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 try to use a regex. 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. For removing stopwords I will use the natural language toolkit.

The dataset has not been cleaned properly, but I will turn in what I have so far. I discovered today that not all of the non english language articles have been properly cleaned out. Dont really know why it works for some articles and some not, will have to look into it closer. There is a vocabulary text file but its definitely not ready yet. All code so far is in “clean.py”.