Document Classification

Project 1: Data Science Practicum

Instructor: Dr. Shannon Quinn

Team: Volans

Members:

1. Ankita Joshi
2. Prajay Shetty
3. Vyom Shrivastava

Our Approach

1. Preprocessing
2. Combining training files
3. TF-IDF thresholding for stopwords
4. Naive Bayes Implementation (with laplacian smoothing)
5. Ignored new words during prediction
6. Started with 3 branches

What worked. What didn’t.

Tf-IDf thresholding helped.

Ignoring new words helped.

The basic naive bayes implementation alone, gave us a 94 something accuracy.

Class based implementation of Naive Bayes - just call .train() or .test() for a given dataset.

Combining training files did not.

Ngram implementation (did not use)

Tried to implement a binary naive bayes classifier, one for each class, couldn’t finish the implementation in time.