CS168 Project 2

Henry Lin, Kaylee Bement

1. A. See code.py

B. See code.py

C. Based on the heatmaps, the Jaccard similarity metric seems to be the most reasonable for this dataset. Since any bag of words model is bound to be sparse by design, it is expected that Jaccard would perform well in comparing the articles of each newsgroup.

The atheism newsgroup seems to share a noticeable similarity with the religion Christian newsgroup. While the two groups may have differing opinions, it is expected that they are rated similar since they discuss the same topics. There is a surprising amount of similarity between science-related and religion-related newsgroups such as religion Christian and medicine, but the two groups definitely have an intersection where there is a significant level of debate, both in the dataset and real life.

1. A. See code.py [answer question as well – avg classification error]

B. 1(b) plots the average similarity between newsgroups. For any given newsgroups A and B, the similarity between A and B is the same as the similarity between B and A, so the plot will be symmetric. However, 2(a) plots the number of articles in one newsgroup that have their nearest neighbor in another newsgroup. For any given newsgroups A and B, the number of articles in A that have a nearest neighbor in B does not correlate with the number of articles in B that have a nearest neighbor in A.

C. Blah

D. Blah