ENSA4/FID2/24-25

المدرسة الوطنية للعلوم التطبيقية - أكادير
ⵜⵉⵎⵙⵙⵓⵜ ⵏ ⵜⵓⵙⵙⵉⵔⵉⵏ ⵜⵉⵖⵉⵔⵉⵜⵉⵏ - ⴰⴳⴰⴷⵉⵔ
ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES - AGADIR

Report
Big Data
2024/2025

Prepared by:

Hassan EL QADI, Achraf HAJJI, Aya HAMZI

IMDB Reviews Sentiment Analysis using Big Data

Supervised by:

M. Aimad QAZDAR

11/12/2024

1 Introduction :

The "IMDB Reviews Sentiment Analysis using Big Data" project leverages Hadoop's MapReduce programming model to process and classify sentiments in IMDB movie reviews. This analysis provides insights into public sentiment trends and consumer perceptions, facilitating better understanding of content reception. By utilizing distributed computing, the project efficiently processes large datasets and classifies sentiments as positive, negative, or neutral.
Key components of the project include:

- Development of a validated sentiment keywords dictionary.
- Implementation of advanced data preprocessing techniques.
- Scalable sentiment analysis using the Hadoop platform.

2 Objectives :

The project aims to achieve the following objectives:

- Utilize Hadoop's distributed processing capabilities for efficient analysis of large datasets.
- Develop a robust sentiment classification system based on validated sentiment keywords.
- Derive insights from sentiment trends in the IMDB reviews dataset.

3 Progress Overview :

3.1 Sentiment Keywords Dictionary :

The sentiment keywords dictionary was created using Python's Natural Language Toolkit (NLTK). Positive and negative keywords were expanded with synonyms and validated using the VADER sentiment analyzer. For example:

- **Positive Keywords**: *brilliant, fantastic, wonderful, happy, love, success.*
- **Negative Keywords**: *terrible, awful, frustrating, regret, worst.*

Validated keywords were saved in a JSON file for future use.

3.2 Text Preprocessing :

Text preprocessing involved:

- Expanding contractions (e.g., "don't" to "do not") using the `contractions` library.
- Converting text to lowercase.
- Removing HTML tags and special characters.

- Tokenizing text into individual words.
- Removing stop words using NLTK's predefined list.
- Lemmatizing words to reduce them to their base forms (e.g., "running" to "run").

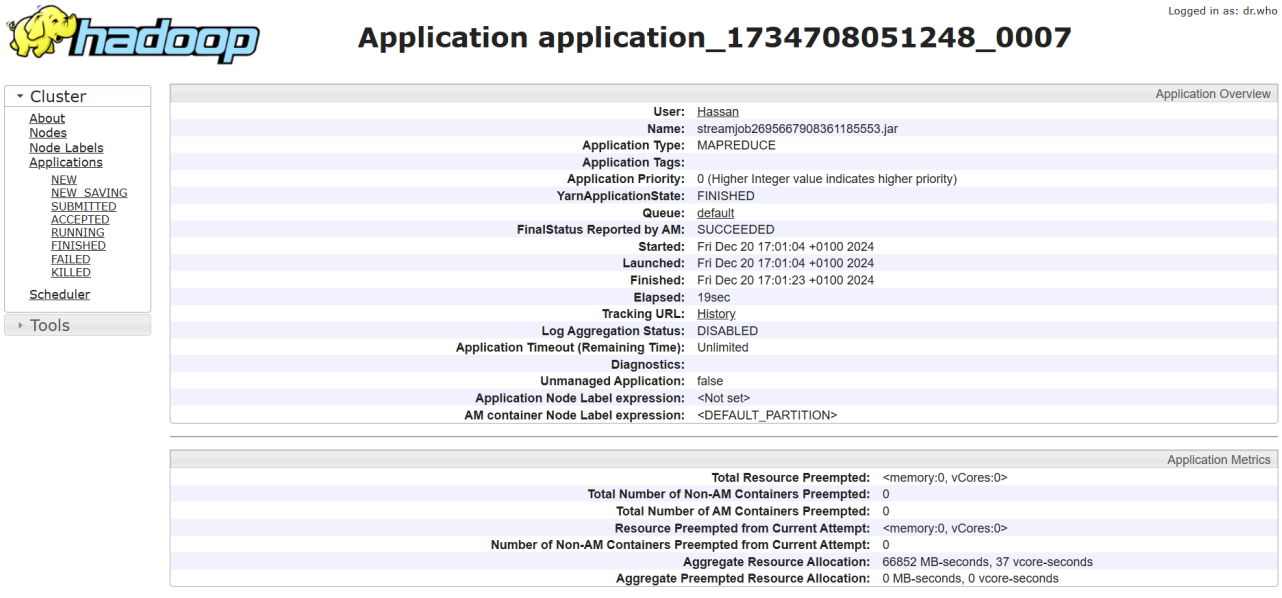
3.3 Sentiment Analysis with MapReduce :

