Netflix Data Cleaning, MapReduce Processing, and Analysis Report

Pegah Moradpour

Introduction

This report describes the process I followed to clean the Netflix dataset, apply a MapReduce job to analyze the genre distribution by country, and interpret the results. The primary hypothesis I aim to investigate is whether Italian Netflix content has a higher proportion of "Drama" compared to global Netflix content.

Data Cleaning

Objective

The purpose of data cleaning is to prepare the Netflix dataset for analysis by removing incomplete records, standardizing formats, and ensuring data quality.

Steps

1. **Load the Data**: I loaded the Netflix dataset from HDFS, specifying a schema to ensure data types were properly defined for each column.
2. **Drop Missing Values**: I used the `dropna()` method to remove any rows containing `NULL` or missing values across all columns. This step ensures that the analysis is based on complete data.
3. **Trim Whitespace**: I standardized the `country` field by trimming any extra whitespace around country names using `F.trim()`.
4. **Filter for Relevant Data**: I filtered the dataset to only include titles released from the year 2000 onwards by casting the `release_year` column to an integer and applying a filter (`F.col("release_year") >= 2000`).
5. **Save the Cleaned Data**: The cleaned dataset was saved back to HDFS in a specified directory to be used for subsequent analysis and MapReduce processing.

Result

The cleaned Netflix dataset is now free of incomplete entries and standardized for further processing. This cleaning helps ensure that the MapReduce job and analysis are based on reliable data.

MapReduce Job

Objective

The purpose of the MapReduce job is to count the number of titles available in different genres for each country, focusing on Italy and global counts, to determine the genre distribution.

Mapper Code Explanation

The Mapper reads each line of the Netflix data and extracts relevant fields: `genre` and `country`. If the `country` field is "Italy," it is labeled as "Italy"; otherwise, it is labeled as "Global" to represent all other countries collectively. For each genre-country combination, the Mapper outputs a key-value pair where the key is the genre-country and the value is `1`.

Example Output:

- Input: `"Drama\tItaly\t1"`
- Output Key-Value Pair: `("Drama\tItaly", 1)`

Reducer Code Explanation

The Reducer takes the Mapper's output and aggregates the counts for each unique genre-country combination. It sums up the values associated with each key to get the total count of titles for each genre by country.

Example Output:

- Input: `("Drama\tItaly", 1), ("Drama\tItaly", 1), ("Comedy\tGlobal", 1)`
- Output: `("Drama\tItaly", 2), ("Comedy\tGlobal", 1)`

Result

The MapReduce job produces a list of genres with their respective counts for Italy and globally. This output is saved in HDFS for further analysis.

Data Analysis

Objective

My hypothesis is to determine if Italian Netflix content has a higher proportion of "Drama" titles than global Netflix content. I use the output from the MapReduce job to calculate the drama proportions for Italy and globally and perform a statistical test to evaluate the hypothesis.

Steps

1. **Load MapReduce Output**: I loaded the MapReduce output from HDFS into a DataFrame in my analysis environment (Jupyter Notebook).
2. **Calculate Genre Counts**:
    - **Total Italian Content**: Sum of all titles for each genre where the country is "Italy."

    - **Total Global Content**: Sum of all titles for each genre where the country is "Global."
3. **Calculate Drama Percentage**:
    - **Italian Drama Percentage**: `(Italian Drama Count / Total Italian Content Count) * 100`
    - **Global Drama Percentage**: `(Global Drama Count / Total Global Content Count) * 100`
4. **Statistical Analysis**:
    - **Chi-Square Test**: I used a Chi-Square test to assess whether the difference in proportions of "Drama" between Italian and global content is statistically significant.
5. **Visualization**: I plotted a bar chart showing the drama percentages for Italian vs. global Netflix content.

Result

The output of the analysis showed:

- Italian Drama Percentage: 21.37%
- Global Drama Percentage: 19.22%

The Chi-Square test yielded a p-value greater than 0.05, suggesting that the difference in drama proportions is not statistically significant. Therefore, I do not have strong statistical evidence to conclude that Italian Netflix content has a higher proportion of drama titles compared to global content.

Conclusion

Summary of Findings

1. **Data Cleaning**: The Netflix dataset was thoroughly cleaned, removing missing values and standardizing formats.
2. **MapReduce Processing**: I successfully counted the number of titles by genre for Italy and globally.
3. **Analysis**: While the Italian Drama percentage is slightly higher than the global percentage, statistical analysis shows this difference is not significant.

Limitations and Future Work

1. **Dataset Scope**: The dataset only covers titles up to a certain point in time. Analyzing an updated dataset could yield more accurate results.
2. **Genre Overlaps**: Some titles may belong to multiple genres, potentially influencing the counts. Future analysis could account for this.
3. **Alternative Hypotheses**: Further analysis could explore other genre preferences or trends within specific time frames.

In conclusion, while my analysis shows a higher percentage of drama content in Italian Netflix titles, it does not statistically support the hypothesis. Further data and refinement could help to achieve more robust insights.