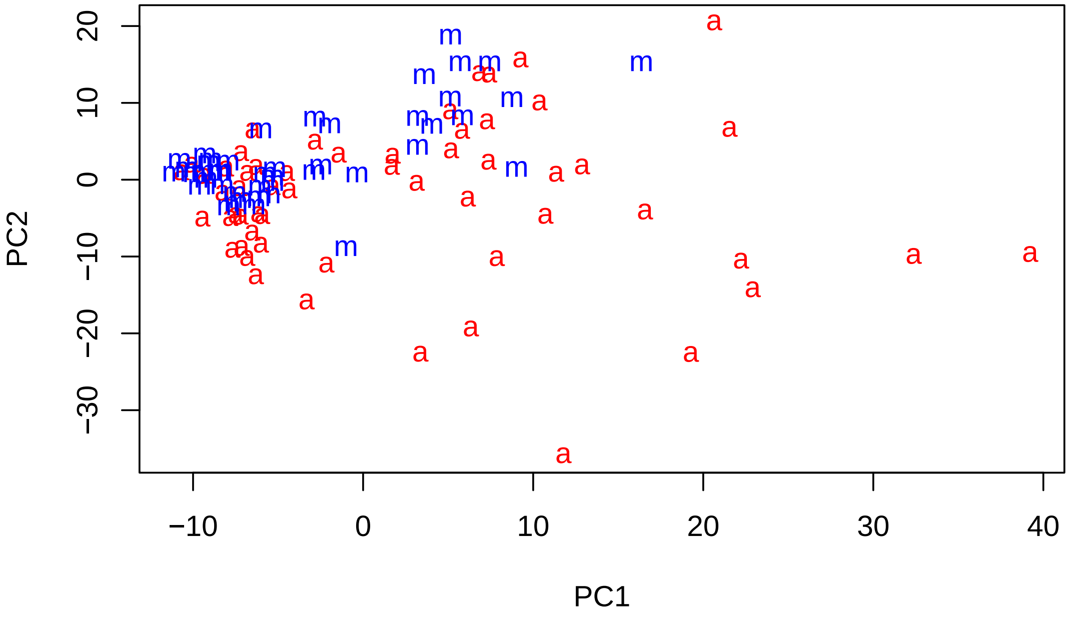
STAT 6390 Mini Project 3 | Brett Walker - 3/28/2023

Question 1.

a. A subject would be the "classification" of a story (e.g. "art" or "music"). Its feature x~ would be the vector of word counts for the selected dictionary. The data is a bag of words vectorization of each article. If I were to construct this dataset, I would use a token bag for times and dates, as the excerpt suggests music articles more often include "words" like those.

b. 5 PCs makes sense. One would expect there to be a few clusters of words which differentiate between art and music stories.

