IT350 Assignment 2 - Report

Sachin Prasanna - 211T058

February 7, 2024

1 Problem Statement

Download and use any news corpus data consisting of sports, politics, and movies (or other domains of your preference) (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 Shingling Matrix.
4. Use any other similarity function and find the Signature matric based similarity.
5. Plot a graph with your conclusions and findings.

2 Dataset

The selected news corpus dataset for this assignment was the AG news dataset which is a collection of more than 1 million news articles. Over the course of its one-year-plus operation, ComeToMyHead, an academic news search engine established in July 2004, has collected news articles from over 2000 sources. This dataset consists of 120,000 training samples and 7600 testing samples of news articles that contain 3 columns. The first column is Class Id, the second column is Title and the third column is Description. The class ids are numbered 1-4 where 1 represents World, 2 represents Sports, 3 represents Business and 4 represents Sci/Tech. The dataset can be accessed here - https://huggingface.co/datasets/ag_news

3 Preprocessing

The testing file was used in the assignment. Since the assignment requires only 3 domains, the 4th domain representing Sci/Tech was dropped. Following this, a 'char_count' feature was engineered for each entry in the 'Description' column, detailing the character count of the descriptions.

Subsequently, a selection process was conducted to obtain a representative subset from each class. This involved randomly choosing 10 instances from each class for character shingles and 50 instances from each class for word shingles. To facilitate this, a DataFrame was initialized, and within a loop iterating over each class index, the DataFrame was filtered to isolate instances belonging to that specific class.

The resulting DataFrame was constructed by concatenating these selected entries across all classes. Extraneous columns such as 'Title' and 'char_count' were removed for clarity. Additionally, a new 'ID' column was appended to facilitate referencing, and the text within the 'Description' column underwent preprocessing steps. These steps involved stripping leading and trailing whitespace, removing non-alphanumeric characters, and converting all text to lowercase.

4 Shingling

For the assignment, two distinct types of shingling methodologies, namely Character Shingling and Word Shingling, were employed for analysis.

In **Character Shingling**, each shingle corresponds to a fixed-size sequence of characters extracted from the text data. This approach is particularly effective for capturing patterns and similarities at a character level, which can be beneficial in scenarios where subtle nuances in text need to be accounted for.

On the other hand, **Word Shingling** involves dividing the text into individual words, with each word serving as a shingle. It allows for a more abstract representation of text data, focusing on the content rather than the specific characters.

Shingles were constructed for three different values of k: k = 5, k = 8, and k = 10 for both Character and Word shingles.

5 Min Hashing

The infinity method of Min Hashing was used to calculate the signatures. This method can be found in the Book **Mining of Massive Datasets** under the chapter Finding Similar Items under the section 3.3.5. The algorithm iterates through each row of the shingle dataframe and examines each column. For each column where the shingle was present, the hash functions were applied. The resulting hash values were then utilized to update the signature matrix, ensuring that the minimum hash value was retained at each corresponding position. Upon completion, the resulting signature matrix was returned as the output of the function.

The number of hash functions was calculated based on the percentage provided and the size of the input shingle dataframe. A list of hash functions, characterized by random coefficients 'a' and 'b', was generated accordingly.

The python code of the algorithm is displayed in the following page:

```
def min_hashing(shingle_df, percentage):

    num_hash_functions = int(shingle_df.shape[0] * percentage)
    hash_functions = [(np.random.randint(1, 100), np.random.randint(1, 100
                                         )) for _ in range(
                                         num_hash_functions)]
    signature_matrix = pd.DataFrame(np.inf, index=range(1,
                                         num_hash_functions + 1), columns=
                                         shingle_df.columns.difference(['
                                         Shingle']))
    shingle_index_map = {shingle: idx for idx, shingle in enumerate(
                                         shingle_df.index)}

    for index, row in shingle_df.iterrows():
        for col in shingle_df.columns:
            if row[col] == 1:
                for i, (a, b) in enumerate(hash_functions, start=1):
                    hash_value = (a * shingle_index_map[index] + b) \%
                                                         shingle_df.shape[
                                                         0]
                    signature_matrix.loc[i, col] = min(signature_matrix.
                                                         loc[i, col],
                                                         hash_value)

    return signature_matrix
```

6 Similarity Metrics

6.1 Jaccard Similarity

Jaccard Similarity is a measure used to compare the similarity and dissimilarity between two sets. It is defined as the size of the intersection of the sets divided by the size of the union of the sets. In other words, it quantifies the proportion of shared elements between two sets relative to the total number of distinct elements in both sets.

Mathematically, the Jaccard Similarity (J) between sets A and B is calculated as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where:
- $|A \cap B|$ represents the size of the intersection of sets.
- $|A \cup B|$ represents the size of the union of sets A and B.

The Jaccard Similarity coefficient ranges from 0 to 1, where a value of 1 indicates that the sets are identical, and a value of 0 indicates that the sets have no elements in common.

6.2 Dice Similarity

Dice Similarity is another measure used to quantify the similarity between two sets. Like Jaccard Similarity, it considers the overlap between sets, but it places more emphasis on the

relative size of the intersection to the sizes of the individual sets. It is particularly useful when dealing with sparse binary data.

Mathematically, the Dice Similarity (D) between sets A and B is calculated as:

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

where:

- $|A \cap B|$ represents the size of the intersection of sets.
- $|A|$ represents the size of set A.
- $|B|$ represents the size of set B.

The Dice Similarity coefficient also ranges from 0 to 1. A value of 1 indicates perfect overlap, meaning that the sets are identical, while a value of 0 indicates no overlap between the sets.

