2IMW30 - FOUNDATION OF DATA MINING
Assignment 1b Report

Duy Pham (0980384, d.pham.duy@student.tue.nl)
Mazen Aly(0978251, m.aly@student.tue.nl)

(Both students contributed equally in this assignment)

EXERCISE 1

In this assignment, we implement the k-means algorithm for clustering the input data points. Clustering is considered as a unsupervised machine learning problem, that means there no labels in the data that we need to predict as classification, but it helps us getting some insights about the input data.

We developed four methods for initializing the first K cluster centroids (where K is an input) The first one is just choosing the first K data points to be the centroids, the second method is randomized selections of the initial centroids. the third method is k-means++ and the fourth method is using Gonzales algorithm.

For each method, we run k-means for different values of k 3,4,5 and in randomized methods, we run that 5 times. we visualise the results to get more understanding and learning experience as shown below.

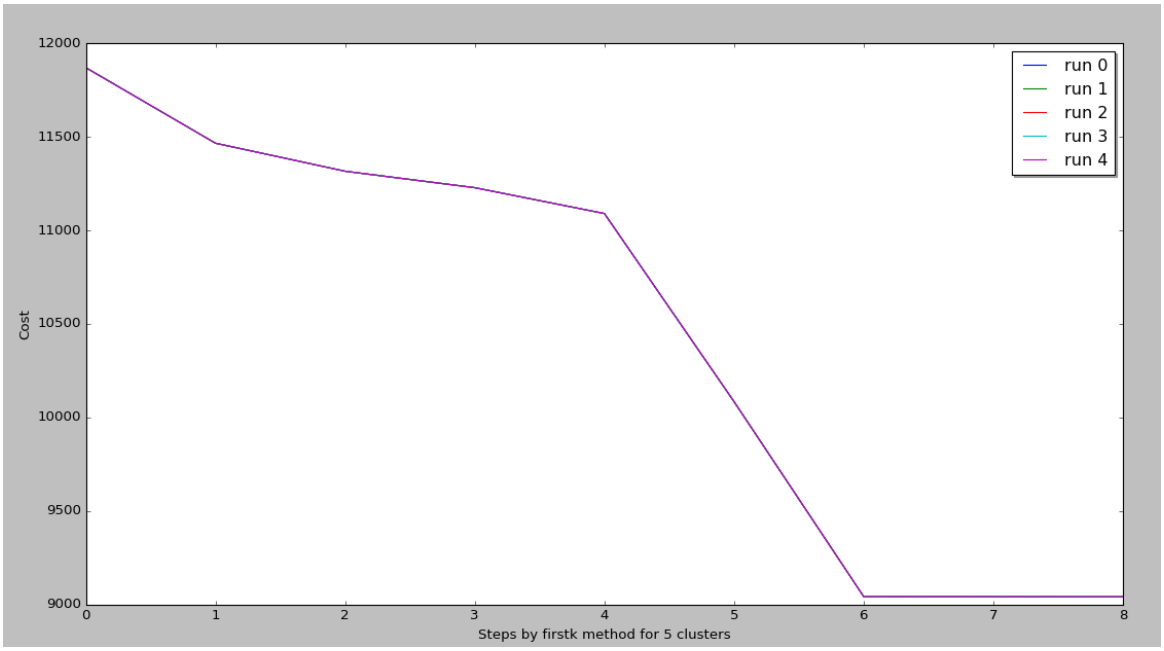

Figure 1

EXERCISE 2

In this excercise, we are proving that if the Jaccard similarity of two sets is 0, then minhashing always gives a correct estimate of the Jaccard similarity.

If the Jaccard similarity of two sets is 0, then the intersection of the two set is empty. This means there are no rows which the elements are both "1" in the characteristic matrix. Hence, when performing minhashing, there cannot be any row with the same values in 2 cells. Thus $P(h(S_i) = h(S_j)) = 0$, which is the same as the Jaccard similarity.

EXERCISE 3

In this exercise, we show an example that the cyclic permutations are not enough to estimate the Jaccard similarity correctly.

Let's take a look at the example in table 1. We will perform a cyclic permutation to estimate the Jaccard similarity.

In this example, we can see that the 2 sets have 2 common words, "b" and "e". In general, $S1 \cap S2 = \{b, e\}$ and $S1 \cup S2 = \{a, b, d, e\}$. Thus the (correct) Jaccard similarity is $|S1 \cap S2|/|S1 \cup S2| = 1/2$.

Now we perform the cyclic permutation.

- a-b-c-d-e: The signature is 1 - 2.
- b-c-d-e-a: The signature is 1 - 1.
- c-d-e-a-b: The signature is 3 - 2.

Table 1: Example

	S1	S2
a	1	0
b	1	1
c	0	0
d	0	1
e	1	1

- d-e-a-b-c: The signature is 2 - 1.
- e-a-b-c-d: The signature is 1 - 1.

We have 2 rows which contain the same values on 2 cells, and 5 rows in total. So the estimated Jaccard similarity is $2/5$, which is not correct. The reason is the $0-0$ row, which makes the next rows to be counted twice in the intersection (while every row is counted only once in the total number of permutations).

EXERCISE 4

In this exercise, we prove the triangle inequality property of Jaccard distance by using the minhash property. That is, the Jaccard distance equals the probability that two sets do not minhash to the same value.

Let's A, B, C be our 3 sets. After constructing the signature matrix, it is inferred from the lecture that:

$$d_j(A,B) = P(minhash(A) \neq minhash(B)) = \frac{|minhash(A) \neq minhash(B)|}{|AllPermutations|}$$

$$d_j(A,C) = P(minhash(A) \neq minhash(C)) = \frac{|minhash(A) \neq minhash(C)|}{|AllPermutations|}$$

$$d_j(C,B) = P(minhash(C) \neq minhash(B)) = \frac{|minhash(C) \neq minhash(B)|}{|AllPermutations|}$$

It is obvious that $minhash(C)$ can be $minhash(A)$, $minhash(B)$, or another value. Then, when $minhash(A) \neq minhash(B)$, either $minhash(C) \neq minhash(A)$, or $minhash(C) \neq minhash(B)$, or $minhash(C) \neq minhash(A) \neq minhash(B)$ is true.

Thus, $|minhash(A) \neq minhash(B)| \leq |minhash(A) \neq minhash(C)| + |minhash(C) \neq minhash(B)|$, which gives:

$$\frac{|minhash(A) \neq minhash(B)|}{|AllPermutations|} \leq \frac{|minhash(A) \neq minhash(C)|}{|AllPermutations|} + \frac{|minhash(C) \neq minhash(B)|}{|AllPermutations|}$$

This is equivalent to:

$$P(minhash(A) \neq minhash(B)) \leq P(minhash(A) \neq minhash(C)) + P(minhash(C) \neq minhash(B))$$

Which completes the proof.