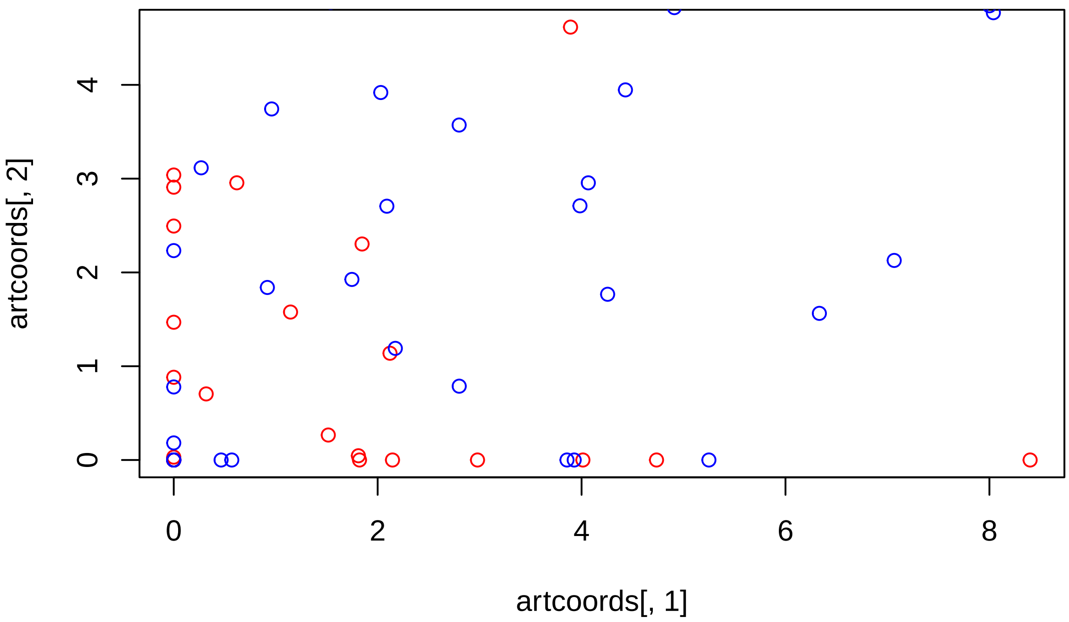


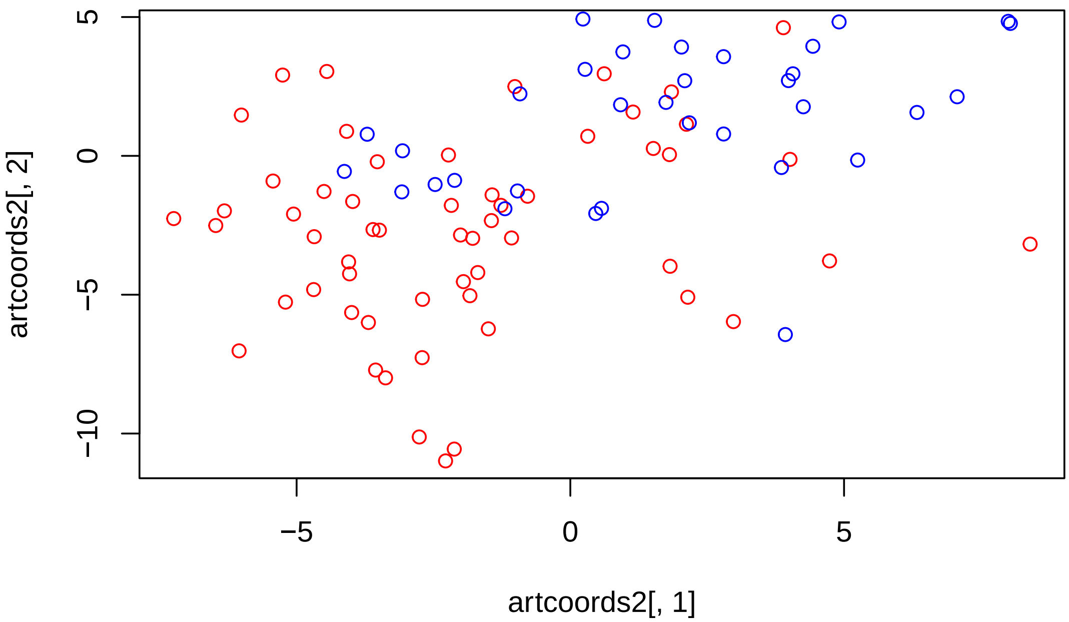
c. In the excerpt, using 2 PCs is shown to be a fairly good predictor–the classes appear almost linearly separable when plotted with the first two PCs. However, with the centered data, this separation becomes less obvious (though there is still a good degree of separation).

d. (see code for implementation of autoencoder in keras)
Autoencoder weights (top 30):
0.12, 0.12, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.1, 0.1, 0.1, 0.099, 0.099, 0.099, 0.098, 0.097, 0.097, 0.096, 0.096, 0.095, 0.095, 0.094, 0.093, 0.093, 0.092, 0.092, 0.092, 0.091
PCA "weights" (top 30):
0.07, 0.065, 0.063, 0.063, 0.063, 0.062, 0.061, 0.06, 0.059, 0.059, 0.058, 0.057, 0.054, 0.054, 0.054, 0.054, 0.053, 0.052, 0.051, 0.05, 0.05, 0.05, 0.049, 0.049, 0.049, 0.048, 0.048, 0.047, 0.046, 0.046

e. Scores are similar, but different between the two methods (approx double from pca to autoencoder). This is likely because the weights on the encoder and decoder are not tied (as they should be if the solution is to be equivalent to PCA),

f & g. (again see code for implementation) Art articles seem to be more crowded toward the axes, with music articles being less clustered, but tending toward the second axis. Oddly enough, without pmax correction on the calculation of the two sets of coordinates, we get good clustering for each group. (resultant graph shown below the first graph). Using this matrix gives ostensibly better results than the matrix generated by PCA.


artcoords[, 2]
artcoords[, 1]


artcoords2[, 2]
artcoords2[, 1]