7 Conclusions and Findings

7.1 Character Shingling

Character Shingling was done and compared using the Jaccard Similiarity and Dice Similarity measures. The were done for k = 5, k = 8 and k = 10 shingles and retaining 10, 20 and 30 percent of the shingles. Figures 1-12 were obtained as a result of visualising the data using Jaccard Similarity. Figures 13-24 were obtained as a result of visualising the data using the Dice Similarity. The analysis yields the following conclusions:

1. When retaining 10% of the data, World news articles had a maximum of 14% column similarity when 5 shingles were taken (Jaccard Similarity). This was well approximated by the signature matrix as well. Similar trends were observed for other classes as well. (For both Similarity Metrics)

2. When retaining 20% and 30% of the data, similar results are observed but the signature matrices better approximate the shingled matrices, showing that more retention is better.

3. Between the 3 classes of documents, documents from the World News category seem to be the most similar to each other. This is followed by the Sports news documents and then finally the Business News documents.

4. The trend of World News articles exhibiting higher similarity within the class compared to Sports and Business News suggests that there may be common themes or topics within World News articles that contribute to this similarity, such as global events or geopolitical issues.

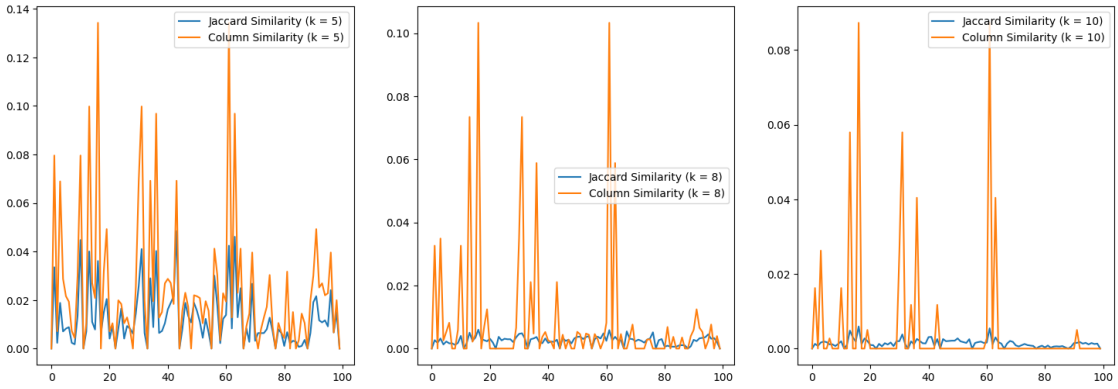

Figure 1: Jaccard analysis of World News Similarity retaining 10%

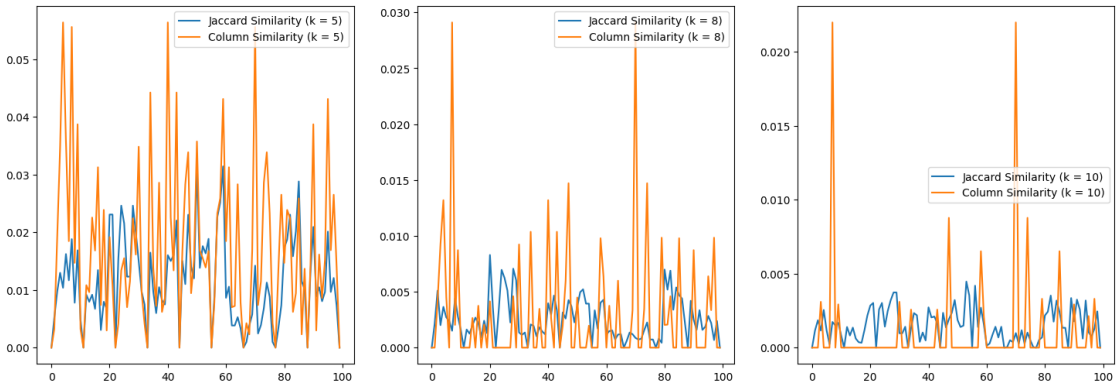

Figure 2: Jaccard analysis of Sports News Similarity retaining 10%

5. Despite the observed similarities within each class, there are still notable differences between documents - which means that running on a larger number of documents had the potential to give better results.

7.2 Word Shingling

Word Shingling was done and compared using the Jaccard Similiarity and Dice Similarity measures. The were done for k = 5, k = 8 and k = 10 shingles and retaining 10, 20 and 30 percent of the shingles. Figures 25 and 26 were the Jaccard and Dice similarity graphs when 10% of the data was retained. All other graphs were omitted because word singling was not that powerful as compared to character shingling for this dataset.

1. Compared to character shingling, word shingling showed much lesser similarity both in signature matrices as well as in the shingle matrices.

