ASSIGNMENT REPORT

*Fundamental of Big Data Analytics*

*(DS2004)*

****

**SALMAN ZAIDI . MARIA KHAN . AMNA SHAHID**

21 I 1738 . 21 I 1352 . 21 I 1713

ASSIGNMENT 2

6.3.2024

DR. MUHAMMAD ISHTIAQ

### 

### Developing a Naïve Search Engine Utilizing Map Reduce

### Technology

…

…

…

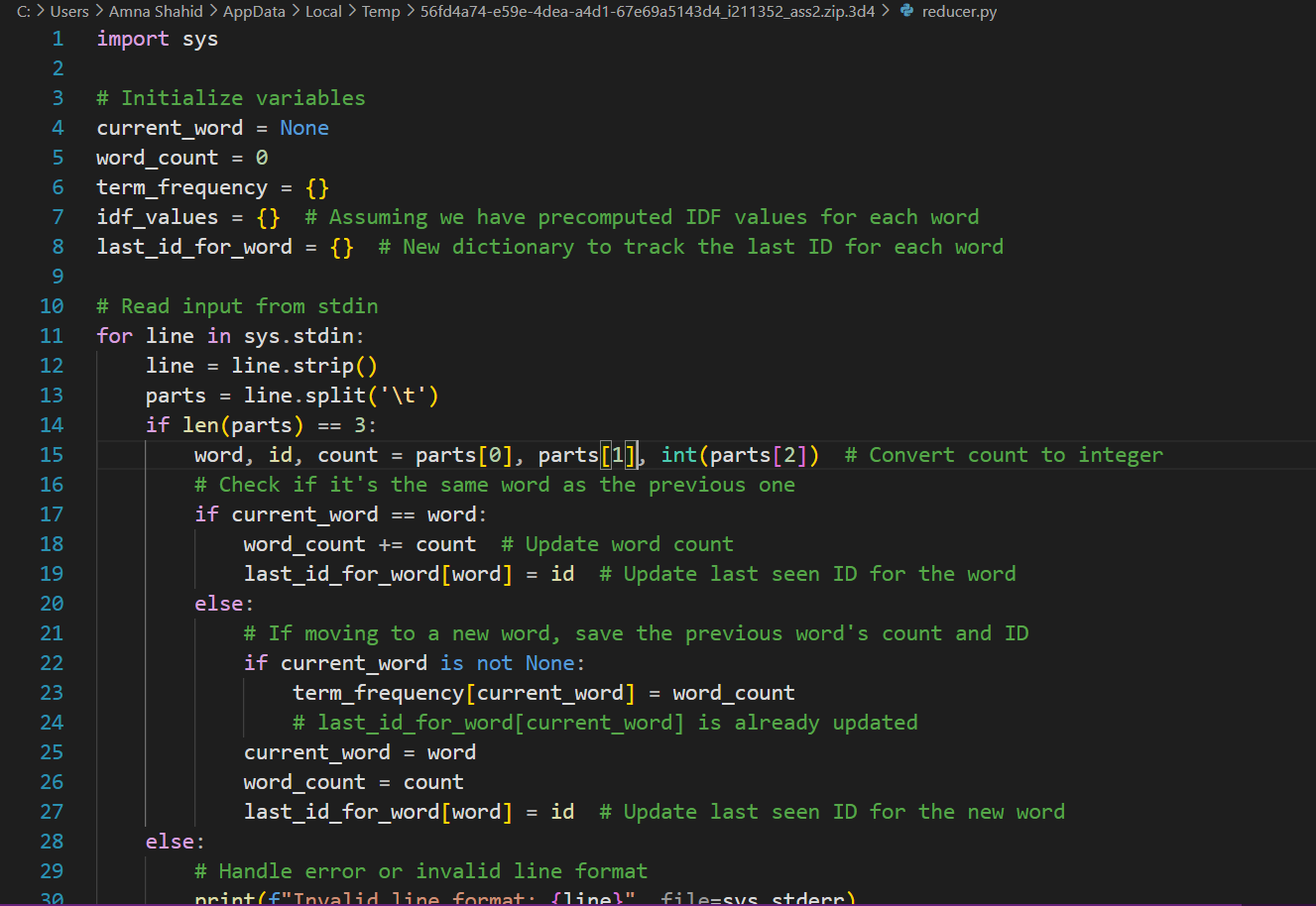
### PART I: QUESTION 1 , 2 and 3

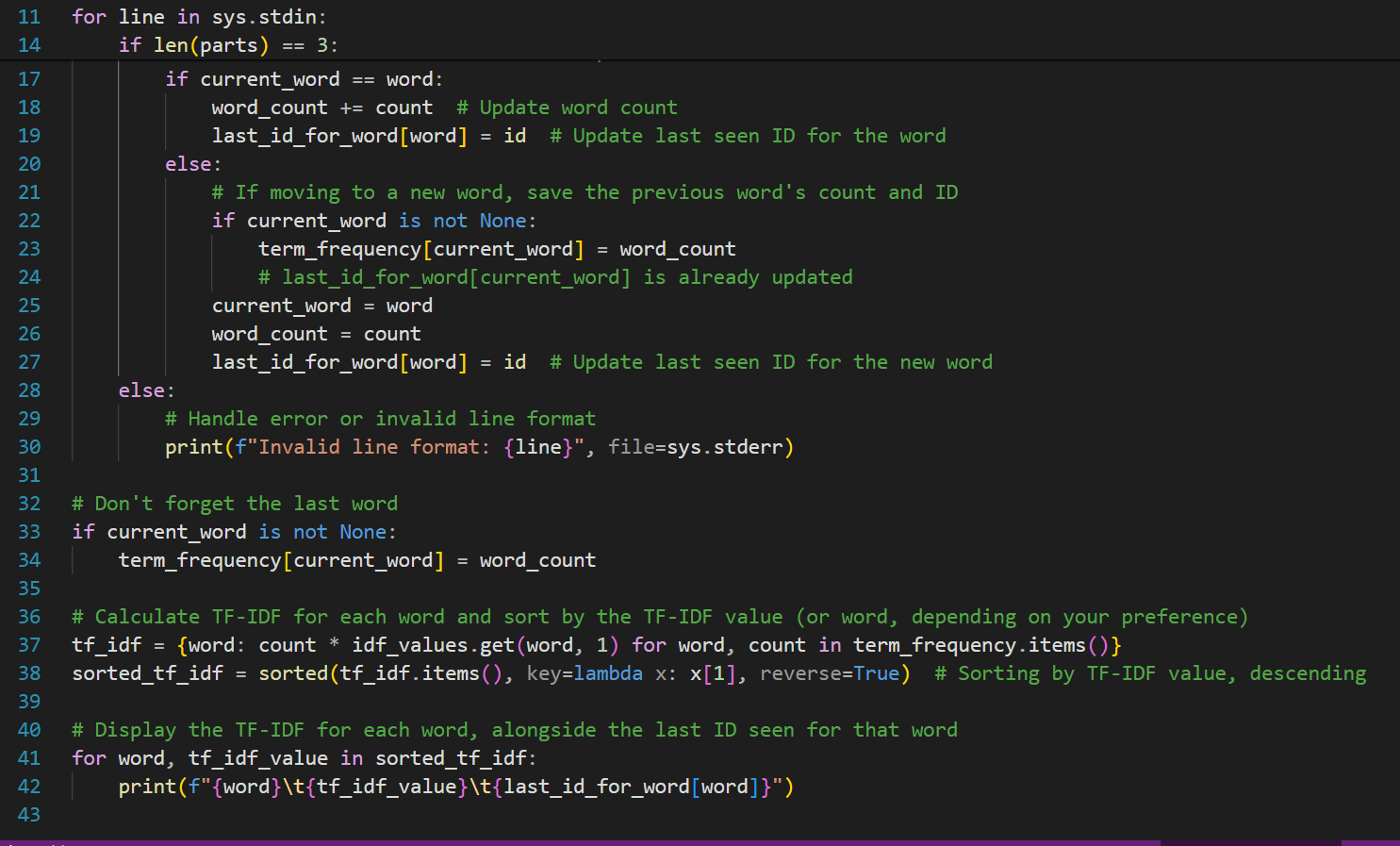
**Mapper.py**

### 

The mapper takes input data, assuming it's in CSV format, and extracts the text from a specific column (SECTION\_TEXT). It then converts the text to lowercase, removes any extra spaces, and splits it into individual words. For each word, it assigns a unique ID (if it's a new word) and counts how many times it appears. Finally, it outputs each word along with its ID and count.

