Mining Massive Data Sets Midterm Report

1st 522H0036 - Luong Canh Phong
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0036@student.tdtu.edu.com

2nd 522H0092 - Cao Nguyen Thai Thuan
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0092@student.tdtu.edu.com

3rd 522H0075 - Tang Minh Thien An
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0075@student.tdtu.edu.com

4th 522H0167 - Truong Tri Phong
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0167@student.tdtu.edu.com

5th Instructor: Nguyen Thanh An
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
nguyenthanhan@tdtu.edu.com

***Abstract*—This project implements and evaluates key techniques in mining massive datasets. It covers hierarchical agglomerative clustering of string shingles using Jaccard distance; PySpark-based linear regression for gold price prediction; CUR decomposition for feature dimensionality reduction on gold price data; and PageRank analysis of the `it.tdtu.edu.vn` web graph using PySpark. The work demonstrates practical applications and provides insights into processing large-scale data.**

I. Introduction

The increasing volume of data requires efficient mining techniques. This project implements and analyzes four core algorithms: (1) hierarchical agglomerative clustering for non-Euclidean text data, using 4-shingles and Jaccard distance on alphabetical strings; (2) PySpark-based linear regression to predict Vietnamese gold prices from historical data; (3) CUR decomposition to reduce the dimensionality of gold price features (from 10 to 5) and assess its impact on regression; and (4) PageRank, implemented in PySpark, to identify influential pages within the `it.tdtu.edu.vn` web graph. Python and PySpark are utilized throughout. This report details the methodologies, implementations, and experimental results for each task.

II. First Task: Hierarchical clustering in non-Euclidean spaces

III. Second Task: Linear Regression – Gold price prediction

IV. Contribution

The following table represents the contribution of each member, note that whichever member handles whichever task will also write the report for that task.

TABLE I
Member Contributions

ID	Member	Contribution	Progress
522H0036	Luong Canh Phong	Task 1 and Handling Report	100%
522H0092	Cao Nguyen Thai Thuan	Overseer and Report Support	100%
522H0075	Tang Minh Thien An	Task 3	100%
522H0167	Truong Tri Phong	Task 2	100%

V. Self-evaluation

The following table is our self-evaluation on our tasks:

TABLE II
Self-evaluation

Task	Task Requirements	Completion Ratio
Task 1	A-Priori Algorithm for Frequent Customers	100%
Task 2	PCY Algorithm for Frequent Items	95%
Task 3	MinHashLSH for Similar Dates	90%
Task 4	Report	100%

VI. Conclusion

References