Assignmnet -2 Similar Items (Min Hashing)

Download and use any news corpus data consisting of sports, politics, and movies (max 200 docs each).

1. Construct several types of K-shingles (5, 8 and 10).

2. Build a minhash signature for the above K-shingles.

3. Compute the Jaccard similarity between all pairs of documents for each type of shingle. Compare the pair-wise similarity between Signature matrix and Singling Matrix.

4. Use any other similarity function and find the Signature matric based similarity.

5. Plot a graph with your conclusions and findings.