Question 1 (marks = 50)

Shingling, Minhashing 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 (the book number for each paragraph needs to be marked) based on their similarity.
2. Use Shingling with shingle size **K** = 5, and resulting Jaccard similarity as the similarity metric.
3. Use minhashing technique, followed by locality sensitive hashing for calculating candidate pairs, which are similar to each other.
4. Define a weighted graph where, vertices are paragraphs, edges are between candidate-pairs defined above, and edge weights are Jaccard similarity.
5. Cluster the paragraphs (vertices) using a graph-clustering algorithm, such as PIC algorithm, to reconstruct the book number for each paragraph. Here the number of clusters = 2, since there are 2 books. For PIC algorithm, refer: http://www.cs.cmu.edu/~wcohen/postscript/icml2010-pic-final.pdf https://spark.apache.org/docs/latest/mllib-clustering.html

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

1. The k-shingles that match and
2. The position indices of the shingles. Indicate their position in the paragraph, starting from beginning of paragraph, where the first index is 0. If the shingle is present more than once in the paragraph, indicate the first occurance.

Output:

Given that you have The output file produced by your code should be a text file containing the following columns:

<Para No **1 –** Para No **2**>   <Shingle**1**>   <Index-**11 –** Index-**21**>
<Para No **1 –** Para No **2**>   <Shingle**2**>   <Index-**12 –** Index-**22**>
…
…
…
<Para No **1 –** Para No **2**>   <Shingle**n**>   <Index-**1n –** Index-**2n**>

(Repeats for the 5 matches).