CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 4: MapReduce-Data Mining

1. Find the set of 2-shingles for the "document":

ABRACADABRA

and also for the "document":

BRICABRAC

Answer the following questions:

i) How many 2-shingles does ABRACADABRA have?
ii) How many 2-shingles does BRICABRAC have?
iii) How many 2-shingles do they have in common?
iv) What is the Jaccard similarity between the two documents?

2. Here is a matrix representing the signatures of seven columns, C1 through C7.

C1	C2	C3	C4	C5	C6	C7
1	2	1	1	2	5	4
2	3	4	2	3	2	2
3	1	2	3	1	3	2
4	1	3	1	2	4	4
5	2	5	1	1	5	1
6	1	6	4	1	1	4

Suppose we use locality-sensitive hashing with three bands of two rows each. Assume there are enough buckets available that the hash function for each band can be the identity function (i.e., columns hash to the same bucket if and only if they are identical in the band). Find all the candidate pairs.

3. Consider the following matrix:

	C1	C2	C3	C4
R1	0	1	1	0
R2	1	0	1	1
R3	0	1	0	1
R4	0	0	1	0
R5	1	0	1	0
R6	0	1	0	0

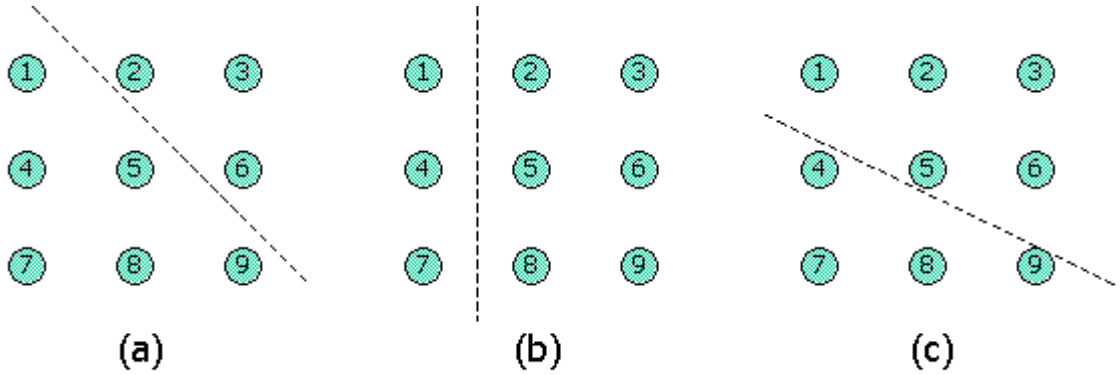
Perform a min hashing of the data, with the order of rows: R4, R6, R1, R3, R5, R2. Compute the signature values of the four columns.

4. We can cluster in one dimension as well as in many dimensions. In this problem, we are going to cluster numbers on the real line. The particular numbers (data points) are 1, 4, 9, 16, 25, 36, 49, 64, 81, and 100, i.e., the squares of 1 through 10. We shall use a k-means algorithm, with **two** clusters. You can verify easily that no matter which two points we choose as the initial centroids, some prefix of the sequence of squares will go into the cluster of the smaller and the remaining suffix goes into the other cluster. As a result, there are only nine different clusterings that can be achieved, ranging from {1}{4,9,...,100} through {1,4,...,81}{100}.

We then go through a reclustering phase, where the centroids of the two clusters are recalculated and all points are reassigned to the nearer of the two new centroids. For each of the nine possible clusterings, calculate how many points are reclassified during the re-clustering phase. List five pairs of initial centroids that results in exactly one point being reclassified.

5. The Bisecting k-Means algorithm starts by dividing the points into two clusters. It may consider several bisections and pick the best one. Let us take "best" to mean the lowest SSE (Sum Squared Error). The SSE is defined to be the sum of the squares of the distances between each of the points of the cluster and the centroid of the cluster.

Suppose that the data set consists of nine points arranged in a square grid, as suggested by the figure below:



Although it doesn't matter for this question, you may take the grid spacing to be 1 (i.e., the squares are 2-by-2) and the lower-left corner to be the point (0,0). We see in the figure three possible bisections. (a) would be the bisection if we chose the two initial centroids to be 3 and 7, for example, and broke ties in favor of 7. (b) would be the split if we chose initial centroids 1 and 2. (c) would be the split for initial choice 2 and 7. Rank these three options from the best to the worse choice.