**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ - ĐẠI HỌC QUỐC GIA HÀ NỘI**

**VIỆN TRÍ TUỆ NHÂN TẠO**

----🙤🙤🕮🙦🙦----

****

**BÁO CÁO TỔNG QUAN**

**BÀI TẬP LỚN LẬP TRÌNH XỬ LÝ DỮ LIỆU**

**ĐỀ TÀI:**

**PHÂN TÍCH DỮ LIỆU GIÁ CỔ PHIẾU**

|  |  |  |
| --- | --- | --- |
| **Họ và tên** |  | **Mã sinh viên** |
| Mai Phan Anh Tùng | : | 23020433 |
| Nguyễn Năng Thịnh | : | 23020439 |
| Trịnh Hoàng Đức | : | 23020359 |

*Hà Nội, ngày 8/12/2024*

# MỤC LỤC

[**MỤC LỤC** 1](#_Toc184148294)

[**1.** **TỔNG QUAN** 3](#_Toc184148295)

[**1.1.** **Giới thiệu** 3](#_Toc184148296)

[**1.2.** **Mục tiêu** 3](#_Toc184148297)

[**1.3.** **Nội dung** 3](#_Toc184148298)

[**2.** **THU THẬP DỮ LIỆU** 4](#_Toc184148299)

[**2.1.** **Thu thập dữ liệu từ Yahoo Finance** 4](#_Toc184148300)

[**2.2.** **Thu thập dữ liệu từ Wikipedia** 4](#_Toc184148301)  
**2.3** [**Lưu trữ dữ liệu bằng Excel 4**](#_Toc184148301)

[**3.** **XỬ LÝ DỮ LIỆU** 5](#_Toc184148302)

[**3.1.** **Thêm những dữ liệu không có sẵn** 5](#_Toc184148303)

[**3.2.** **Loại bỏ những dữ liệu không sử dụng** 5](#_Toc184148304)

[**3.3.** **Định dạng lại dữ liệu chưa phù hợp** 6](#_Toc184148305)

[**3.4.** **Tối ưu hóa thời gian thu thập** 6](#_Toc184148306)

[**4.** **PHÂN TÍCH, TRỰC QUAN HÓA DỮ LIỆU** 6](#_Toc184148307)

[**4.1.** **Phân tích tổng quan** 6](#_Toc184148308)

[**4.1.1.** **Thời điểm xu hướng biến động cổ phiếu** 6](#_Toc184148309)

[**4.1.2.** **Thời gian để đánh giá hiệu suất**  7](#_Toc184148310)

[**4.2.** **Phân tích theo từ khóa** 7](#_Toc184148311)

[**4.3.** **Phân tích theo chủ đề** 8](#_Toc184148312)

[**4.3.1.** **Trung bình lượt tương tác theo chủ đề** 8](#_Toc184148314)

[**4.4.** **Phân tích theo thời gian** 8](#_Toc184148315)

[**4.4.1.** **Chủ đề cụ thể trong dữ liệu thời gian** 9](#_Toc184148317)

[**5.** **KẾT LUẬN** 9](#_Toc184148319)

**TỔNG QUAN**

* 1. **Giới thiệu**

**- Phân tích dữ liệu** là quá trình thu thập, xử lý, kiểm tra và mô hình hóa dữ liệu để rút ra thông tin hữu ích, hỗ trợ đưa ra quyết định hoặc hiểu rõ hơn về một vấn đề cụ thể.

- Dự án của chúng em nhằm mục đích thu thập và phân tích dữ liệu cổ phiếu của các công ty lớn tại Mỹ. Việc phân tích dữ liệu này giúp hiểu rõ hơn về xu hướng thị trường và hiệu suất của các công ty, từ đó cung cấp thông tin hữu ích cho các nhà đầu tư và các bên liên quan.

- Python là một ngôn ngữ lập trình mạnh mẽ, linh hoạt và đặc biệt phù hợp với các tác vụ về xử lý dữ liệu, trực quan hóa, cũng như xây dựng mô hình dự đoán.

- Vì vậy, trong dự án này, chúng em sẽ sử dụng **Python** như công cụ chính và kết hợp với kiến thức về xác suất thống kê để thực hiện các bước xử lý và phân tích dữ liệu.

* 1. **Mục tiêu**

**-** Mục tiêu dự án của chúng em hướng tới bao gồm:

* Thu thập giá cổ phiếu và khối lượng giao dịch của các công ty lớn thông qua việc sử dụng thư viện `yfinance` và các nguồn dữ liệu trực tuyến khác.
* Phân tích và trực quan hoá dữ liệu để rút ra các kết luận về hiệu suất của các công ty, bao gồm việc xác định các xu hướng và mô hình quan trọng trong dữ liệu
* Dự đoán xu hướng giá cổ phiếu trong tương lai dựa trên dữ liệu đã phân tích, giúp các nhà đầu tư có cái nhìn tổng quan về thị trường.
  1. **Nội dung**

- Nội dung dự án của chúng em bao gồm:

* Thu thập dữ liệu
* Xử lý dữ liệu
* Phân tích dữ liệu
* Trực quan hóa dữ liệu
* Kết luận

