CMPS 470 PAs Project Report

Team: ChatBLT

Team Member: Zachary Gros, Rhett Hill

To generate features for our dataset, we chose to use the scikit-learn feature extraction library’s text tools to detect which words were the most frequent and most important in our dataset. We specifically used the TfidfVecorizer (Term Frequency-Inverse Document Frequency) to determine the frequency of every word in the dataset with each word being assigned a TF-IDF score by the vectorizer. This helps us to better detect which emails are spam or not by comparing the TF-IDF scores of inputted emails to those in our dataset. We can visualize the data features using Principal Component Analysis as shown below. Looking at the plot below, we can see that the emails labeled ‘ham’ are much more spread out than those labeled ‘spam’ which are all clustered in one concentrated area.

