**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

**VIỆN KỸ THUẬT CÔNG NGHỆ**

****

**BÁO CÁO KẾT THÚC HỌC PHẦN  
THỰC HÀNH CHUYÊN ĐỀ DỮ LIỆU LỚN**

**BÁO CÁO TỔNG KẾT THỰC HÀNH**

**Sinh viên thực hiện:** Nguyễn Hữu Nghĩa

**MSSV:** 2124802050013

**Lớp:** D21TTNT01

**GVHD:** ThS. Nguyễn Thế Bảo

Học kỳ I năm học 2024 – 2025

10/2024

**BÌNH DƯƠNG - 2024**

**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

**VIỆN KỸ THUẬT CÔNG NGHỆ**

****

**BÁO CÁO KẾT THÚC HỌC PHẦN  
THỰC HÀNH CHUYÊN ĐỀ DỮ LIỆU LỚN**

**BÁO CÁO TỔNG KẾT THỰC HÀNH**

**Sinh viên thực hiện:** Nguyễn Hữu Nghĩa

**MSSV:** 2124802050013

**Lớp:** D21TTNT01

**GVHD:** ThS. Nguyễn Thế Bảo

Học kỳ I năm học 2024 – 2025

10/2024

**BÌNH DƯƠNG - 2024**

MỤC LỤC

[DANH MỤC CÁC HÌNH 2](#_Toc179033081)

[CHƯƠNG 1. CÀI ĐẶT HADOOP - SPARK TRÊN WINDOWS 2](#_Toc179033082)

[1.1. Chuẩn bị môi trường 2](#_Toc179033083)

[1.2. Cài đặt Java 2](#_Toc179033084)

[1.3. Cài đặt python 2](#_Toc179033085)

[1.4. Cài đặt Apache Spark 3](#_Toc179033086)

[1.5. Cài đặt Hadoop 4](#_Toc179033087)

[1.6. Thiết lập biến môi trường 5](#_Toc179033088)

[1.7. Kiểm tra và khởi động Apache Spark 6](#_Toc179033089)

[1.8. Khởi động SPARK Master 8](#_Toc179033090)

[1.9. Khởi động SPARK Worker 9](#_Toc179033091)

[1.10. Kết nối Shell với Spark Master 9](#_Toc179033092)

[CHƯƠNG 2. LAUNCHING APPLICATION ON HADOOP – SPARK WINDOWS 11](#_Toc179033093)

[2.1. Khởi động hệ thống: Hadoop, Spark. 11](#_Toc179033094)

[CHƯƠNG 3. XỬ LÝ DỮ LIỆU VĂN BẢN VỚI PYSPARK 13](#_Toc179033095)

[3.1. Mô tả dữ liệu 13](#_Toc179033099)

[3.2. Bài tập xử lý dữ liệu: 14](#_Toc179033100)

[CHƯƠNG 4. ỨNG DỤNG SPARK ĐỂ TRIỂN KHAI CÁC MÔ HÌNH HỌC MÁY (MACHINE LEARNING) PHÂN TÁN 15](#_Toc179033101)

[4.1 Kmeans\_pyspark 15](#_Toc179033102)

[4.2 Kmeans\_scikit-learn 16](#_Toc179033103)

[4.3 So sánh tâm cụm 16](#_Toc179033104)

[4.5 Nhận xét 17](#_Toc179033105)

[CHƯƠNG 5. PHÂN TÍCH MẠNG XÃ HỘI 18](#_Toc179033106)

[TÀI LIỆU THAM KHẢO 23](#_Toc179033107)

# DANH MỤC HÌNH

[Image 1 Cài đặt java 2](#_Toc179026106)

[Image 2 Install Python 3](#_Toc179026107)

[Image 3 Download Apache Spark 4](#_Toc179026108)

[Image 4 Giải nén Apache Spark 4](#_Toc179026109)

[Image 5 Thiết lập Hadoop 5](#_Toc179026110)

[Image 6 Thiết lập biến môi trường 5](#_Toc179026111)

[Image 7 Thêm spark và hadoop vào system path 6](#_Toc179026112)

[Image 8 Kiểm tra lại sau khi cài đặt spark, hadoop, java, python 7](#_Toc179026113)

[Image 9 Khởi động Spark-shell 7](#_Toc179026114)

[Image 10 Khởi động Pyspark 8](#_Toc179026115)

