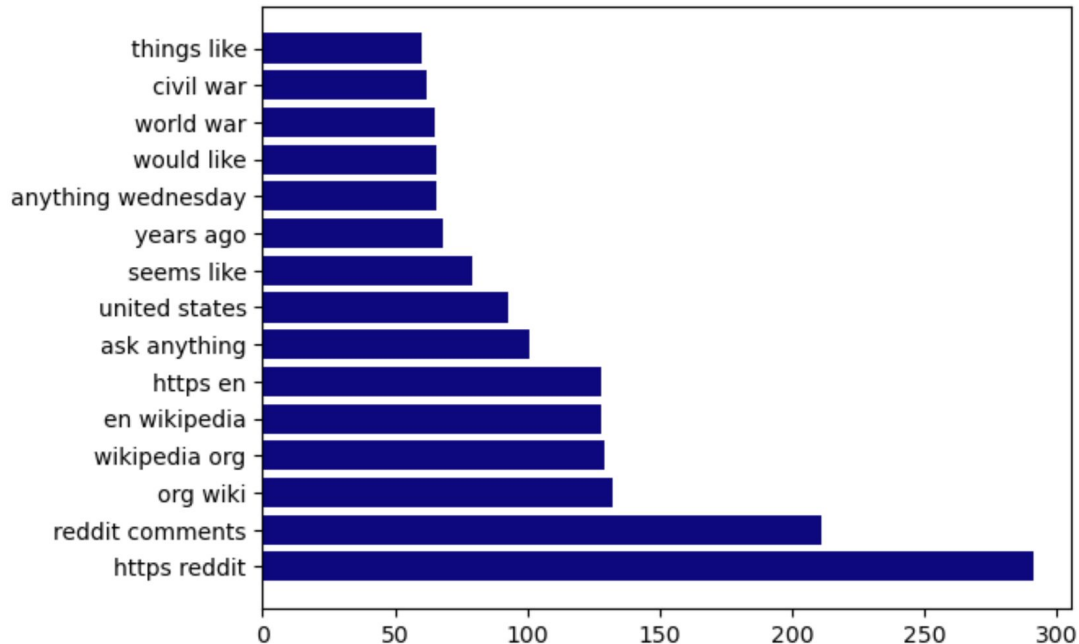


```
<bound method CountVectorizer.get_feature_names_out of CountVectorizer(stop_words=['i', 'me', 'my', 'myself', 'we', 'our', 'ours',  
    'ourselves', 'you', "you're", "you've", "you'll",  
    "you'd", 'your', 'yours', 'yourself', 'yourselves',  
    'he', 'him', 'his', 'himself', 'she', "she's",  
    'her', 'hers', 'herself', 'it', "it's", 'its',  
    'itself', ...])>
```

