Question 1 (marks = 50)

Shingling and Locality Sensitive Hashing

Problem Statement:

Given:

1. A number of paragraphs from two different books on two different topics.
2. The paragraphs are jumbled up and in no particular order.
3. The paragraphs are of varying length.

Input:

Input will be a text file ***Data.txt*** containing two columns, ***Para No*** (indicates the serial number of the paragraph) and ***Para**.*

To Do:

1. The paragraphs belonging to each book need to be separated based on their similarity.
2. Use Shingling with shingle size **K** = 5.
3. Cluster the similar paragraphs together to reconstruct the book, using k-means algorithm, where **k** = 2.

Output:

The output file produced by your code should be a text file containing the *Para No*s belonging to each book in separate lines. Each Para No belonging to a particular book should be separated by a comma.

Book 1: 1,4,6…
Book 2: 2,3,5….

Question 2 (marks = 50)

Problem Statement:

For the most similar 5 candidate pairs, of each book (set of paragraphs derived above), give the textual overlap regions:

1. The k-shingles that match and
2. Their position indices. Indicate their position in the paragraph, as nth Shingle. If the shingle is present more than once in the paragraph, indicate the first position.

Output:

The output file produced by your code should be a text file containing the following columns

<Para No **m** – Para No **n**><Shingle**1**><Index-**m** – Index-**n**>
<Para No **m** – Para No **n**><Shingle**2**><Index-**m** – Index-**n**>

<Para No **i** – Para No **j**><Shingle**3**><Index-**i** – Index-**j**>
<Para No **i** – Para No **j**><Shingle**4**><Index-**i** – Index-**j**>….