Info284 Fev004

First obligatory assignment

The text in the test.csv file is organized as: id, title, author, text. I will be using the Pandas module to convert the csv files to a data-structure called DataFrame. This will allow me to conveniently access each column separately.

First task is to remove any non English articles from the dataset. For this I will use the list of stopwords from the natural language toolkit.
I have decided to not use the metadata, not sure if that means that I am supposed to remove it from the dataset. In this we have two options either remove metadata, stopwords… and spit out a new processed csv file or keep the original dataset as is and pre-process it each time the python script is executed. I think it is at least save to remove non-english articles, stopwords and rows where the text field is empty. I am not going to use the metadata but I will still keep it in the dataset until I get more information.

Non-english articles, stopwords, empty fields, metadata in text field and everything else other than words have now been removed so that a vocabulary can be made.

I stopped using pandas and decided to process the files without it, this turned out much better than than my first attempt. Clean.py is the old code no longer in use(but I will deliver it also), clean2.py contains code that actually works and is without the use of the pandas module. There is now a file, Vocabulary.txt, which contains the vocabulary. I have not yet counted how often each word appears in the dataset. It is unclear to me if I should count how many times x word appears in each of the articles, or how many times the word appears in all of the dataset.