BU.330.740 Large Scale Computing on the Cloud

**Assignment 1. Big Data Solution – MapReduce**

Question 1. Market Basket Analysis (MBA) is a data mining technique used to uncover patterns in customer purchasing behavior. Specifically, it identifies relationships between items that are frequently bought together. By analyzing transactional data, MBA helps businesses understand associations between products and provides actionable insights to improve sales and customer satisfaction. MBA has significant applications across industries, particularly in retail, e-commerce, and marketing, delivering measurable business value in cross-selling, targeted promotions and discounts, product bundling, store layout optimization, inventory management, and so on.

In this exercise, you will use the MapReduce model to identify frequent item pairs from a dataset of transactions. You are provided a dataset containing a large list of transactions. Each line represents a transaction, with items separated by commas. Example:

Transaction 1: {milk, bread, cereal}

Transaction 2: {milk, sugar, bread, eggs}

Transaction 3: {milk, bread, butter}

Transaction 4: {sugar, eggs}

Transaction 5: {strawberry, yogurt, cereal, blueberry}

Transaction 6: {strawberry, milk}

Describe your MapReduce solution, including the keys, values, and operations used in the mapper and reducer.

Question 2. The plagiarism check problem involves detecting instances where text, code, or other content in a document has been copied from another source without proper attribution. This problem is significant in education, where students are expected to submit original work in essays, assignments, projects, or research papers. Plagiarism detection is essential for maintaining academic integrity, fostering a culture of originality, and ensuring fair assessment.

In this exercise, you will use the MapReduce model to identify documents having highly matching content. Suppose we have:

Article 1: sentence 1, sentence 2, sentence 3, …

Article 2: sentence 4, sentence 5, sentence 6, …

Article 3: sentence 7, sentence 8, sentence 9, …

Article 4: sentence 1, sentence 5, sentence 8, …

Your output will be each pair of articles and the number of their matching sentences. Describe your MapReduce solution, including the keys, values, and operations used.

**Hint**: You will need two pairs of mappers and reducers to solve this problem.