2. This is because a set of words appearing back to back commonly is much rarer than characters.

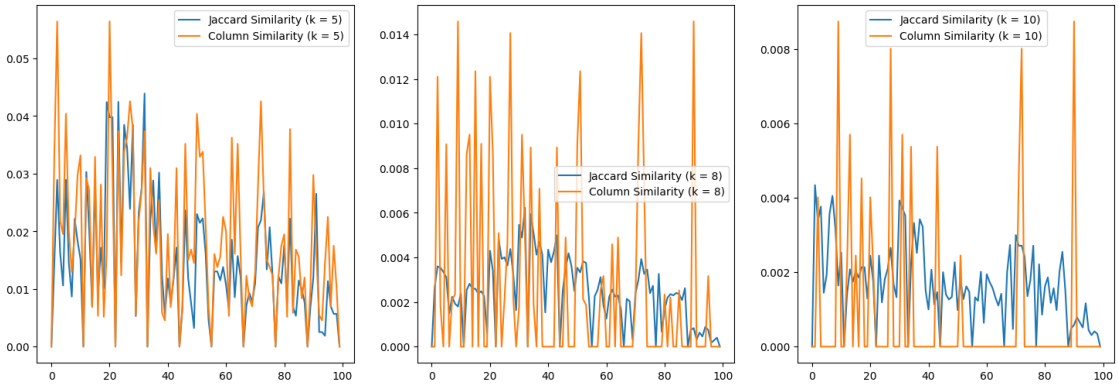

Figure 3: Jaccard analysis of Business News Similarity retaining 10%

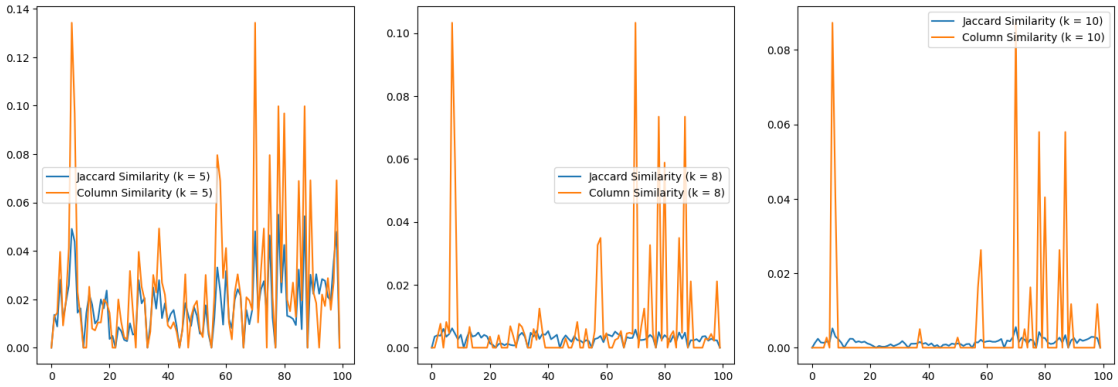

Figure 4: Jaccard analysis of World News Similarity retaining 20%

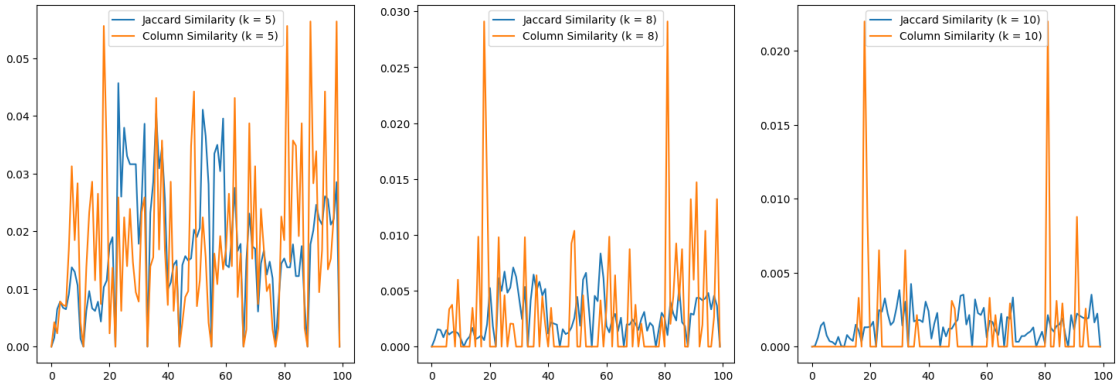

Figure 5: Jaccard analysis of Sports News Similarity retaining 20%

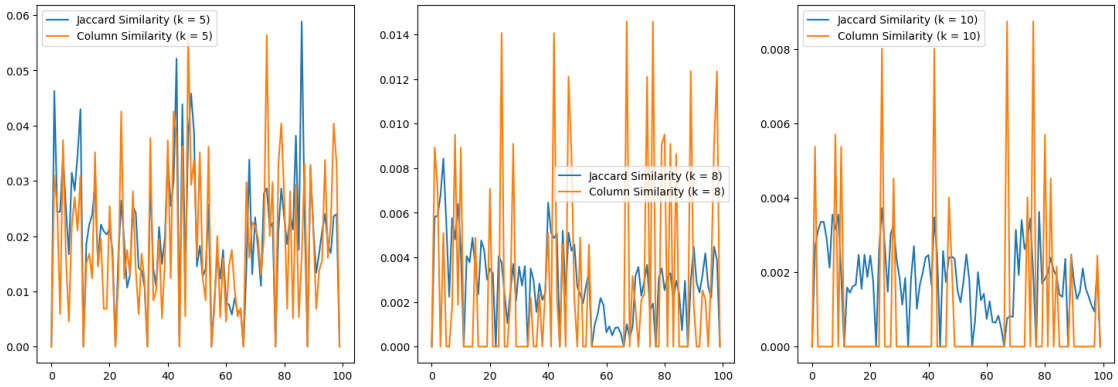

Figure 6: Jaccard analysis of Business News Similarity retaining 20%

