CSI 5810 (Assignment # 2)

1. The folder "CSI5810TextFiles" posted on Moodle contains 8 text files. You are to apply text-processing steps including stop word filtering to obtain term-document matrix under Boolean Model. Using this matrix, calculate similarity between all document pairs and show your results in the form of an 8x8 matrix. Use Jaccard's similarity measure.

2. This is a continuation of Exercise #1. In this case, determine the vector space representation for each document and calculate the 8x8 document similarity matrix using Cosine measure of similarity.

3. In this exercise, you will use "Wheat Data" posted at Moodle. The data consists of 32 training examples each from three classes. Using these training examples, you will perform classification of 3 test examples by k-NN classification (k= 1, 3, and 5), and by Naïve Bayes classifier. Compare and comment on your results.

4. In this exercise, you will again use 32 training examples of wheat data and project them into two-dimensions using the Fisher's LDA method for multiple classes. Next, you will apply PCA on the same 32 examples to reduce the data to two dimensions. You will show your result by creating two scatter plots, one for LDA and the other for PCA. Make sure to color code the project points with their respective class labels.