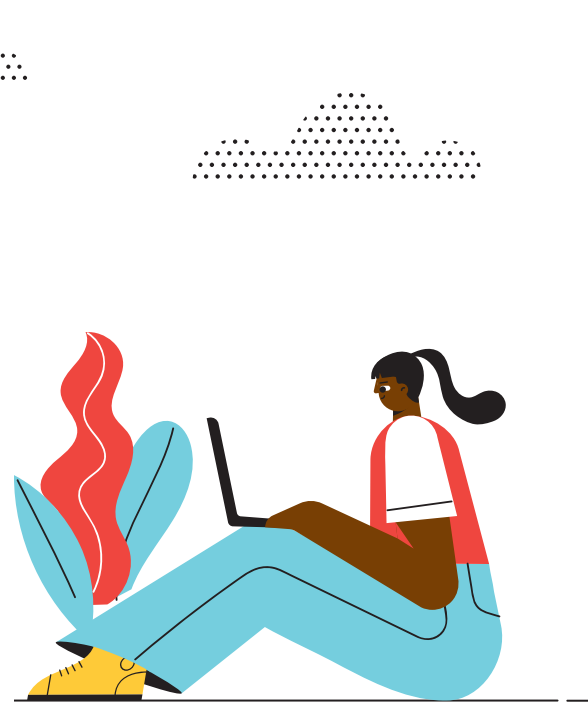
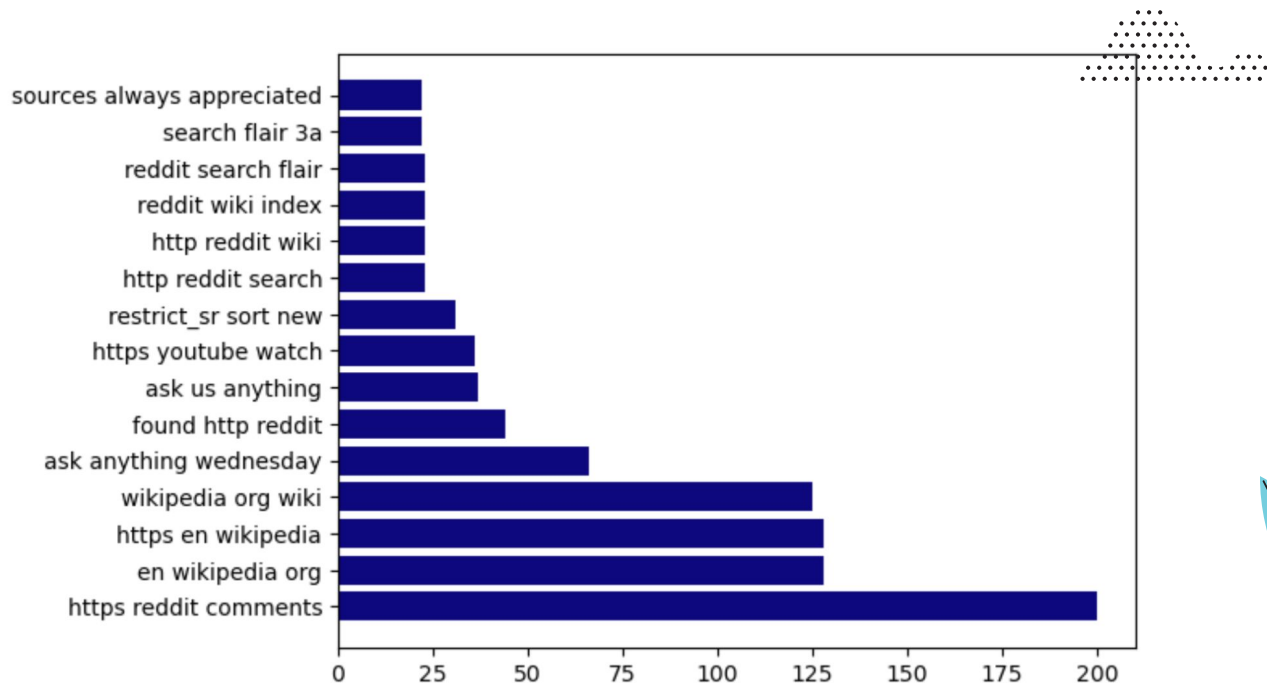
Unigrams



Bi-grams



Tri-grams



Let's Do Some Modeling



The Stop Words

science

history



askscience

askhistorians

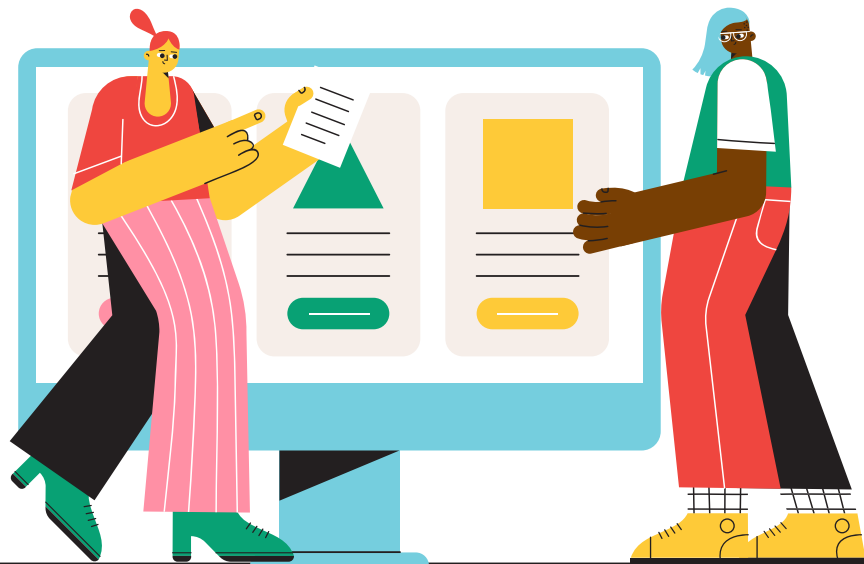


The Data Stats

subreddit
askhistorians: 0.578244
askscience: 0.421756

df.shape
3815, _

Model
Multinomial Naive Bayes



Outcome: Training vs Test Scores

(0.9993009437259699,
0.9475890985324947)



Basic

(0.9905627403005942,
0.9685534591194969)



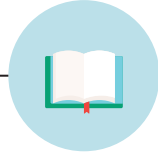
OG Text

(0.9902132121635792,
0.9664570230607966)



Stem

(0.9891646277525341,
0.960167714884696)



Lemmatize

Best Params:

```
{'cvec__binary': False,  
'cvec__max_df': 0.5,  
'cvec__max_features': None,  
'cvec__min_df': 1,  
'cvec__ngram_range': (1, 1)}
```



Thanks!

Tanya Do
Data Person
hey@tanyado.com