[Image 11 Khởi động SPARK Master 8](#_Toc179026116)

[Image 12 Kết quả trên web localhost 8](#_Toc179026117)

[Image 13 Khởi động SPARK Worker 9](#_Toc179026118)

[Image 14 Kết quả trên web localhost 9](#_Toc179026119)

[Image 15 Kết nối Shell với Spark Master 9](#_Toc179026120)

[Image 16 Kết quả trên web localhost:8080 10](#_Toc179026121)

[Image 17 Khởi động hệ thống: Hadoop, Spark. 11](#_Toc179026122)

[Image 18 Submit wordcount\_demo.py và kết quả. 11](#_Toc179026123)

[Image 19 Submit als.py. 12](#_Toc179026124)

[Image 20 Xử lý dữ liệu văn bản với pyspark 13](#_Toc179026125)

[Image 21 Xử lý dữ liệu văn bản với pyspark 13](#_Toc179026126)

[Image 22 Liệt kê 20 từ được nhắc nhiều nhất 14](#_Toc179026127)

[Image 23 Liệt kê 5 tài khoản được nhắc nhiều nhất 14](#_Toc179026128)

[Image 24 Kết quả Kmeans-pyspark 15](#_Toc179026129)

[Image 25 Kết quả Kmeans-sklearn 16](#_Toc179026130)

[Image 26 Cài đặt thư viện networkx 18](#_Toc179026131)

[Image 27 Kết quả phương pháp truyền thống 20](#_Toc179026132)

[Image 28 Kết quả phương pháp dữ liệu lớn 21](#_Toc179026133)

1. CÀI ĐẶT HADOOP - SPARK TRÊN WINDOWS

Chuẩn bị môi trường

Để làm việc với PySpark trên Windows cần chuẩn bị các môi trường sau:

- Hệ điều hành Windows: Đảm bảo bạn đang sử dụng một phiên bản Windows

hỗ trợ cài đặt và chạy PySpark.

- Cài đặt Java 8s: Tải Java JDK 8 từ trang chính thức của Oracle hoặc từ một

nguồn uy tín khác.

- Cài đặt Python 3: Tải Python 3 từ trang chính thức của Python.

Cài đặt Java

- Truy cập trang web chính thức của Oracle hoặc OpenJDK để tải JDK. Chạy jre-8u202-windows-i586.exe để tiến hành cài đặt. Sau khi cài đặt xong chạy lệnh để kiểm tra: java -version.

- Hoặc cài java qua Command Prompt: Nhấn Windows + R, gõ cmd, rồi nhấn Enter.

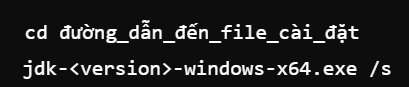


Image 1 Cài đặt java

Cài đặt python

Truy cập đường dẫn https://www.python.org/downloads/release/python-383/ và tải

tập tin **Windows x86 executable installer**



Image 2 Install Python

Hoặc mở cmd nhập lệnh:

curl -O https://www.python.org/ftp/python/3.x.x/python-3.x.x-amd64.exe

Cài đặt Apache Spark

* Truy cập đường dẫn <https://spark.apache.org/downloads.html>
* Chọn phiên bản Spark: 3.5.2 (có thể chọn phiên bản khác)
* Chọn gói Pre-built for Apache Hadoop 2.7 để nhận bản build sẵn
* Nhấp vào liên kết spark-3.5.2-bin-hadoop2.7.tgz để tải về

