CSI 5810

Assignment #4

1. In this exercise, you will perform k-means clustering on the seed data at the following link:
https://archive.ics.uci.edu/ml/datasets/seeds
You will perform clustering using the following values of k: 2,3, 4, and 5. In each case you will determine the SSE value and calculate the value of Rand index and tabulate your results.

2. In this exercise, you will build a linear predictive model to predict crime rate based on a number of factors. The data is in the "crime-rate" file. You will build the model by writing your own script for gradient search. Experiment with 2-3 learning rates to see the effect of learning rate on the search.

3. A transaction database is given below. Using the A-priori algorithm, determine all frequent item-sets with minimum support of 30%. Show results at each step of the algorithm.

TID#	Items Bought
1	A, B, D, E
2	B, C, D
3	A, B, D, E
4	A, C, D, E
5	B, C, D, E
6	B, D, E
7	C, D
8	A, B, C
9	A, D, E
10	B, D

4. Consider the following simple IR situation. We have five keywords and six documents. The term-document matrix is given by the following matrix F.

	D1	D2	D3	D4	D5	D6
K1	1	0	1	0	0	0
K2	0	1	0	0	0	0
K3	1	1	0	0	0	0
K4	1	0	0	1	1	0
K5	0	0	0	1	0	1

(i) Obtain the singular value decomposition of F.
(ii) Reconstruct F using only the top two singular values.
(iii) Show the representation of the documents and the keywords in the 2-D space after SVD application.
(iv) Using the cosine similarity measure in the LSI space, calculate the document similarity matrix.