README

1. I used a list comprehension to filter out punctuation and numbers as they do not really carry any semantic meaning and only take space. Removing punctuation does have its flaws though, since words that include a punctuation mark like “can’t” or “U.S.” will not appear in the data. I also lowercased the data since I did not really see a difference in differentiating between lowercase and uppercase words. Finally, I filtered out words that appear less than 50 times in the corpus.
2. I used the Truncated SVC class to perform dimensionality reduction on the table from part 1.
3. I used the same models I used for assignment 1, Gaussian Naïve Bayes and Quadratic Discriminant Analysis.