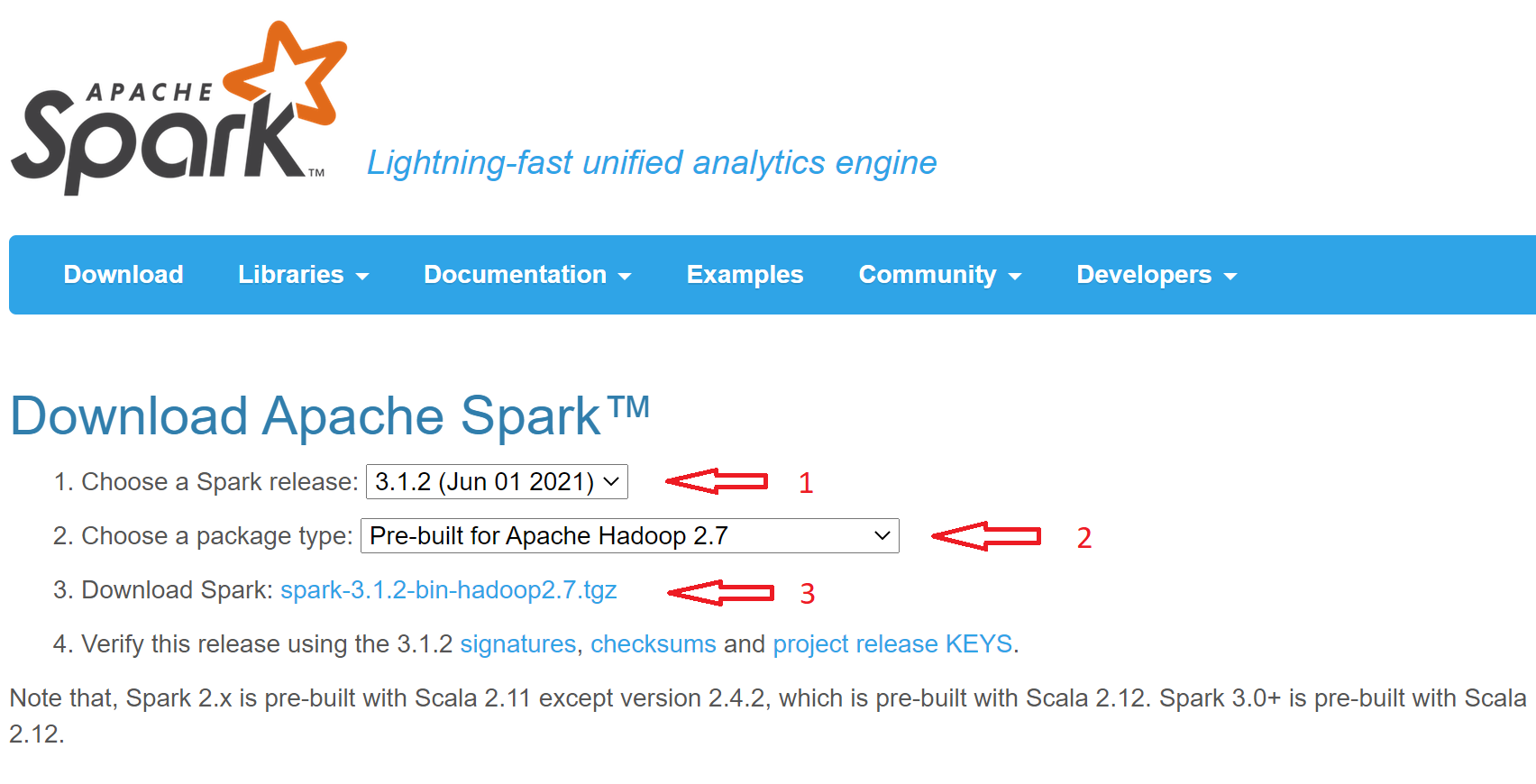


Image 3 Download Apache Spark

* Giải nén và copy vào folder **C:\spark\spark-3.5.2-bin-hadoop2.7**

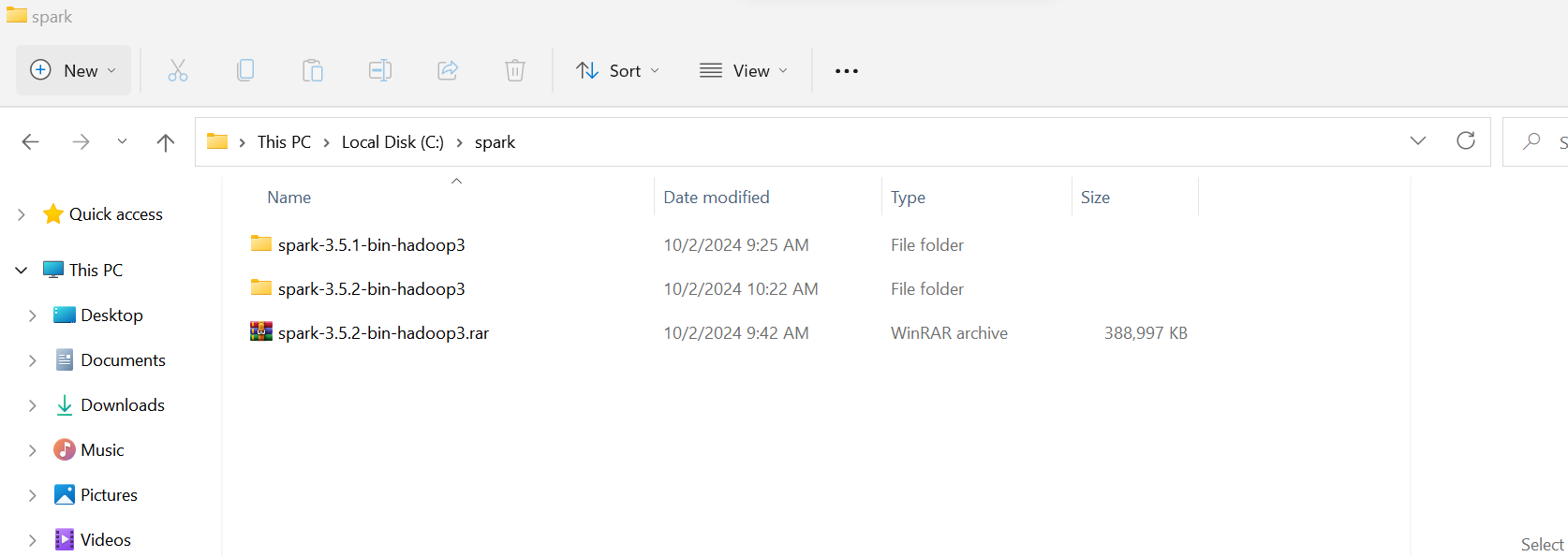


Image 4: Giải nén apache spark

Cài đặt Hadoop

* Tạo folder chứa hadoop: **C:/hadoop/**
* Truy cập đường dẫn và tải về <https://github.com/cdarlint/winutils/blob/master/>
* Giải nén tập tin **winutils-master.zip**
* Chọn Hadoop có version tương ứng với Spark bên trên
* Copy vào **C:/hadoop/**

