Project Big Data
Sentiments Analysis

Objective:
The objective of this project is to perform **Sentiment Analysis** using the MapReduce programming model. You will analyze a dataset of IMDB reviews and classify the sentiment (positive, negative, neutral) based on the presence of predefined keywords.

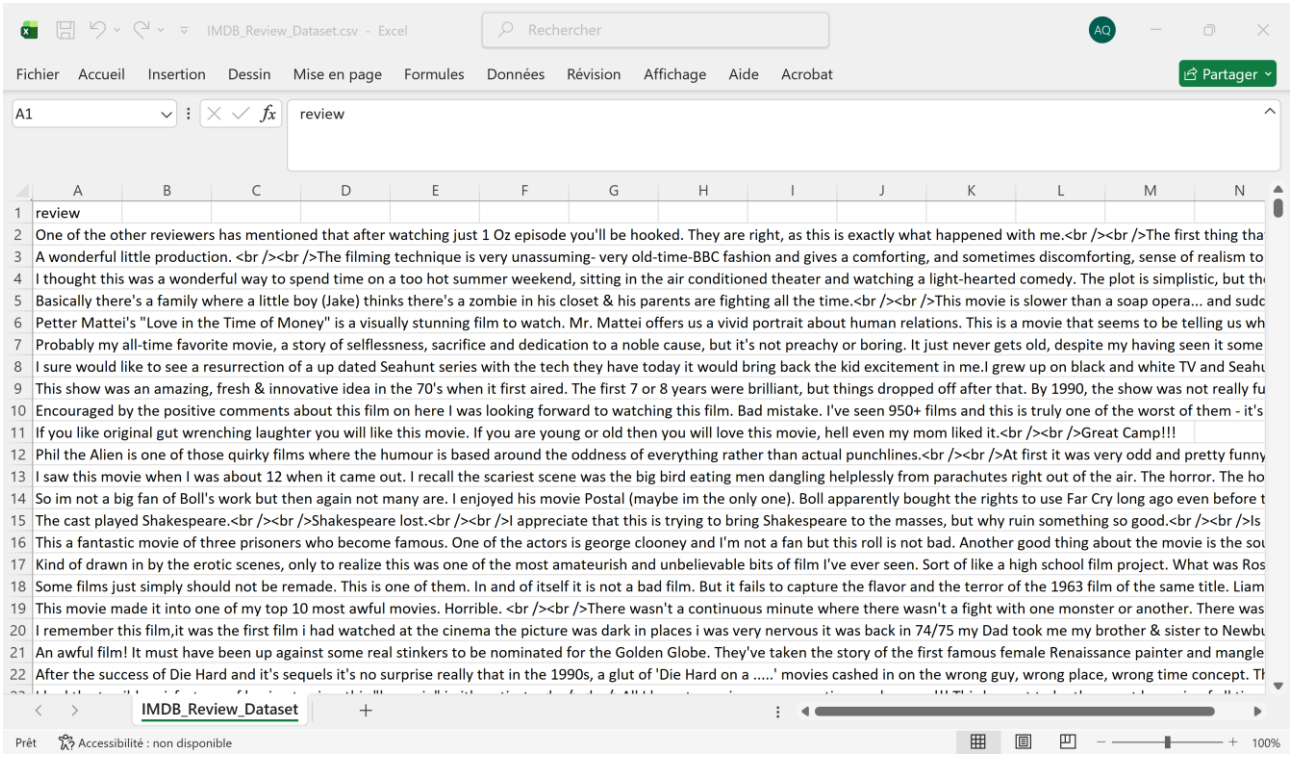
Prerequisites:

- Basic understanding of the MapReduce framework.
- Python installed (if using Hadoop streaming) or access to a Hadoop cluster.

Dataset Example

We will use the **IMDB_Review_Dataset.csv**, which contains customer reviews of various movies.

- Each line in the dataset represents a single review.
- Example



Tasks

1. Collect the Sentiment Keywords Dictionary

- **Positive Keywords**:
  Examples include *love, amazing, happy, good, excellent, wonderful, fantastic, brilliant, satisfied, success*.
- **Negative Keywords**:
  Examples include *horrible, frustrating, terrible, bad, worst, disappointing, awful, regret*.

2. Preprocess the Dataset

- Explore preprocessing techniques to clean the data:
  - Remove irrelevant information, such as **stop words** (e.g., *the, is, and, or*).
  - Strip special characters and punctuation to standardize the text.

3. Create the MapReduce Program

- Implement a program to process the dataset and classify reviews based on the sentiment dictionary.

4. Upload Dataset to HDFS

- Load the IMDB_Review_Dataset.csv file into the Hadoop Distributed File System (HDFS).

5. Run MapReduce Job

- Execute the MapReduce program to classify reviews as **Positive**, **Negative**, or **Neutral**.

6. Fetch Results

- Collect the sentiment analysis results from the HDFS output.

7. Display Results in a Graph

- Use visualization tools (e.g., Matplotlib, Excel, or Tableau) to create a bar chart or pie chart showing the count of positive, negative, and neutral reviews.

Analysis Questions

1. **Which sentiment dominates the dataset?**
   Analyze the visualized results to determine whether the majority of reviews are positive, negative, or neutral.
2. **What does this imply about the overall sentiment in the IMDB reviews?**
   Reflect on the broader implications of your findings (Consumer Perception, Platform Trustworthiness, Economic Insights).