Data Wrangling and Data Analysis
Lab4 Exercises: Heterogeneous Data Integration

Question 1.
Compute the Jaccard similarities of each pair of the following three sets: {1, 2, 3, 4}, {2, 3, 5, 7}, and {2, 4, 6}.

Question 2.
Compute the Jaccard bag similarity of each pair of the following three bags: {1, 1, 1, 2}, {1, 1, 2, 2, 3}, and {1, 2, 3, 4}.

Question 3.
Let C(D1) = {aa, bb, ab , ba}, C(D2) = {aa, ac, ca, ba}, C(D3) = {ab, ba, ca} be the 2-shingle representation of documents D1, D2, D3. Create the matrix representation of the shingles-documents relationship.

Question 4.
What is the linked open data cloud?

Question 5.
Compute the Jaro and Jaro-Winkler similarity between arnab and urban. Why do you think you got this result?

Question 6.
Compute the edit distance between Recreation and Regeneration assuming that substitution cost is 1. What if we consider the substitution as insertion and deletion (cost 2)?

Question 7.
Compute the gap distance between "Journal of Knowledge and Data Engineering" and "J. of Knowl. and Data Eng." assuming that the open gap cost = 1 and extend gap cost = 0.1.

Question 8.
Consider the following document "need an efficient technique to group records if they match". What will be the cardinality of the set that contains the 6-shingles of the document?

Question 9.
Use python libraries to compute three types of similarities between Journal and Formal.

Question 10.
Consider the shingle-document matrix that you produced in Question 3 and assume the following permutations:
p1 = {aa,bb,ab,ba,ac,ca},
p2 = {ca, ac, ba, ab, bb, aa},
p3 = {ac, ca, ab, ba, bb, aa}.
The hash functions are defined to be the first non-0 row of the document representation.
a. Create the signature matrix of the documents.
b. Compare the Jaccard similarity of each pair of documents with the Jaccard similarity of their signatures.

Question 11.
For the information to be integrated, heterogeneity can be in a set of levels. What are these levels?

Question 12.
Where the mappings in the information integration system are implemented (found)?

Question 13.
What is the pairwise Jaccard dissimilarity between C(D1), C(D2) and C(D3) in Question 3?