The sentiment analysis process included:

- Uploading the cleaned dataset to the Hadoop Distributed File System (HDFS).
- Executing a MapReduce job with Python scripts for mapping and reducing. The command used:

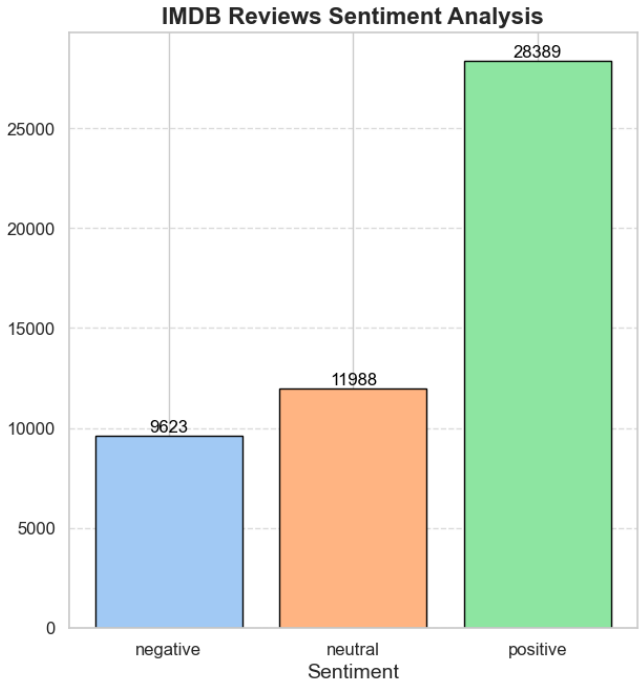
```
hadoop jar C:/hadoop-3.2.2/share/hadoop/tools/lib/hadoop-streaming-3.2.2.jar \
    -file mapper.py \
    -file reducer.py \
    -mapper "python mapper.py" \
    -reducer "python reducer.py" \
    -input /user/hassan/imdb_reviews/Cleaned_IMDB_Reviews.csv \
    -output /user/hassan/imdb_sentiment_output
```

- Mapper script identified keywords to classify sentiments as positive, negative, or neutral.
- Reducer script aggregated results to determine sentiment distribution.



3.4 Results and Visualization :

The sentiment analysis results are as follows:



4 Analysis Questions :

4.1 Which sentiment dominates the dataset? :

The analysis reveals that **positive sentiments dominate** the dataset, with the majority of reviews classified as positive. This suggests that most users have favorable opinions about the movies they reviewed.

4.2 What does this imply about the overall sentiment in the IMDB reviews? :

The findings imply a generally positive outlook among IMDB users regarding movies. This reflects positively on the platform's trustworthiness, as a large number of positive reviews can enhance user confidence. From an economic perspective, such sentiment trends may indicate higher engagement levels and potentially better box office performance for movies with overwhelmingly positive feedback.

5 Conclusion :

This project successfully demonstrated the use of Hadoop's MapReduce model for scalable sentiment analysis of large datasets. The results reveal a predominantly positive sentiment among IMDB reviews, highlighting the importance of consumer perception in shaping the platform's reputation. Future work could involve:

- Implementing machine learning models for sentiment analysis to enhance accuracy.
- Expanding the dataset to include reviews from other platforms.

- Analyzing temporal trends in sentiment for predictive insights.

The source code and detailed implementation are available in the **GitHub Repository**.