MVP for Natural Language Processing project

The goal is to compare data from New York Times articles and Onion articles to be able to predict which is which based on either title or content.

I pulled data from Kaggle for from both the New York Times dataset and the Onion datasetmushroom and imported it into a Jupyter Notebook. I did some basic exploratory data analysis on it and added a target column using 1 for the Onion and 0 for NYT.

So far, I have been mostly engaged in pre-processing the data.

1) Removed punctuation
2) Tokenized the data
3) Removed Stop Words
4) Vectorized the data

My document-term matrix is about 84,000 columns right now so I definitely need to do dimensionality reduction and probably need to go back and do additional pre-processing and test the different stemming options.

After completing dimensionality reduction, the plan is to work through topic modeling on each data set and then moving to build a predictive classification model.