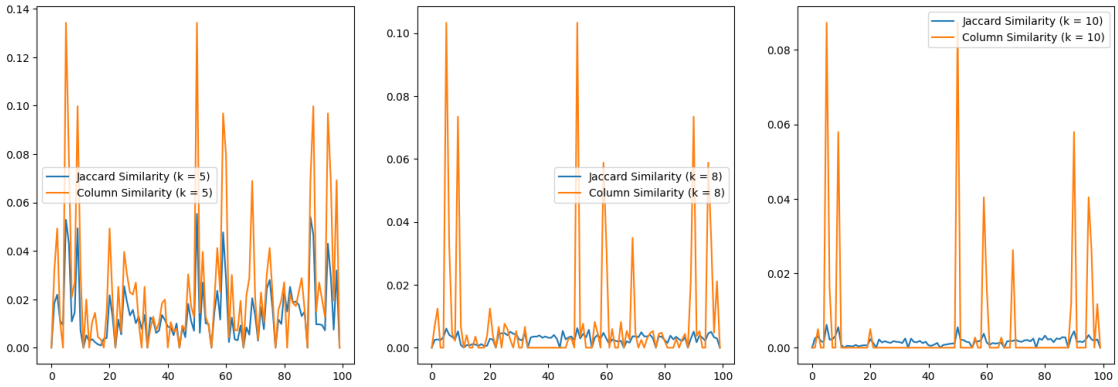

Figure 7: Jaccard analysis of World News Similarity retaining 30%

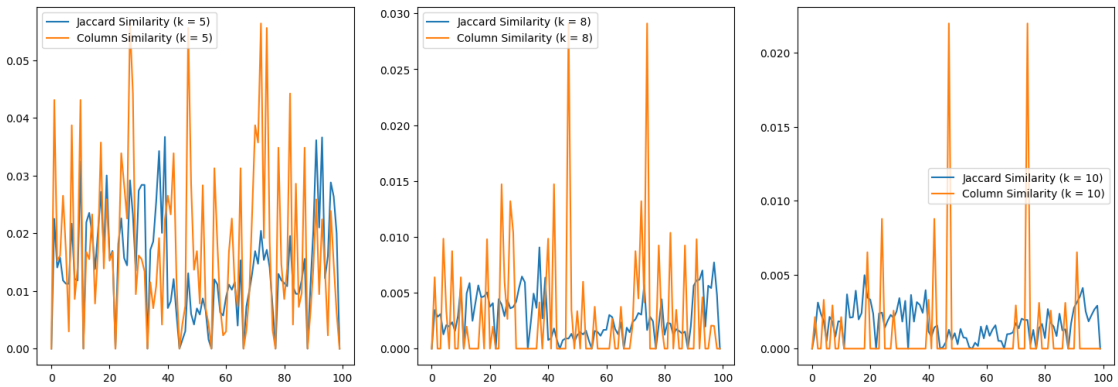

Figure 8: Jaccard analysis of Sports News Similarity retaining 30%

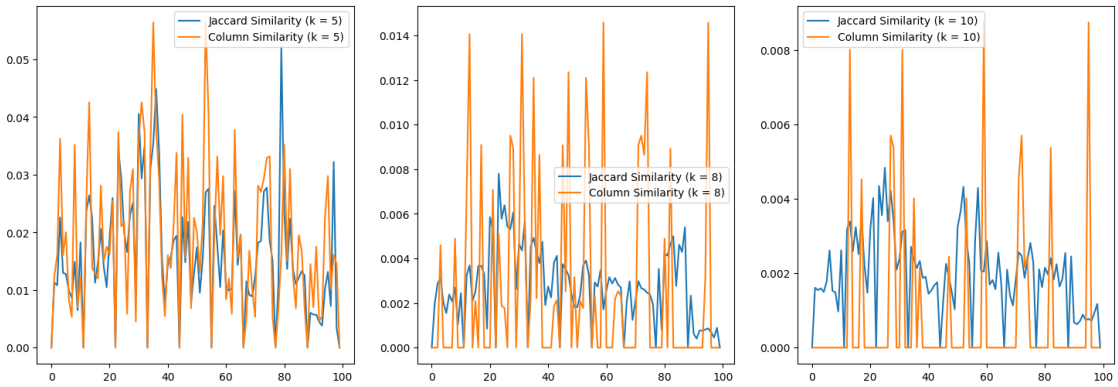

Figure 9: Jaccard analysis of Business News Similarity retaining 30%

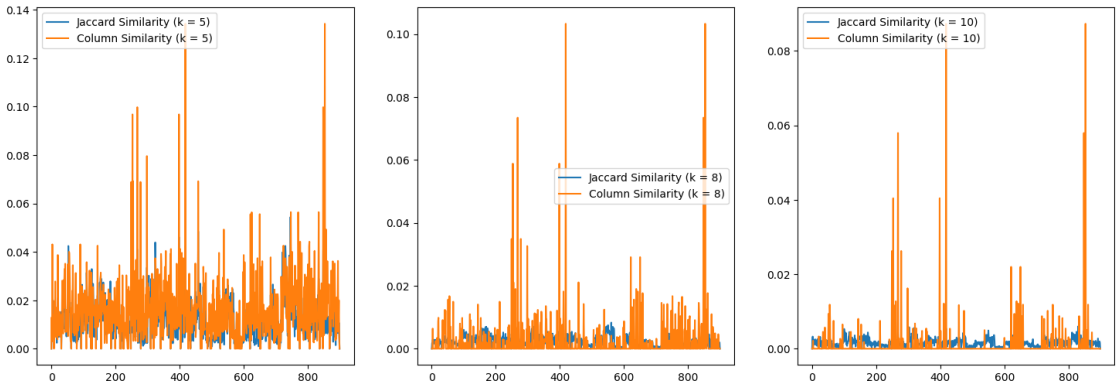

Figure 10: Jaccard analysis of all News Articles retaining 10%

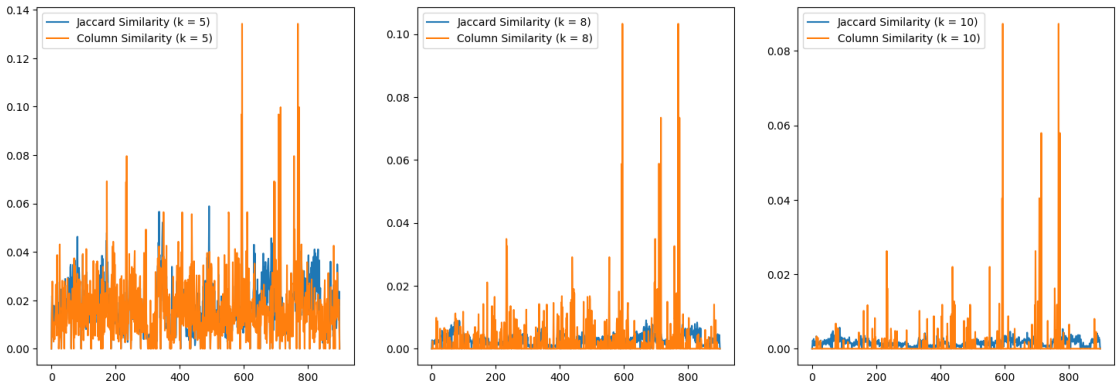