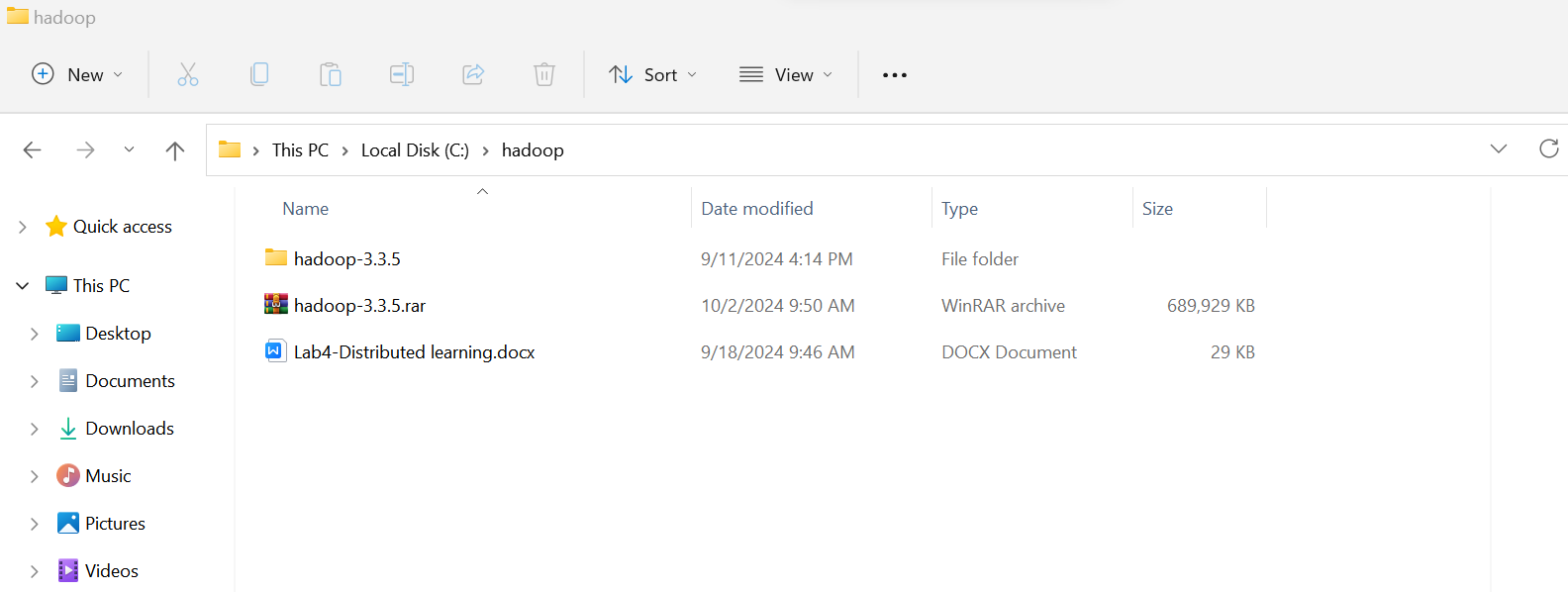


Image 5: Thiết lập Hadoop

Thiết lập biến môi trường

* **SPARK\_HOME**=C:\spark\spark-3.5.2-bin-hadoop3
* **HADOOP\_HOME**=C:\hadoop\hadoop-2.7.7

A screenshot of a computer

Description automatically generated

Image 6: Thiết lập biến môi trường

* Thêm Spark và Hadoop vào system PATH

A screenshot of a computer program

Description automatically generated

Image 7: Thêm spark và hadoop vào system path

Kiểm tra và khởi động Apache Spark

A screenshot of a computer program

Description automatically generated

Image 8: Kiểm tra lại sau khi cài đặt spark, hadoop, java, python

A screenshot of a computer program

Description automatically generated

Image 9 Khởi động Spark-shell

A screenshot of a computer program

Description automatically generated

Image 10 Khởi động Pyspark

Khởi động SPARK Master

* Mở 1 cửa sổ Command Prompt mới, chạy lệnh: spark-class org.apache.spark.deploy.master.Master

A screen shot of a computer program

Description automatically generated

Image 11 Khởi động SPARK Master

A screenshot of a computer

Description automatically generated

Image 12 Kết quả trên web localhost

Khởi động SPARK Worker

A screenshot of a computer program

Description automatically generated

Image 13 Khởi động SPARK Worker

A white background with colorful objects

Description automatically generated

A white background with text

Description automatically generated Image 14 Kết quả trên web localhost

Kết nối Shell với Spark Master

A computer screen shot of a black screen

Description automatically generated

Image 15 Kết nối Shell với Spark Master

A screenshot of a computer

Description automatically generatedImage 16 Kết quả trên web localhost:8080

1. LAUNCHING APPLICATION ON HADOOP – SPARK WINDOWS
2. Khởi động hệ thống: Hadoop, Spark.

A screenshot of a computer application

Description automatically generated

Image 17 Khởi động hệ thống: Hadoop, Spark.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Image 18 Submit wordcount\_demo.py và kết quả

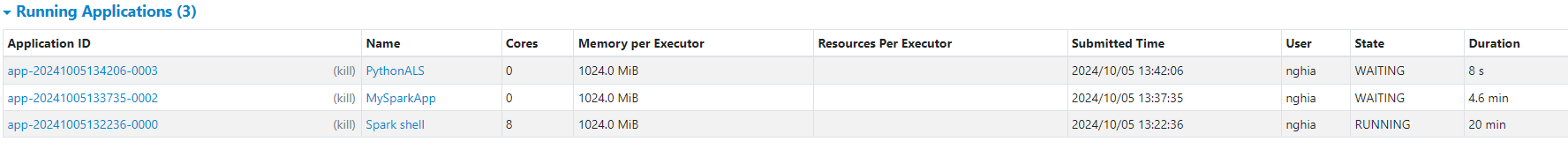


Image 19 Khởi động hệ thống: Hadoop, Spark.

