Carter Auer

Results

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\carte\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
Preprocessing: none, Vectorization: bow, Accuracy: 0.875976676146079, F1 Micro: 0.875976676146079, F1 Macro: 0.5995626753240965
Preprocessing: none, Vectorization: tfidf, Accuracy: 0.7490467010613746, F1 Micro: 0.7490467010613746, F1 Macro: 0.5013670216607017
Preprocessing: none, Vectorization: lsa, Accuracy: 0.3929793575449829, F1 Micro: 0.3929793575449829, F1 Macro: 0.1383184835070326
Preprocessing: porter, Vectorization: bow, Accuracy: 0.872256554154896, F1 Micro: 0.872256554154896, F1 Macro: 0.5953111946047801
Preprocessing: porter, Vectorization: tfidf, Accuracy: 0.7388243786695939, F1 Micro: 0.7388243786695939, F1 Macro: 0.4917134104624368
Preprocessing: porter, Vectorization: lsa, Accuracy: 0.3929793575449829, F1 Micro: 0.3929793575449829, F1 Macro: 0.1383184835070326
Preprocessing: snowball, Vectorization: bow, Accuracy: 0.8726429017341262, F1 Micro: 0.8726429017341262, F1 Macro: 0.5957375987673805
Preprocessing: snowball, Vectorization: tfidf, Accuracy: 0.7382834630542384, F1 Micro: 0.7382834630542384, F1 Macro: 0.4909793497751417
Preprocessing: snowball, Vectorization: lsa, Accuracy: 0.3929793575449829, F1 Micro: 0.3929793575449829, F1 Macro: 0.1383184835070326
Preprocessing: lemmatizer, Vectorization: bow, Accuracy: 0.876440332726182, F1 Micro: 0.876440332726182, F1 Macro: 0.6004502336386367
Preprocessing: lemmatizer, Vectorization: tfidf, Accuracy: 0.7498525599861032, F1 Micro: 0.7498525599861032, F1 Macro: 0.5014759087861245
Preprocessing: lemmatizer, Vectorization: lsa, Accuracy: 0.3929793575449829, F1 Micro: 0.3929793575449829, F1 Macro: 0.1383184835070326
Top Five Categories and their Counts:
category
POLITICS          35600
WELLNESS          17931
ENTERTAINMENT     17361
TRAVEL             9883
STYLE & BEAUTY     9813
Name: count, dtype: int64


Is dataset balanced or imbalanced? =
Imbalanced
```

1.Look at the category data

a.What are the five categories we are classifying?

#1 POLITICS
#2 WELLNESS
#3 ENTERTAINMENT
#4 TRAVEL
#5 STYLE & BEAUTY

b.How many instances of each class are there?

POLITICS 35600
WELLNESS 17931
ENTERTAINMENT 17361
TRAVEL 9883
STYLE & BEAUTY 9813

c.What type of dataset are we dealing with based on these numbers?

It seems we are dealing with articles that are heavily leaning towards political topics.

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\carte\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
Preprocessing: none, Vectorization: bow, Accuracy: 0.875976676146079, F1 Micro: 0.875976676146079, F1 Macro: 0.5995626753240965
Preprocessing: none, Vectorization: tfidf, Accuracy: 0.7490467010613746, F1 Micro: 0.7490467010613746, F1 Macro: 0.5013670216607017
Preprocessing: none, Vectorization: lsa, Accuracy: 0.3929793575449829, F1 Micro: 0.3929793575449829, F1 Macro: 0.1383184835070326
Preprocessing: porter, Vectorization: bow, Accuracy: 0.872256554154896, F1 Micro: 0.872256554154896, F1 Macro: 0.5953111946047801
Preprocessing: porter, Vectorization: tfidf, Accuracy: 0.7388243786695939, F1 Micro: 0.7388243786695939, F1 Macro: 0.4917134104624368
Preprocessing: porter, Vectorization: lsa, Accuracy: 0.3929793575449829, F1 Micro: 0.3929793575449829, F1 Macro: 0.1383184835070326
Preprocessing: snowball, Vectorization: bow, Accuracy: 0.8726429017341262, F1 Micro: 0.8726429017341262, F1 Macro: 0.5957375987673805
Preprocessing: snowball, Vectorization: tfidf, Accuracy: 0.7382834630542384, F1 Micro: 0.7382834630542384, F1 Macro: 0.4909793497751417
Preprocessing: snowball, Vectorization: lsa, Accuracy: 0.3929793575449829, F1 Micro: 0.3929793575449829, F1 Macro: 0.1383184835070326
Preprocessing: lemmatizer, Vectorization: bow, Accuracy: 0.876440332726182, F1 Micro: 0.876440332726182, F1 Macro: 0.6004502336386367
Preprocessing: lemmatizer, Vectorization: tfidf, Accuracy: 0.7498525599861032, F1 Micro: 0.7498525599861032, F1 Macro: 0.5014759087861245
Preprocessing: lemmatizer, Vectorization: lsa, Accuracy: 0.3929793575449829, F1 Micro: 0.3929793575449829, F1 Macro: 0.1383184835070326
```

2.Look at the results for vectorization

a.Which of the 3 vectorization methods performs “best”?
bow/ Bag of words performed the best on average by a lot compared to tfidf and lsa

b.Why do you think this method performs so well in conjunction with Naïve Bayes? Think about the assumptions being made by the Naïve Bayes model as well as the distribution of classes.

BOW way of representing words coincides well with Naïve Bayes. The way BOW manages each word as a separate independent feature in order to represent each as a probability works well when combined with Naïve Bayes probability based modeling.

c. Explain why it makes sense that the other two vectorization methods have worse performance. Hint: for each of the vectorization models we discussed potential pitfalls in class, and we still want to bear the Naïve Bayes assumptions in mind.

tfidf focuses more on word importances which isn't terrible for Naïve Bayes but still doesn't play well with its model building. As for lsa when intaking words does something called “dimensionality reduction” which in this case can lead to a loss in frequency data which is critical for Naïve Bayes to train properly.

3.Look at the results for stemming and lemmatization

a. Is there an improvement when applying stemming and lemmatization as compared to not preprocessing the data? Which method had the highest improvement?

Yes it is important to apply stemming and lemmatization instead of not preprocessing at all. From my results I saw a marginal increase in F1 scores when using preprocessing.

b. Explain why you think this is the case.
I believe since the preprocessing is allowing for words to be more accurately categorized will cause for more accurate predictions since some words will now be misrepresentation when training the model.