Figure 11: Jaccard analysis of all News Articles retaining 20%

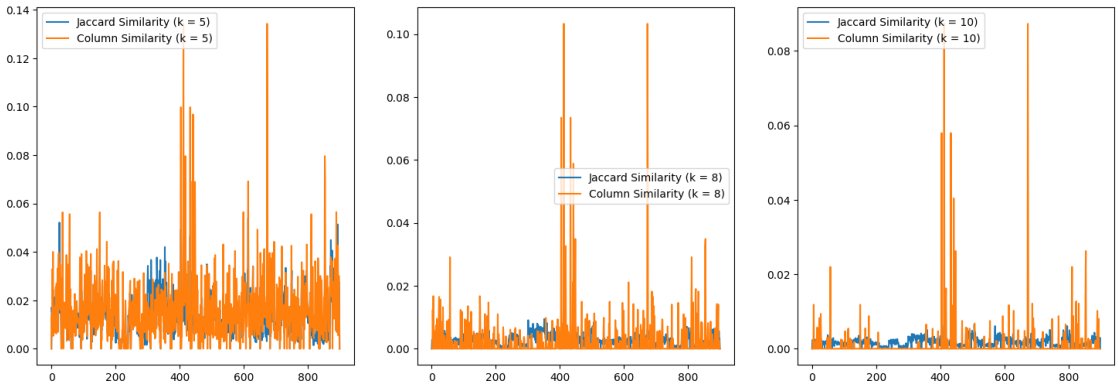

Figure 12: Jaccard analysis of all News Articles retaining 30%

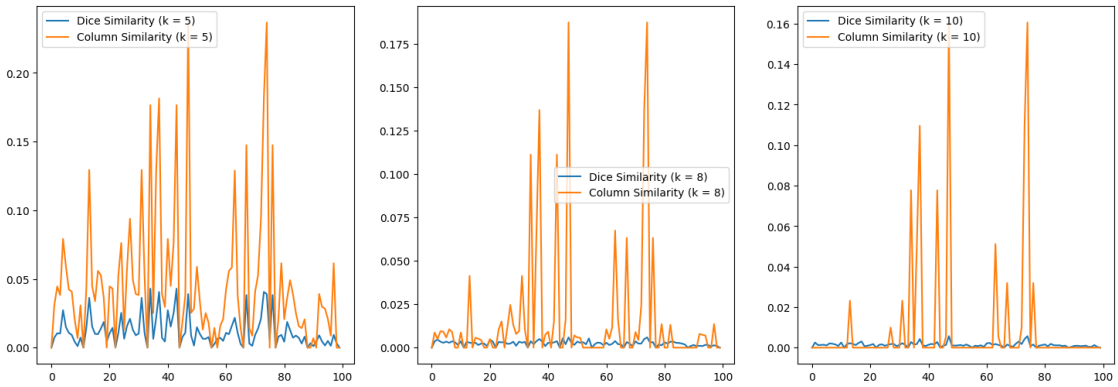

Figure 13: Dice analysis of World News Similarity retaining 10%

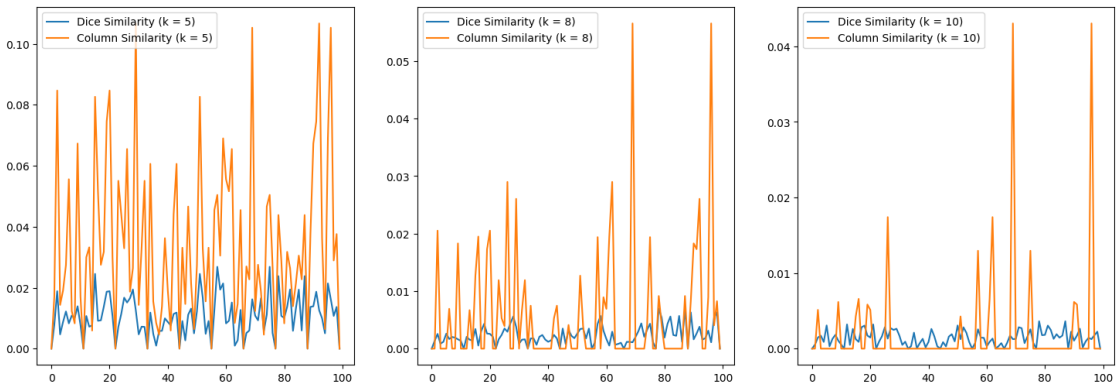

Figure 14: Dice analysis of Sports News Similarity retaining 10%

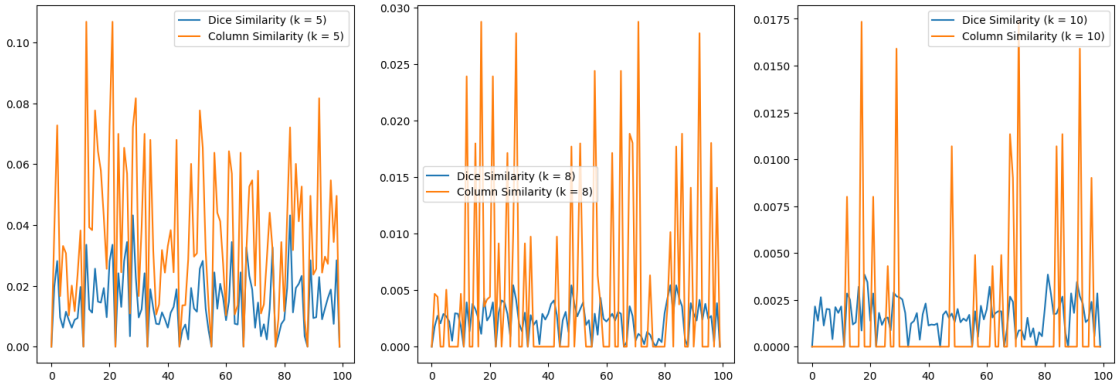

Figure 15: Dice analysis of Business News Similarity retaining 10%

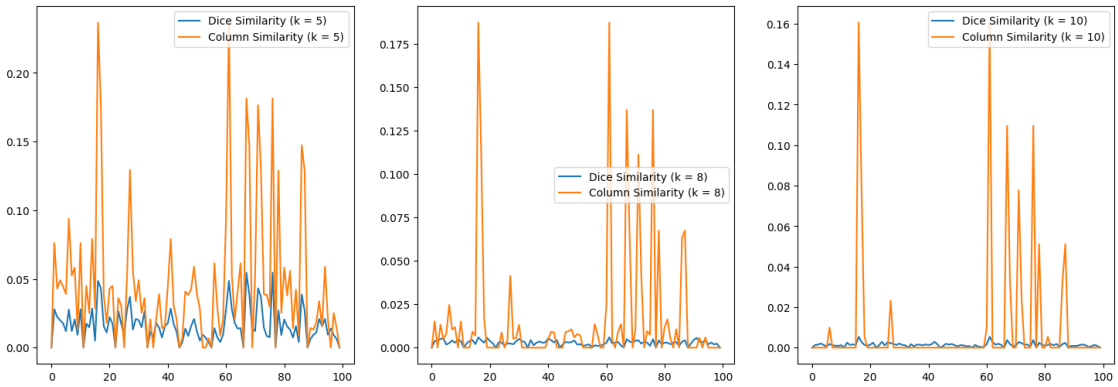