1. XỬ LÝ DỮ LIỆU VĂN BẢN VỚI PYSPARK

4. 1. Mô tả dữ liệu

B1: Khởi tạo ứng dụng PySpark

B2: Tạo RDD

B3: Xử lý trên RDD

A computer screen shot of a program code

Description automatically generated

Image 20 Xử lý dữ liệu văn bản với pyspark

A computer screen shot of a program code

Description automatically generated

Image 21 Xử lý dữ liệu văn bản với pyspark

* 1. Bài tập xử lý dữ liệu:

Cho file văn bản elonmusk\_tweets.csv chứa các dòng tweets của Elon Musk từ 2011-2017. Dữ liệu được chia sẻ bởi Adam Helsinger. Từ file dữ liệu trên, hãy thực hiện Viết chương trình lab3.py thực hiện các xử lý sau với PySpark:

* Liệt kê 20 từ được nhắc đến nhiều nhất

A screenshot of a computer

Description automatically generated

Image 22 Liệt kê 20 từ được nhắc nhiều nhất

* Liệt kê 10 tài khoản được nhắc đến nhiều nhất

A screenshot of a computer

Description automatically generated

Image 23 liệt kê 5 tài khoản được nhắc nhiều nhất

1. ỨNG DỤNG SPARK ĐỂ TRIỂN KHAI CÁC MÔ HÌNH HỌC MÁY (MACHINE LEARNING) PHÂN TÁN

4.1 Kmeans\_pyspark

**Soucer code:**

A black background with colorful lines

Description automatically generated

**Kết quả:**

* **Submit lên spark-master**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA black background with white text

Description automatically generated

Image 24 kết quả Kmeans-pyspark

**4.2 Kmeans\_scikit-learn**

**Soucer code:**

**A computer screen shot of text

Description automatically generated**

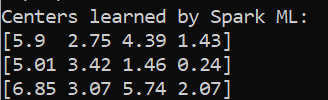
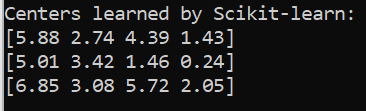
**Kết quả:**

A screen shot of a computer screen

Description automatically generated

Image 25 Kết quả Kmeans-sklearn

**4.3 So sánh tâm cụm**



* Dựa vào kết quả trên, có thể thấy rằng chênh lệch Kmeans\_spark, và Kmeans\_scikitlearn không đáng kể

**4.5 Nhận xét**

- Hiệu suất: Nếu bạn làm việc với tập dữ liệu lớn, PySpark sẽ hiệu quả hơn vì nó được thiết kế cho xử lý phân tán. Scikit-learn thường hoạt động tốt trên dữ liệu nhỏ hơn.

- Dễ sử dụng: Scikit-learn có thể dễ sử dụng hơn cho người mới bắt đầu nhờ vào API trực quan và tài liệu phong phú.

- Khả năng mở rộng: PySpark cho phép bạn mở rộng xử lý dữ liệu lớn mà không gặp vấn đề về hiệu suất.

1. PHÂN TÍCH MẠNG XÃ HỘI

Cho file văn bản ***data/twitter\_following.txt*** chứa thông tin về việc tài khoản người dùng theo dõi tài khoản khác trên cùng mạng xã hội. Mỗi dòng file văn bản có dạng:

**<user\_id1> <user\_id2>**

cho biết người dùng với user\_id1 theo dõi người dùng user\_id2. Nếu user\_id2 cũng theo dõi user\_id1 thì cặp tài khoản này được gọi là theo dõi lẫn nhau (mutual followers).

**Yêu cầu:**

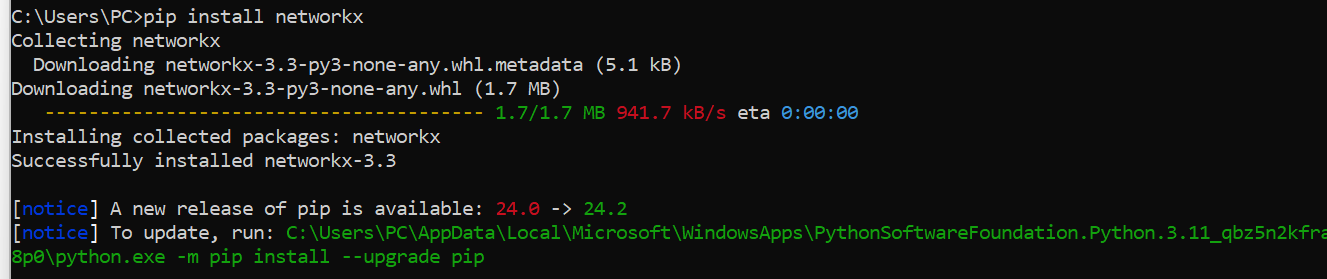
Vận dụng cả 2 phương pháp (1) truyền thống và (2) dữ liệu lớn thực hiện các xử lý sau:

- Thống kê số lượng người theo dõi (followers) của mỗi tài khoản người dùng.

