CSI 5810 (Assignment # 2)

1. (a) An information retrieval system has a certain pair of average precision and recall values when the system returns 10 documents in response queries. Would the precision and recall rate remain unchanged if the system were modified to return 20 documents in response to queries?
   (b) A system retrieves 20 documents with precision being 80% and recall being 50% for a given query. What is the number of relevant documents *that are not retrieved by the system*?

2. 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 a 8x8 matrix.

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

4. Consider the following set of seven two-dimensional records:
   (1 0)' (0 1)' (0 –1)' (0 0)' (0 2)' (0 –2)' (-2 0)'

   The first three records are examples of class 1 and the other four are from class 2. (i) Sketch the decision boundary due to 1-NN rule. (ii) Find the sample means for two classes and sketch the minimum distance decision boundary.

5. In this exercise, you will use "Wheat Data" posted at Moodle. The data consists of 25 training examples each from two classes. Using these training examples, you will perform classification of 4 test examples by 1-NN classification and by Naïve Bayes classfier.