**Reducer.py**

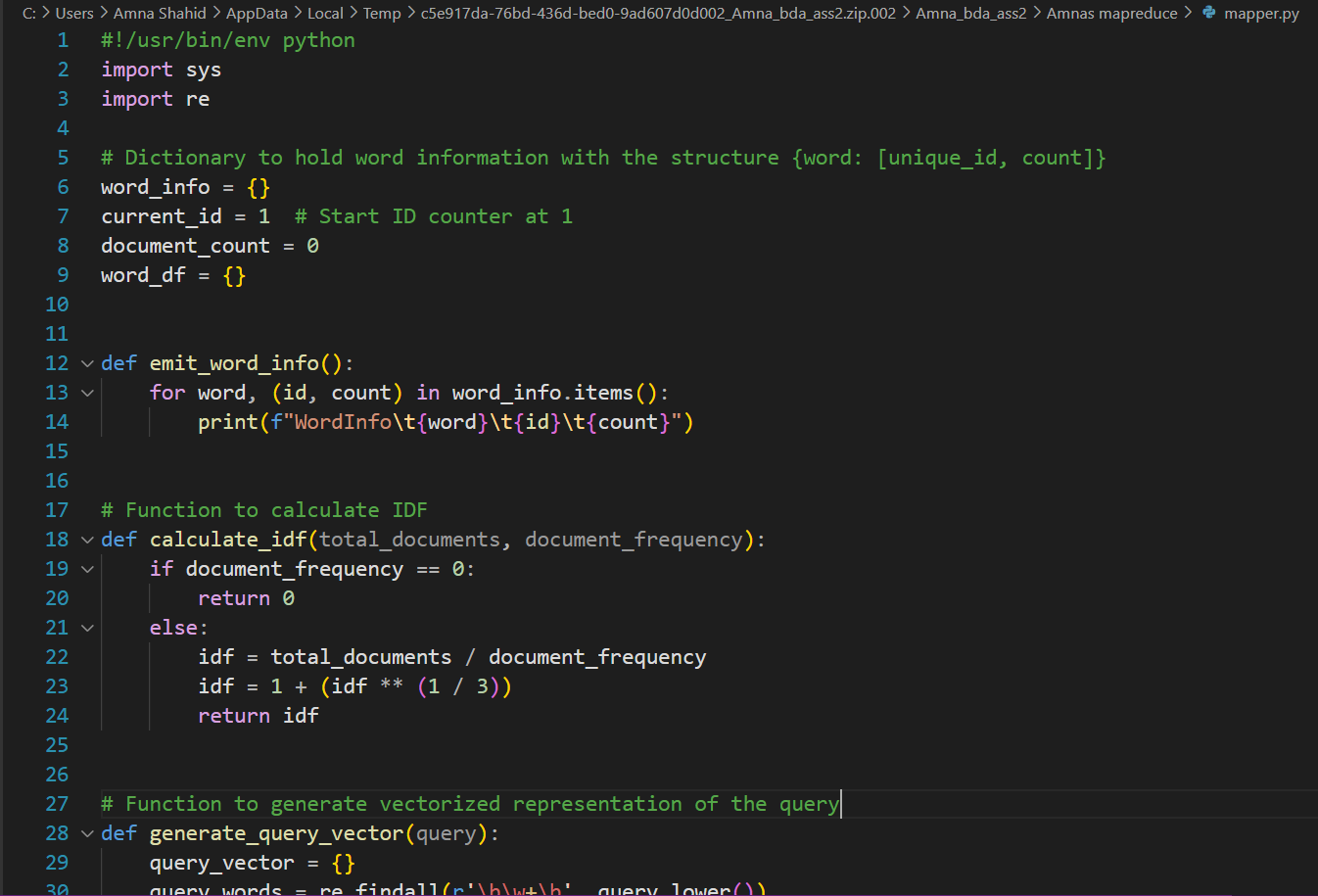


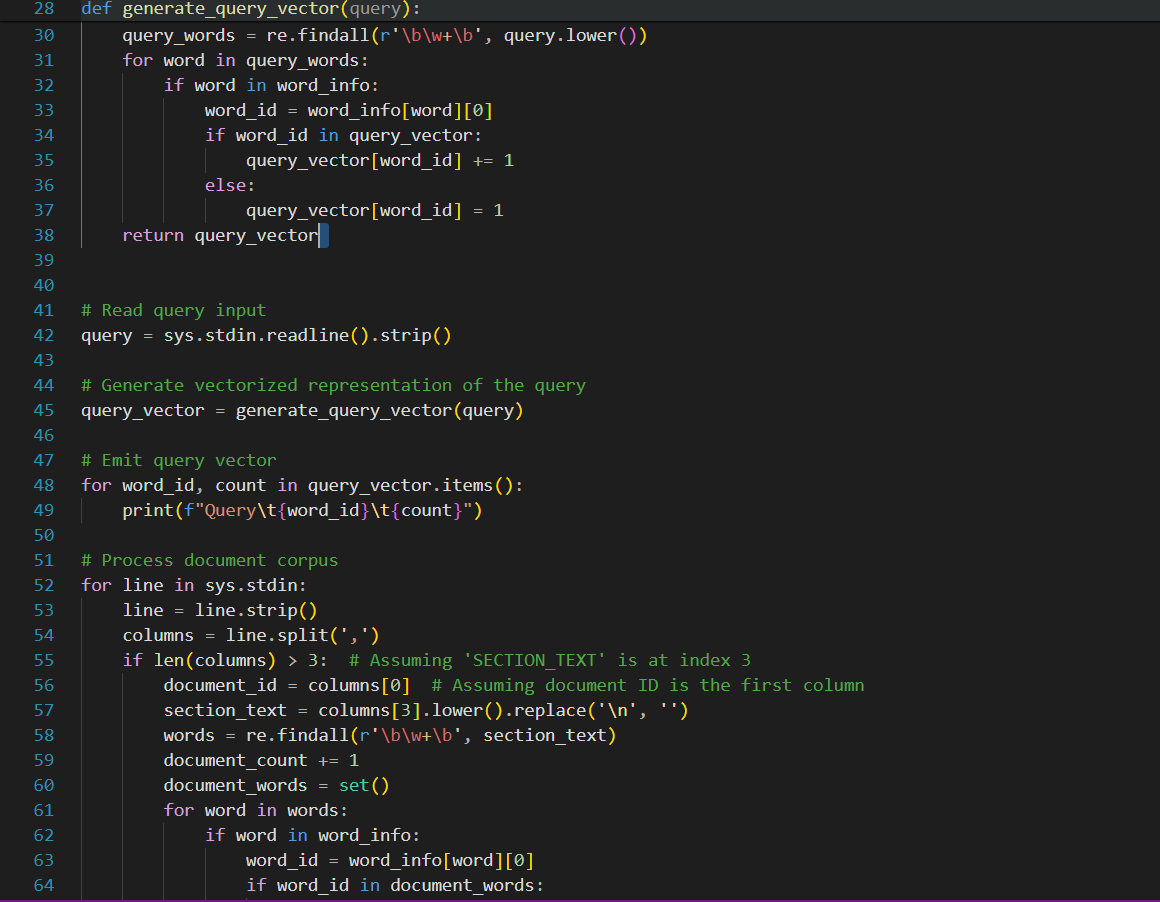


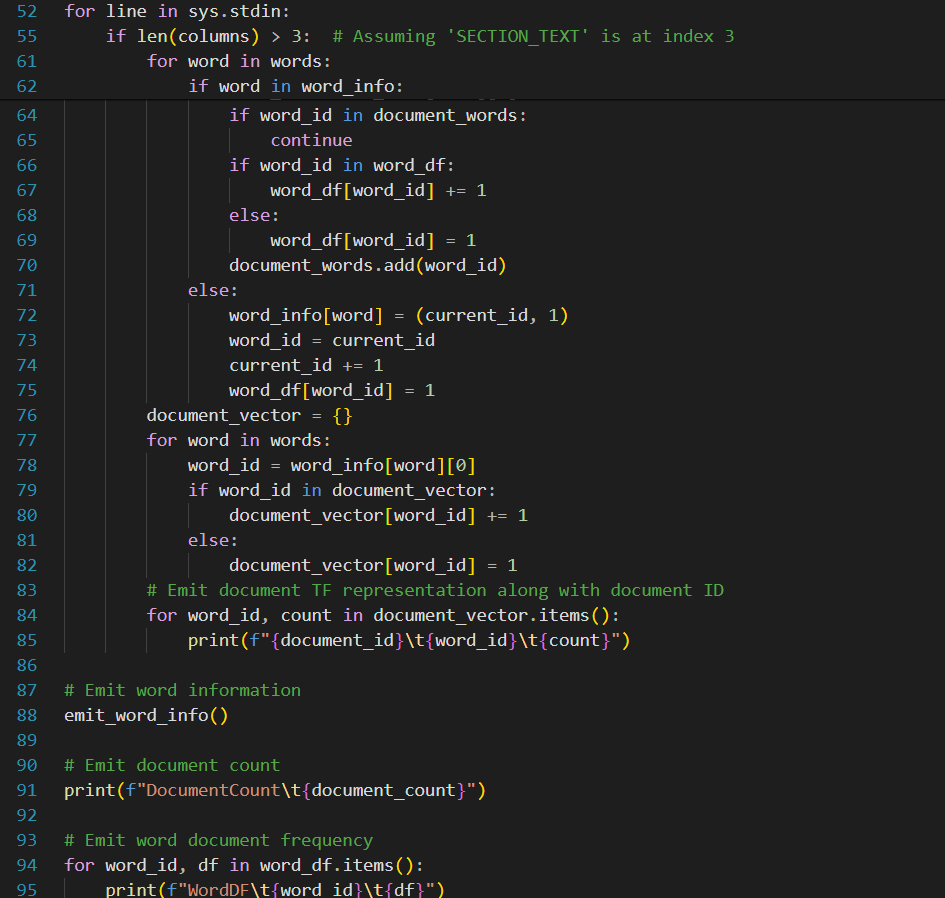
The reducer receives the output from the mapper, which consists of words, their IDs, and counts. It calculates the Term Frequency (TF) for each word (how often it appears in a document). It also assumes there are precomputed IDF (Inverse Document Frequency) values for each word. Then it computes the TF-IDF for each word by multiplying TF with IDF.After calculating TF-IDF for each word, it sorts them in descending order based on their TF-IDF values. Finally, it prints each word along with its TF-IDF value and the last seen ID for that word. In simpler terms, the mapper counts how many times each word appears in the input data, and the reducer calculates how important each word is in the context of the entire dataset.