- Liệt kê top 5 người dùng có nhiều theo dõi nhất.

- Liệt kê tất cả các cặp người dùng theo dõi lẫn nhau trong file dữ liệu được cho.

**Cài đặt thư viện networkx**



*Image 26 Cài đặt thư viện networkx*

**Chạy Phương pháp Truyền thống**

Soucer code:

A screen shot of a computer screen

Description automatically generated A screen shot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

*Image 27 Kết quả phương pháp truyền thống.*

**Chạy Phương pháp Dữ liệu Lớn**

Soucer code:

A computer screen shot of a program code

Description automatically generated

A computer screen shot of text

Description automatically generated

* A screenshot of a computer

  Description automatically generatedSubmit lên spark-master

A screenshot of a computer screen

Description automatically generated

Image 28 Kết quả phương pháp dữ liệu lớn

KẾT LUẬN

Báo cáo tổng kết thực hành chuyên đề dữ liệu lớn đã trình bày những khía cạnh quan trọng trong việc cài đặt và sử dụng Hadoop, Spark, cũng như ứng dụng của chúng trong phân tích và xử lý dữ liệu. Qua các chương, chúng ta đã thực hiện các bước cụ thể để thiết lập môi trường, xử lý dữ liệu văn bản và phân tích mạng xã hội, đồng thời áp dụng hai phương pháp truyền thống và dữ liệu lớn.

Cài đặt và Thiết lập Môi Trường: Việc cài đặt Hadoop và Spark trên Windows đã được thực hiện thành công, giúp chúng ta có nền tảng vững chắc để làm việc với dữ liệu lớn. Điều này cũng minh họa cho sự cần thiết của việc thiết lập biến môi trường và cấu hình phù hợp để các công cụ này hoạt động hiệu quả.

Xử Lý Dữ Liệu với PySpark: Qua việc thực hiện các bài tập xử lý dữ liệu văn bản, chúng ta đã có cái nhìn sâu sắc hơn về cách PySpark xử lý các tập dữ liệu lớn một cách hiệu quả. Việc thống kê tần suất từ ngữ và tài khoản đã cho thấy khả năng xử lý nhanh chóng của PySpark so với phương pháp truyền thống.

Phân Tích Mạng Xã Hội: Chương phân tích mạng xã hội đã giúp chúng ta hiểu rõ hơn về cách mà các tài khoản người dùng tương tác với nhau trên nền tảng mạng xã hội. Việc xác định số lượng người theo dõi, tìm kiếm top người dùng có nhiều theo dõi nhất, cũng như phát hiện các cặp tài khoản theo dõi lẫn nhau đã cho thấy tính ứng dụng cao của các kỹ thuật phân tích mạng.

So Sánh Hai Phương Pháp: Qua việc áp dụng cả phương pháp truyền thống và dữ liệu lớn, chúng ta nhận thấy rằng PySpark vượt trội hơn trong việc xử lý các tập dữ liệu lớn, trong khi phương pháp truyền thống dễ hiểu và phù hợp với dữ liệu nhỏ hơn.

**Định Hướng Tương Lai**

Từ những kinh nghiệm và kiến thức thu được, chúng ta có thể mở rộng nghiên cứu sang các lĩnh vực khác như học máy (machine learning) hoặc phân tích dự đoán, nhằm khai thác tối đa giá trị từ dữ liệu lớn. Sự phát triển của công nghệ sẽ tiếp tục mở ra nhiều cơ hội và thách thức, đòi hỏi những người làm trong lĩnh vực này không ngừng học hỏi và thích nghi.

TÀI LIỆU THAM KHẢO

[1] File bài giảng

[2] <https://www.who.int/vietnam/vi/health-topics/cardiovascular-disease>

[3] <https://en.wikipedia.org/wiki/ID3_algorithm>

[4] <https://en.wikipedia.org/wiki/C4.5_algorithm>

[5] <https://machinelearningcoban.com/2017/08/31/evaluation/>

[6] <https://machinelearningcoban.com/2018/01/14/id3/>

[7] <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

[8]<https://www.kaggle.com/code/tanmay111999/heart-failure-prediction-cv-score-90-5-models>