TRANSPORT AND TELECOMMUNICATION INSTITUTE

BIG DATA

COURSEWORK

Student: Sergejs Kopils

Student Code: St83519

RIGA 2024

Contects

Introduction

In today's data-driven world, the ability to extract meaningful insights from vast amounts of information has become a cornerstone of decision-making across industries. Big Data Analytics plays a crucial role in uncovering patterns, trends, and relationships within structured and unstructured datasets, empowering organizations to make informed strategic choices. This project focuses on leveraging Big Data methodologies to analyze customer transaction trends from a structured dataset and sentiment patterns from an unstructured dataset of customer reviews.

The structured dataset contains detailed records of customer transactions, including transaction amounts, customer demographics, and product categories. This data provides an opportunity to explore spending behaviors, identify key purchase patterns, and analyze seasonal trends. On the other hand, the unstructured dataset consists of customer reviews of musical instruments, offering rich textual data to assess customer sentiment, review helpfulness, and yearly review trends.

The objective of this project is twofold:

1. To analyze customer spending habits and identify key insights from structured transaction data.
2. To perform sentiment analysis on customer reviews, evaluating overall satisfaction levels and identifying trends in helpfulness scores.

The analysis is performed using Python in Google Colab, with tools such as Pandas, NumPy, Seaborn, TextBlob, and rdflib. MongoDB Atlas is used for data storage and retrieval, while RDF ontology modeling helps represent semantic relationships between customers, products, and reviews.

This report will detail the methodologies employed in data exploration, preprocessing, analysis, visualization, and ontology creation, followed by a discussion of the results and their implications. Through this study, we aim to highlight the power of Big Data Analytics in extracting actionable insights from diverse datasets, contributing to data-driven decision-making processes.

Background and Context

The rapid growth of data in recent years has led to the emergence of Big Data as a critical component of modern analytics. Organizations across various sectors, including retail, healthcare, and finance, are leveraging Big Data technologies to gain valuable insights and maintain a competitive edge. Big Data refers to datasets that are too large, complex, or fast-changing to be processed effectively using traditional data management tools.

The concept of the "4Vs" of Big Data—Volume, Velocity, Variety, and Veracity—highlights the challenges associated with managing and analyzing such datasets. Volume refers to the sheer size of data, Velocity represents the speed at which data is generated and processed, Variety indicates the diverse data formats (structured, unstructured, semi-structured), and Veracity focuses on the quality and reliability of data.

In this project, the structured dataset provides transactional data, which is inherently numerical and categorical, making it suitable for statistical analysis and visualization. Conversely, the unstructured dataset consists of customer reviews in textual format, requiring natural language processing techniques to extract meaningful insights.

Modern tools and platforms, such as Google Colab and MongoDB Atlas, play a crucial role in managing and analyzing Big Data. Google Colab provides a cloud-based Python development environment, enabling collaborative coding and access to powerful computing resources. MongoDB Atlas serves as a scalable and flexible NoSQL database, ideal for storing and querying large datasets.

The integration of semantic web technologies, such as RDF and ontology modeling, further enhances the ability to represent relationships between different data entities. This approach allows for a more intuitive understanding of data connections and facilitates advanced querying and reasoning capabilities.

This section sets the stage for the methodologies employed in this study, highlighting the importance of Big Data technologies and the tools used to process and analyze both structured and unstructured datasets.

Methodology

The methodology adopted in this project follows a structured approach to handle both structured and unstructured datasets, ensuring consistency, reproducibility, and accuracy in analysis. The following steps outline the key stages of the methodology:

1. Dataset Selection

Two datasets were selected to represent structured and unstructured data:

- **Structured Dataset:** A CSV dataset containing customer transaction records, including fields such as Customer ID, Transaction Amount, Category, and Date.
- **Unstructured Dataset:** A JSON dataset containing customer reviews of musical instruments, including fields such as reviewerID, reviewText, overall rating, and helpfulness score.

2. Environment Setup

The analysis was performed in a cloud-based environment using **Google Colab**, ensuring access to Python libraries and scalable computational resources. Essential libraries such as Pandas, NumPy, Seaborn, TextBlob, and rdflib were installed and configured.

3. Data Exploration and Cleaning

- **Structured Dataset:**
    - Identified and handled missing values in fields such as Gender.
    - Converted date fields to a standardized datetime format.
    - Removed inconsistencies and ensured data integrity.

- **Unstructured Dataset:**
    - Handled missing values in reviewer names.
    - Standardized date formats.
    - Preprocessed text data for sentiment analysis.

4. Data Processing and Transformation

- **Structured Dataset:**
    - Feature engineering: Extracted age from birthdates and transaction year, month, and day from date fields.
    - Normalization: Scaled transaction amounts for consistency.
- **Unstructured Dataset:**
    - Sentiment analysis using TextBlob to extract sentiment polarity from review text.
    - Helpfulness score calculated as a ratio of helpful votes to total votes.

5. Data Analysis

- **Structured Dataset:**
    - Analyzed spending patterns across categories, genders, and age groups.
    - Identified monthly spending trends.
- **Unstructured Dataset:**
    - Examined sentiment distribution.
    - Analyzed helpfulness score trends.
    - Identified review patterns over time.

6. Ontology Creation

An RDF ontology was created to model semantic relationships between entities:

- **Customer:** Represented as FOAF.Person with attributes such as age and gender.
- **Review:** Linked to customers and products with properties such as sentiment and rating.
- **Product:** Connected via reviews and purchases.

Conclusion and Future Work

This project successfully demonstrated the application of Big Data Analytics techniques to analyze both structured and unstructured datasets. Key insights were derived from customer spending trends, sentiment analysis, and review helpfulness scores. Future work could focus on integrating advanced machine learning models for predictive analytics and enhancing scalability through more robust cloud infrastructure. Additionally, expanding the ontology to include more complex relationships could further improve data interpretation and analysis.

References and Citations

1. MongoDB Atlas Documentation: https://www.mongodb.com/docs/
2. Google Colab Documentation: https://colab.research.google.com/
3. Kaggle Datasets: https://www.kaggle.com/
4. rdflib Documentation: https://rdflib.readthedocs.io/en/stable/
5. TextBlob Documentation: https://textblob.readthedocs.io/en/dev/
6. Seaborn Documentation: https://seaborn.pydata.org/