### PART II: QUESTION 4 and 5

**Mapper.py:**

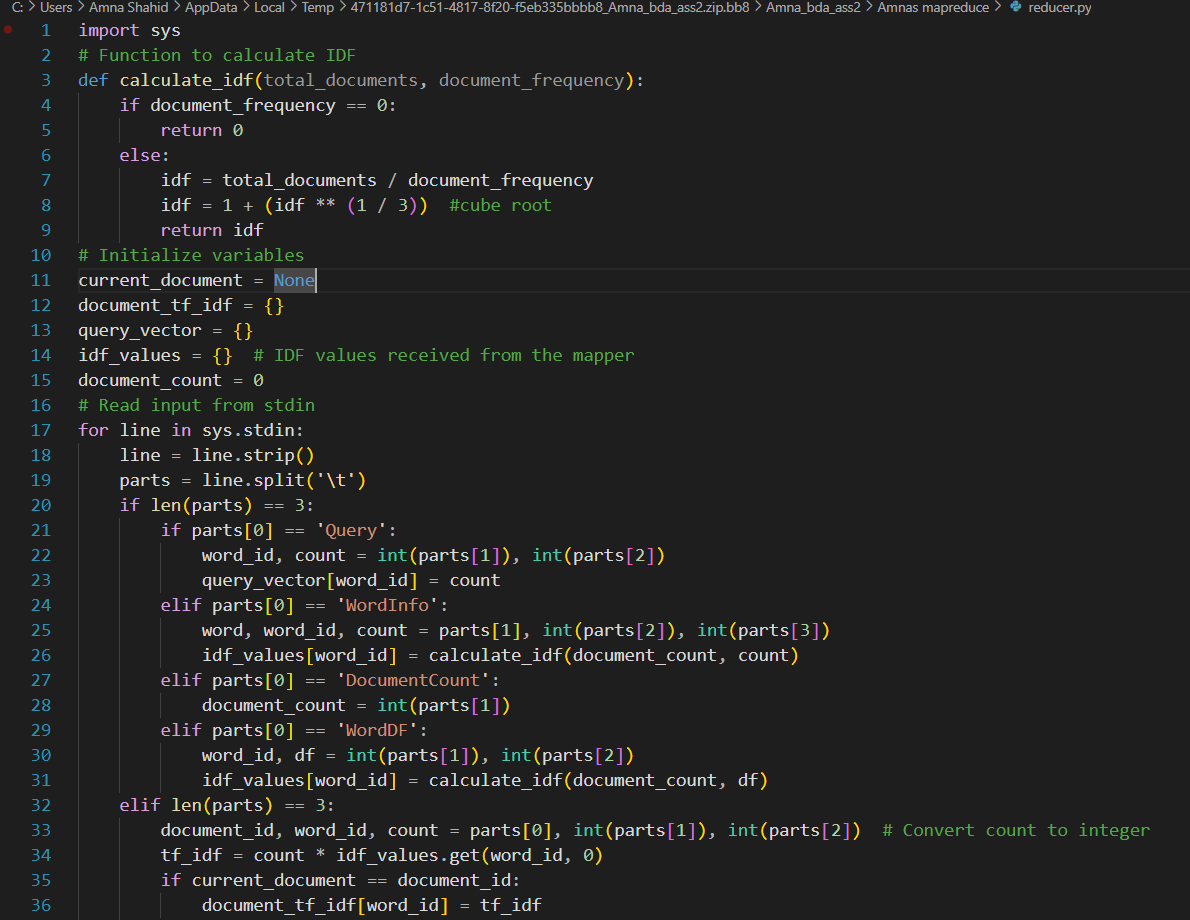


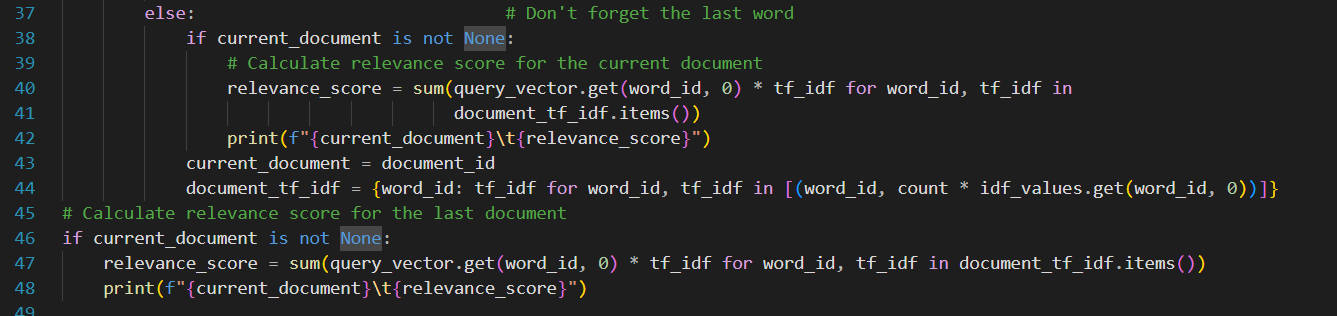




It reads a query (a question or a search term) from the input. Then it goes through each document's text and counts how many times each word appears. For each word, it assigns a unique ID and counts how many documents it appears in. Finally, it outputs information about words, documents, and their counts.

**Reducer.py:**





It calculates the importance of words in the query and documents. It receives IDF (Inverse Document Frequency) values calculated in the mapper. It goes through each document's words and computes TF-IDF (Term Frequency-Inverse Document Frequency) scores. It multiplies TF (how often a word appears in a document) with IDF (how rare the word is across all documents) to get TF-IDF.It calculates the relevance of each document to the query by summing up the TF-IDF scores of words in the query found in the document. Finally, it prints out the relevance score for each document. In simple terms, the mapper counts how often words appear in documents, while the reducer calculates how relevant each document is to a given query based on those word counts.

### PART III: QUESTION 6 and QUESTION 7

### **OVERVIEW AND LEARNING**

OVERVIEW:

These codes illustrate the MapReduce framework applied to the task of determining document relevance to a given query. In the initial phase, the mapper script receives the query input and converts it into a vectorized representation. Concurrently, it processes each document from the input, extracting words and assigning unique identifiers while tallying their occurrence across documents. Emitting data regarding words, documents, and their respective counts, the mapper lays the groundwork for subsequent analysis. Upon receiving this processed data, the reducer script begins its task by calculating the significance of words in both the query and documents, utilizing Inverse Document Frequency (IDF) values provided by the mapper. Through the computation of Term Frequency-Inverse Document Frequency (TF-IDF) scores, which weigh the frequency of word occurrence against its rarity across documents, the reducer determines the relevance of each document to the query. Summing the TF-IDF scores for words shared between the query and documents yields a measure of their relevance. This calculated relevance score is then printed out for each document. These codes offer an educational exploration into fundamental concepts such as the MapReduce paradigm, TF-IDF analysis, and document relevance assessment, all of which are integral to tasks in information retrieval and text processing.

THE END!