Project Proposal

Inho Yong and Yasaswy Kota

For the project, our team is going to parse the data using python. The parsing will remove certain stop words, like ‘the’, ‘that’, ‘a’, etc. which don’t serve much use for text classification purposes. We will then end up with a large feature matrix which will be thousands of dimensions long. We will reduce this multi dimensional data structure using PCA. We aim to keep around 99% of the variance of the original data after dimensionality reduction. This reduced data will be processed using TF-IDF to “rank” the importance of each word.

We will take the first 50 features of each article to form the data to be used for our classifier that is written using MATLAB. We will perform KNN classifier on this data to perform the final classification.