NAME: Tarani Chilamkoti

NJIT UCID: tc533

Email Address: tc533@njit.edu

Data: 28th feb 2025

Professor: Yasser Abduallah

SP25-CS634854 Data Mining

**Midterm Project Report**

*Implementation and Code Usage*

**Apriori Algorithm Implementation in Retail Data Mining**

**Abstract:**

This project investigates frequent itemset mining and association rule building using both brute-force methods and the Apriori algorithm. The idea is to analyze retail transaction data to discover significant relationships between items. The performance of the two approaches is compared using execution time and output similarity.

**Introduction:**

Hidden patterns in large datasets can be discovered using data mining techniques. The Apriori algorithm is a popular method for association rule mining. In this project, we implemented:A brute-force approach for frequent item set mining.The Apriori algorithm is implemented using the MLX tend Python library.Performance analysis and comparison of the two approaches.Datasets are manually built and saved as CSV files including transactional data from various retail establishments.

**Core Concepts and Principles:**

**Frequent Itemset Discovery**Frequent itemset discovery entails recognizing groups of items that frequently appear together in transactions. This allows firms to better identify purchasing habits and optimize inventory or marketing efforts.

**Project Workflow:**

Data Loading and Preprocessing:

Data is loaded from CSV files of various retail retailers. Each dataset contains transactions where numerous goods were purchased together.

The preprocessing phase guarantees that:

Duplicate transactions were removed.

Formatting consistency.

If there are any missing values, they will be handled.

**Determination of Minimum Support and Confidence:**

Data is loaded from CSV files of various retail retailers. Each dataset contains transactions where numerous goods were purchased together.

The preprocessing phase guarantees

Users specify:

Minimum support: The frequency threshold at which an itemset is considered frequent.

Minimum confidence: refers to the strength of the link between itemsets.

These settings have a direct impact on the number and quality of the rules generated.

**Iteration Through Candidate Item sets:**

Data is loaded from CSV files of various retail retailers. Each dataset contains transactions where numerous goods were purchased together.

The preprocessing phase guarantees

Both the brute-force and Apriori methods iterate over possible item sets:

Creating one-item sets, then growing to larger sets.

Filtering itemsets by support values.

Iterate until no more frequent itemsets are identified.

**Support Count Calculation:**

Support is calculated as: Support(x)= number of transactions containing X / total transactions

This determines how frequently an itemset appears within transactions.

**Confidence Calculation:**

Confidence is calculated as:  Confidence(X=>Y) = support (X U Y) / support(X)

It measures the likelihood that if a customer buys item X, they will also buy item Y.

**Association Rule Generation:**

Association rules are created from frequent itemsets where confidence exceeds the user-defined threshold. These rules reveal valuable insights, such as "Customers who buy laptops often buy external hard drives."

**Results and Evaluation:**

The brute-force approach successfully builds frequent itemsets but is computationally expensive.

Apriori algorithm generates the same frequent item sets in substantially less time.

Performance comparison: The Apriori algorithm is significantly faster due to its pruning method.

**Conclusion:**

This project demonstrates the application of frequent itemset mining using two approaches. The Apriori algorithm significantly outperforms the brute-force method in execution time while maintaining accuracy. These techniques have practical applications in retail analytics and recommendation systems.

*Screenshots*

Here are what the csv files contain

Figure 1: Nike Item Names CSV file

Figure 2: K Mart Items Names CSV file

Figure 3: Amazon Item Names CSV files

Figure 4: Best Buys Item Names CVS files

Figure 5: Generic Items Names CSV files

A screenshot of a bedding list

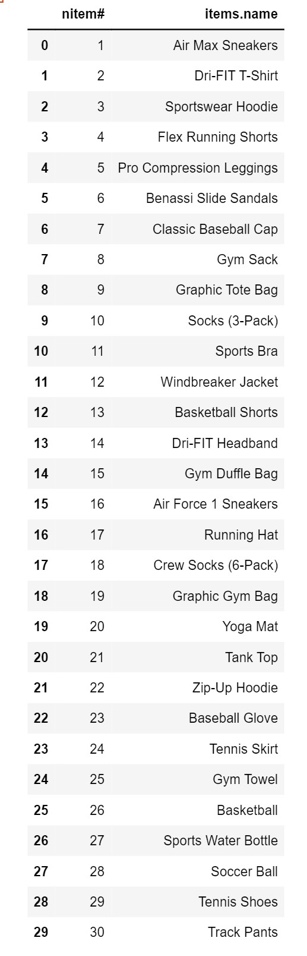
AI-generated content may be incorrect.

Fig.1 Fig.2

A screenshot of a computer

AI-generated content may be incorrect.

Fig. 3

A screenshot of a computer

AI-generated content may be incorrect.

Fig. 4

A screenshot of a cell phone

AI-generated content may be incorrect.

Fig.5

Below are screenshots of the python code:

To Load the Datasets

A screenshot of a computer

AI-generated content may be incorrect.

Checking if the file is existing:

A screenshot of a computer

AI-generated content may be incorrect.

Loading the datasets:

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

Apriori rule:

A screen shot of a computer

AI-generated content may be incorrect.

Generating Association rule:

A screen shot of a computer

AI-generated content may be incorrect.

Final Output:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

***Other***

The source code (.py file) and data sets (.csv files) will be attached to the zip