Figure 16: Dice analysis of World News Similarity retaining 20%

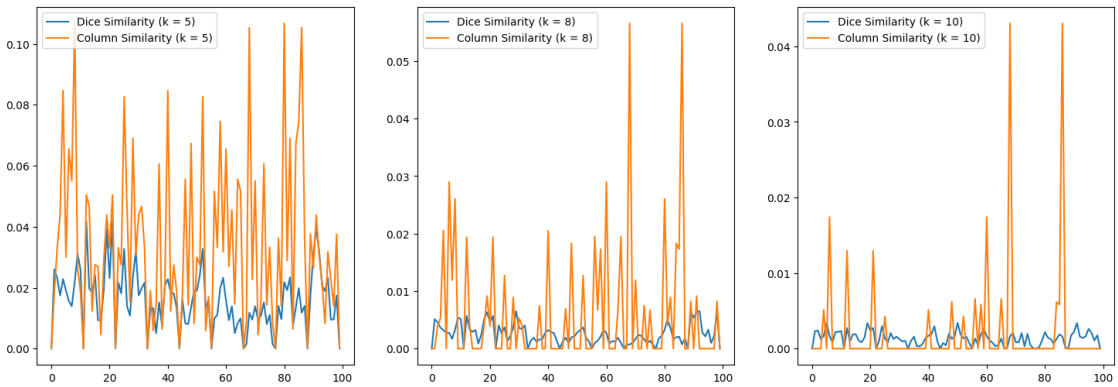

Figure 17: Dice analysis of Sports News Similarity retaining 20%

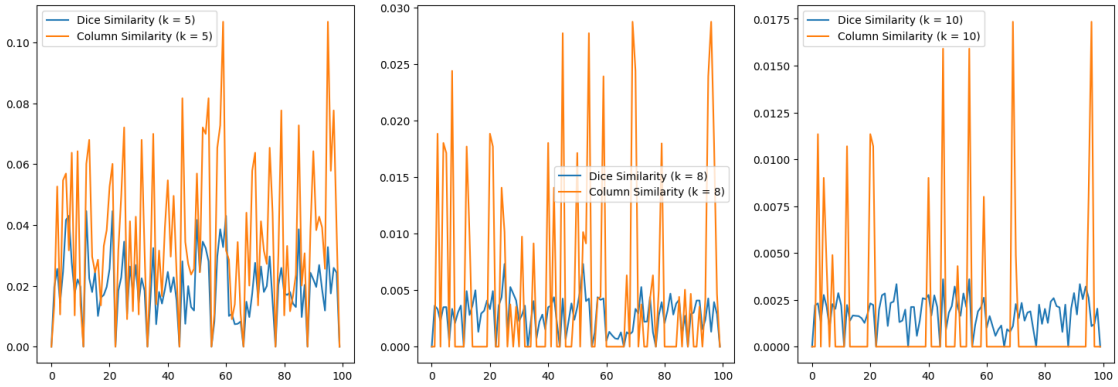

Figure 18: Dice analysis of Business News Similarity retaining 20%

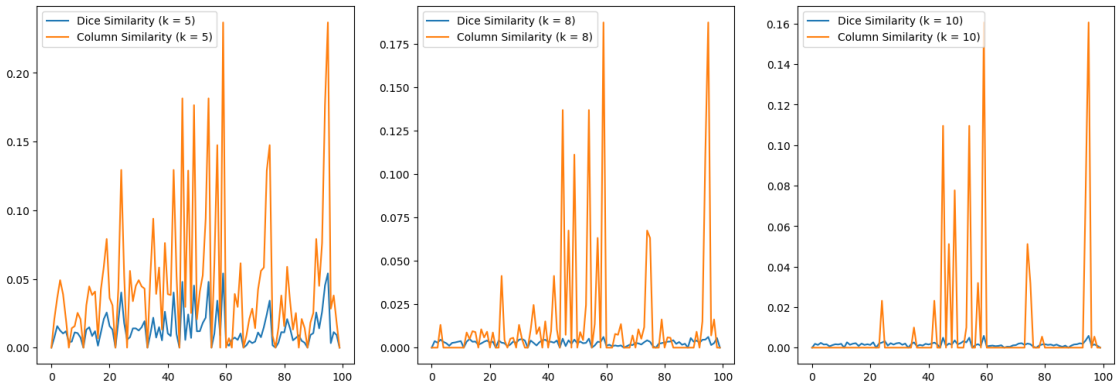

Figure 19: Dice analysis of World News Similarity retaining 30%

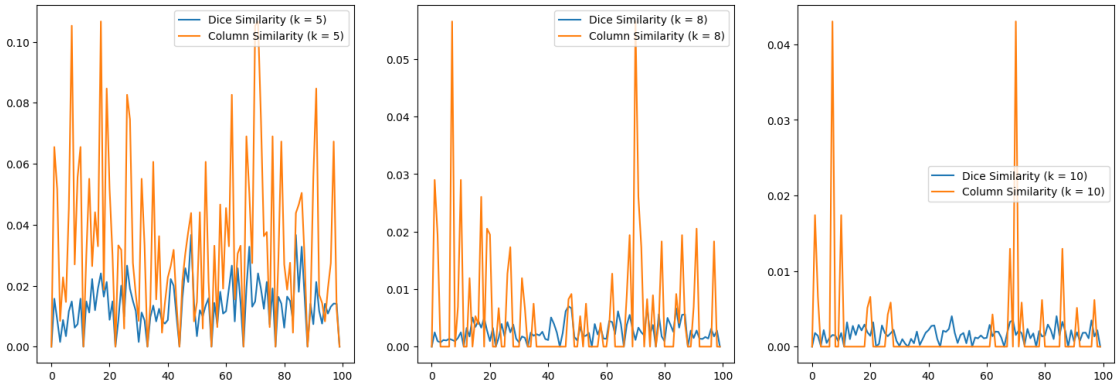

Figure 20: Dice analysis of Sports News Similarity retaining 30%

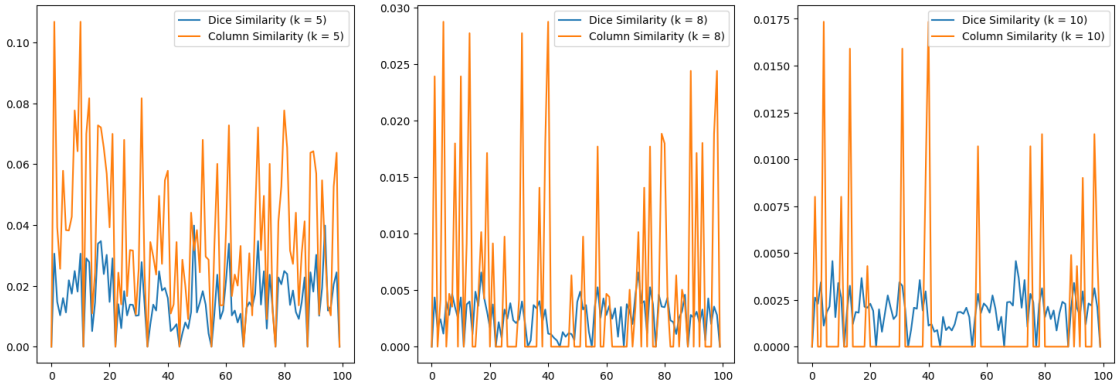