1. **THU THẬP DỮ LIỆU**
   1. **Thu thập dữ liệu từ Yahoo Finance**

**-** Chúng em sử dụng thư viện `yfinance` để tải dữ liệu giá cổ phiếu và khối lượng giao dịch của các công ty. Dữ liệu được thu thập cho khoảng thời gian từ đầu năm 2020 đến cuối năm 2024, cho phép phân tích hiệu suất theo thời gian

* 1. **Thu thập dữ liệu từ Wikipedia**
* Sử dụng thư viện **pandas** để đọc bảng dữ liệu về các công ty lớn nhất theo doanh thu từ trang Wikipedia. Dữ liệu này được sử dụng để xác định danh sách các công ty hàng đầu mà chúng ta sẽ phân tích.
  1. **Lưu trữ dữ liệu**
* Dữ liệu được lưu trữ trong các file Excel cho từng công ty, giúp chúng em dễ dàng truy cập và phân tích sau này. Mỗi file Excel chứa thông tin chi tiết về giá cổ phiếu và khối lượng giao dịch, được lưu trong thư mục cụ thể trên máy tính.

1. **XỬ LÝ DỮ LIỆU**

**-** Sau khi thu thập dữ liệu, chúng em đã xây dựng được một lượng dữ liệu lớn nhưng lượng dữ liệu đó còn có một số vấn đề như:

* Một số dữ liệu không có sẵn cần được tính toán từ những cột dữ liệu đã có.
* Một số dữ liệu không được sử dụng trong dự án cần được loại bỏ.
* Một số dữ liệu vẫn chưa để đúng định dạng.
* Thời gian lấy dữ liệu chưa được tối ưu.
  1. **Thêm những dữ liệu không có sẵn**

- Dữ liệu sau khi thu thập của chúng em vẫn chưa đầy đủ thông tin và cần được bổ sung như sử dụng các API như `yfinace` để tải dữ liệu giá cổ phiếu và khối lượng giao dịch của các công ty.

* 1. **Loại bỏ những dữ liệu không sử dụng**

**-** Trong quá trình thu thập dữ liệu, các dữ liệu không cần thiết hoặc không liên quan đến chúng em sẽ loại bỏ. Điều này giúp giảm thiểu sự phức tạp của dữ liệu và tập trung vào các thông tin quan trọng, như giá cổ phiếu và khối lượng giao dịch.

* 1. **Định dạng lại dữ liệu chưa phù hợp**

- Dữ liệu thu thập sẽ được chúng em chuẩn hoá và định dạng lại để phù hợp với yêu cầu phân tích

* 1. **Tối ưu hóa thời gian thu thập**

**-** Chúng emsử dụng thu thập theo lô để giảm thiểu thời gian thu thập dữ liệu. Trong hàm `get\_stock\_names()`, dữ liệu cho nhiều công ty được thu thập trong một lần gọi API, giúp tiết kiệm thời gian và tài nguyên

1. **PHÂN TÍCH, TRỰC QUAN HÓA DỮ LIỆU**

* Cung cấp cái nhìn tổng quan về dữ liệu đã thu thập, bao gồm việc phân tích giá cổ phiếu và khối lượng giao dịch của các công ty trong khoảng thời gian từ năm 2020 đến năm 2024. Dữ liệu được phân tích theo từng công ty, cho phép so sánh hiệu suất giữa các công ty khác nhau.
  1. **Phân tích tổng quan**
     1. **Thời điểm**
* Phân tích dữ liệu theo thời điểm cụ thể, chẳng hạn như ngày, tháng, năm, để xác định các xu hướng và biến động trong giá cổ phiếu.
  + 1. **Thời gian**
* Phân tích dữ liệu theo khoảng thời gian, ví dụ như theo quý hoặc năm, để đánh giá hiệu suất dài hạn của các công ty.
  1. **Phân tích theo từ khóa**
* Phân tích dữ liệu dựa trên các từ khóa cụ thể liên quan đến các công ty hoặc lĩnh vực, giúp xác định các yếu tố ảnh hưởng đến giá cổ phiếu và khối lượng giao dịch.
  1. **Phân tích theo chủ đề**

**4.3.1 Trung bình**

**-** Tính toán và phân tích các giá trị trung bình của dữ liệu, chẳng hạn như giá cao nhất và thấp nhất của cổ phiếu, để có cái nhìn tổng quan về hiệu suất của từng công ty.

* 1. **Phân tích theo thời gian**
     1. **Chủ đề**

**-** Phân tích các chủ đề cụ thể trong dữ liệu theo thời gian, giúp xác định các xu hướng và mô hình trong hành vi của thị trường.

1. **KẾT LUẬN**

Tóm tắt các phát hiện chính từ phân tích dữ liệu, nhấn mạnh các xu hướng và mô hình quan trọng:

* Dữ liệu giá cổ phiếu và khối lượng giao dịch của các công ty hàng đầu đã được thu thập và phân tích.
* Các xu hướng và mô hình quan trọng đã được xác định, bao gồm sự tăng trưởng của giá cổ phiếu trong một số công ty và sự giảm sút của khối lượng giao dịch trong một số khác.
* Các phân tích đã chỉ ra rằng một số công ty có hiệu suất tốt hơn so với các công ty khác, và các yếu tố như doanh thu, lợi nhuận và tỷ lệ tăng trưởng đã đóng vai trò quan trọng trong việc xác định hiệu suất của các công ty.