Sam Miller

CS 584

HW 3

My experimental setup for this homework involved using Python with sklearn to process the news group data and evaluate the performance of a Naïve Bayes classifier on differentiating between documents labelled as ‘motorcycle’ and ‘auto.’ The first step involved fetching the already vectorized news group data set from sklearn using the sklaern.datasets.fetch\_20newsgroups\_vectorized function. This data also already comes partiioned into training and testing datasets. From here, I used numpy to identify which documents were labelled with categories 7 and 8, which correspond to auto and motorcycle documents respectively, and filtered out everything that did not fit into these labels. Finally, using sklearn.naive\_bayes.MultinomialNB class, I fit a naïve bayes classifer to the training set and evaluated its performance over the testing set using F1 score as the main metric.