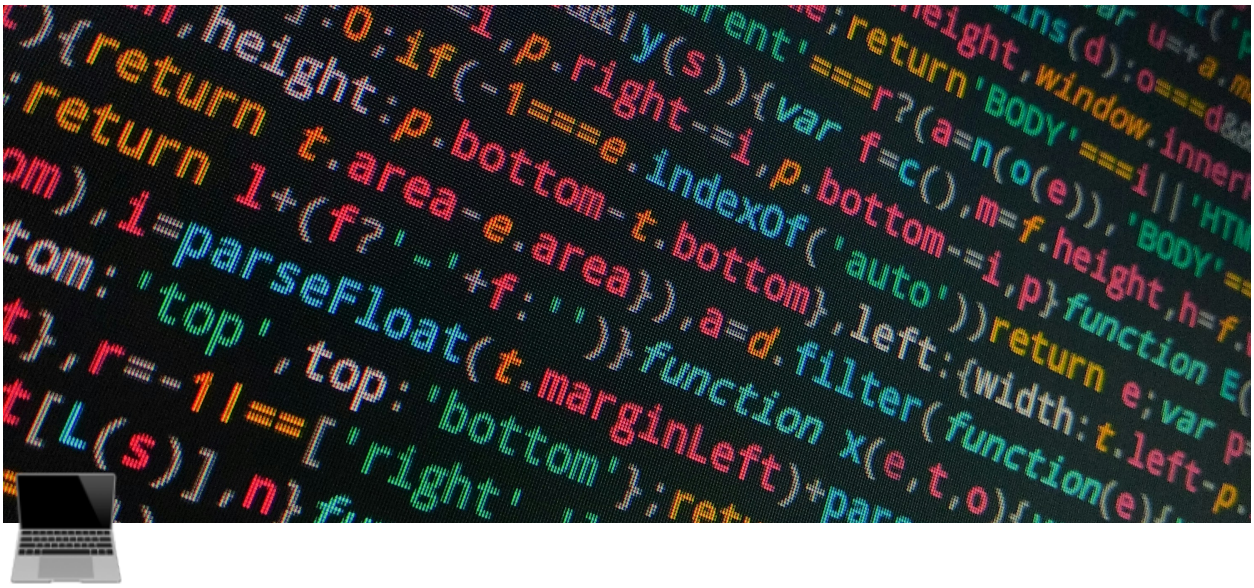

BDA Assignment #02

Created	@March 28, 2024 2:30 AM
Class	Big Data

Project Report: Text Mining and Analysis Pipeline

Date: March 31, 2024

Team Members:

- Muhammad Rayyan Mohsin 22i-2052
- Sufyan Nasr 22i-1898
- Haider Farooq 22i-1943

Overview:

Our project focuses on the analysis and extraction of insights from a massive dataset comprising textual information. Leveraging advanced techniques in text

mining and distributed computing, we aim to derive meaningful conclusions and valuable knowledge from the corpus.

Phase 1: Data Preprocessing and TF-IDF Vector Generation:

Significant progress has been made in the initial phase of our project, which involved meticulous data preprocessing using the Natural Language Toolkit (NLTK). This step was essential in ensuring that the text data was clean, standardized, and ready for further processing.

Following successful data preprocessing, we developed code for generating TF-IDF (Term Frequency-Inverse Document Frequency) vectors for the documents. TF-IDF is a fundamental technique in text mining, allowing us to quantify the importance of words within documents relative to the entire corpus.

Hadoop Clusters on Microsoft Azure VMs:

One of the key challenges we encountered was handling the sheer volume of data. To address this challenge, we leveraged the scalability and parallel processing capabilities of Hadoop, deploying local clusters on Microsoft Azure Virtual Machines (VMs). This strategic decision enabled us to distribute data processing tasks efficiently across multiple nodes, resulting in faster computations and efficient resource utilization.

Extraction and Cleaning of SECTION_TEXT and ARTICLE_ID Columns:

With the infrastructure in place, our focus shifted towards extracting pertinent information from our large dataset. Specifically, we targeted the SECTION_TEXT and ARTICLE_ID columns, recognizing their importance in subsequent analyses. Through a series of extraction and cleaning procedures, we meticulously curated the relevant data, ensuring accuracy and consistency throughout the process.

Input Preparation for Mapper and Reducer Files:

To facilitate further data processing within the Hadoop framework, we prepared the extracted and cleaned data for ingestion into mapper and reducer files. This involved formatting the data into a structured format compatible with Hadoop's MapReduce paradigm, effectively streamlining the subsequent stages of analysis.

Mapper and Reducer Files for Hadoop Implementation:

To accommodate our code for implementation on Hadoop clusters, we developed three pairs of mapper and reducer files tailored to our specific data processing tasks:

1. Mapper1.py / Reducer1.py: Initial processing of extracted data, including tokenization, lowercasing, and removal of stop words.
2. Mapper2.py / Reducer2.py: Calculation of inverse document frequency (IDF) values for each term across the entire corpus.
3. Mapper3.py / Reducer3.py: Combination of TF (term frequency) and IDF values to calculate TF-IDF scores for each term-document pair.

Conclusion:

In conclusion, our project has made significant strides in the analysis and extraction of insights from a vast corpus of textual information. Through meticulous data preprocessing, TF-IDF vector generation, and the deployment of Hadoop clusters, we have established a robust foundation for advanced analysis. Moving forward, we are poised to implement MapReduce tasks, leverage TF-IDF vectors for advanced analyses, and iteratively refine our methodologies to optimize performance and accuracy. Our ultimate goal is to document and present comprehensive findings to stakeholders, showcasing key insights and implications derived from our analysis. Through these efforts, we aim to deliver a robust and insightful analysis that contributes valuable knowledge and understanding to our domain of interest, paving the way for informed decision-making and further exploration in the field of text mining and analysis.