Figure 21: Dice analysis of Business News Similarity retaining 30%

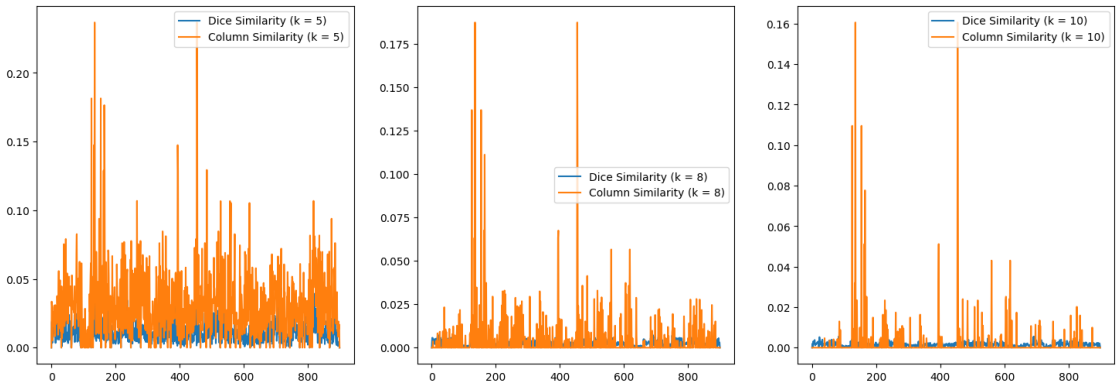

Figure 22: Dice analysis of all News Articles retaining 10%

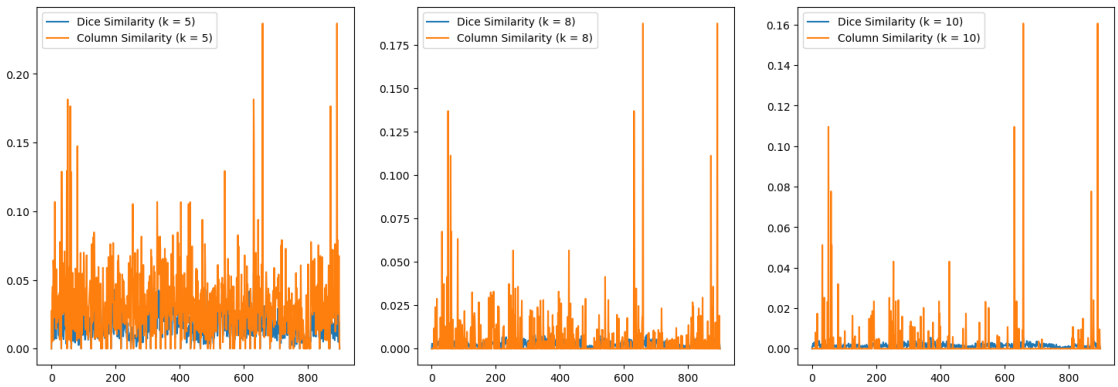

Figure 23: Dice analysis of all News Articles retaining 20%

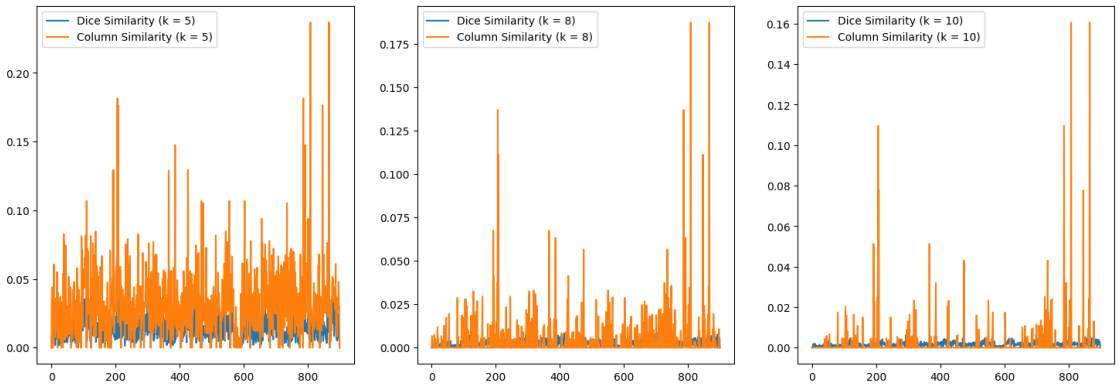

Figure 24: Dice analysis of all News Articles retaining 30%

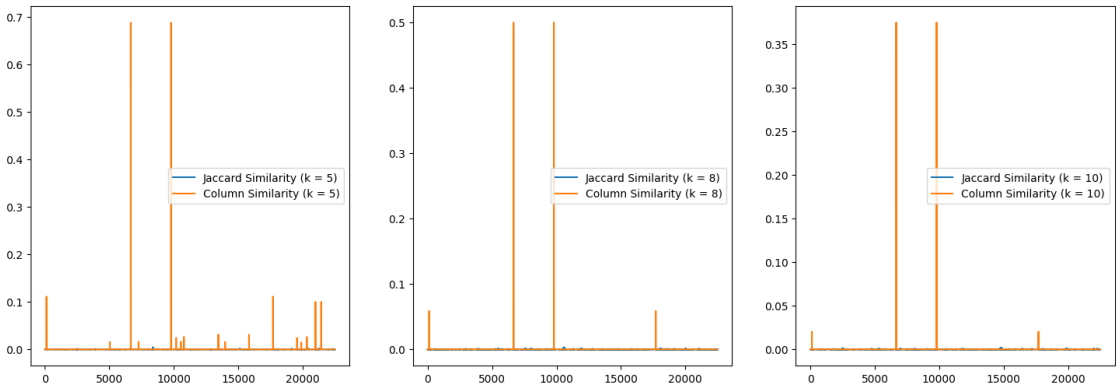

Figure 25: Jaccard analysis of all News Articles retaining 10% - Word Shingling

3. Exploring hybrid approaches that combine both character and word shingling techniques could offer a more comprehensive analysis of document similarities, leveraging the advantages of each approach to capture both fine-grained and high-level textual patterns effectively.

8 Code

The code associated with this report and assignment can be found on the GitHub Repository - https://github.com/sachinprasanna7/Min-Hashing

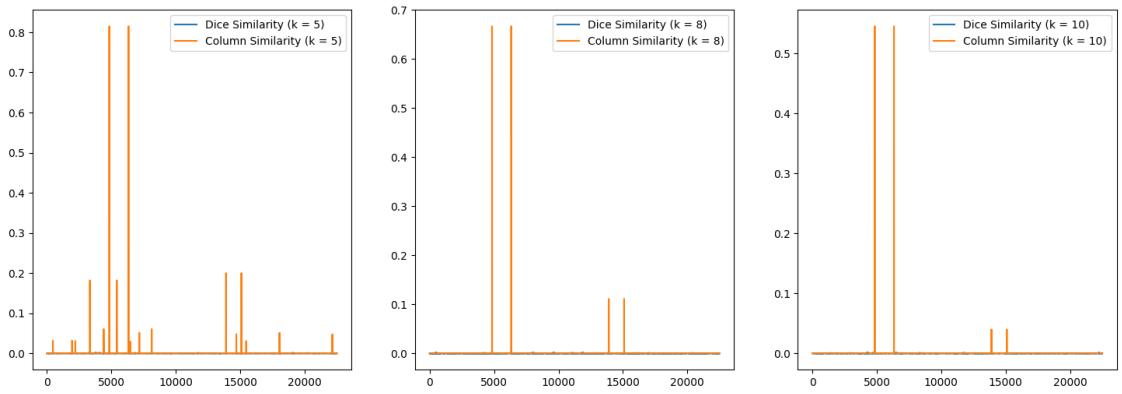

Figure 26: Dice analysis of all News Articles retaining 10% - Word Shingling