TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO CUỐI KỲ MÔN HỌC**

**KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**THU THẬP VÀ PHÂN CỤM GIÁ CỔ PHIẾU TỪ SỞ GIAO DỊCH CHỨNG KHOÁN THÀNH PHỐ HỒ CHÍ MINH (HOSE) TRONG NĂM 2021**

*Người hướng dẫn*: **TS. LÊ CUNG TƯỞNG**

*Người thực hiện*: **HUỲNH MINH HẢI – 51800373**

**NGUYỄN TẤN TÀI – 51800112**

**NGUYỄN QUỐC BẢO – 51800011**

Lớp: **18050201**

Khoá**: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO CUỐI KỲ MÔN HỌC**

**KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**THU THẬP VÀ PHÂN CỤM GIÁ CỔ PHIẾU TỪ SỞ GIAO DỊCH CHỨNG KHOÁN THÀNH PHỐ HỒ CHÍ MINH (HOSE) TRONG NĂM 2021**

*Người hướng dẫn*: **TS. LÊ CUNG TƯỞNG**

*Người thực hiện*: **HUỲNH MINH HẢI – 51800373**

**NGUYỄN TẤN TÀI – 51800112**

**NGUYỄN QUỐC BẢO – 51800011**

Lớp: **18050201**

Khoá**: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn **TS. Lê Cung Tưởng** – giảng viên bộ môn *Khai thác dữ liệu và Khai phá tri thức (505043)* đã nhiệt tình giúp đỡ cũng như giải đáp những thắc mắc của chúng em trong suốt quá trình chúng em thực hiện bài báo cáo cuối kỳ này. Nhờ có sự hướng dẫn, gợi ý tận tình của Thầy về những kiến thức, tài liệu và các kỹ năng liên quan mà chúng em mới hoàn thành được bài báo cáo cuối kỳ với đề tài *“Thu thập và phân cụm giá cổ phiếu từ Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh (HOSE) trong năm 2021”.*

Dĩ nhiên, với khả năng và kiến thức còn hạn hẹp, bài báo cáo của chúng em vẫn còn nhiều hạn chế và thiếu sót. Do vậy, chúng em mong rằng sẽ nhận được những lời góp ý và đánh giá từ Thầy để bài báo cáo cuối kỳ của chúng em được hoàn thiện hơn.

Một lần nữa, chúng em xin chân thành cảm ơn Thầy!

BÁO CÁO CUỐI KỲ ĐƯỢC HOÀN THÀNH

TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Chúng tôi xin cam đoan đây là báo cáo cuối kỳ của riêng chúng tôi và được sự hướng dẫn của TS. Lê Cung Tưởng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa từng công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính chúng tôi thu thập từ các nguồn khác nhau và có ghi rõ trong phần tài liệu tham khảo.

**Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung báo cáo cuối kỳ của mình.** Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 20 tháng 12 năm 2021*

*Tác giả*

*(ký và ghi rõ họ tên)*

*Huỳnh Minh Hải*

*Nguyễn Tấn Tài*

*Nguyễn Quốc Bảo*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(ký và ghi rõ họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(ký và ghi rõ họ tên)

TÓM TẮT

Từ những năm 60 của thế kỷ trước, người ta đã bắt đầu sử dụng các công cụ máy tính để tổ chức và khai phá các cơ sở dữ liệu. Cùng với sự phát triển vượt bậc của công nghệ thông tin và truyền thông, khả năng thu thập, lưu trữ và xử lý dữ liệu cho các hệ thống thông tin không ngừng được tăng cao. Ở thời điểm này, người ta đã bắt đầu đề cập đến khái niệm khủng hoảng phân tích dữ liệu để cung cấp thông tin với yêu cầu chất lượng phải ngày càng cao cho người ra quyết định trong các tổ chức tài chính, kinh tế, khoa học,... Lượng dữ liệu khổng lồ này thực sự là một nguồn tài nguyên có nhiều giá trị bởi lẽ thông tin là yếu tố then chốt trong mọi hoạt động quản lý, kinh doanh, phát triển sản xuất và dịch vụ của mọi tổ chức, doanh nghiệp. Nó giúp người điều hành hiểu biết về môi trường và tiến trình hoạt động của tổ chức trước khi đưa ra quyết định để tác động đến quá trình hoạt động nhằm đạt được mục tiêu một cách hiệu quả và bền vững.

Khai phá dữ liệu (Data Mining) là một lĩnh vực mới nổi nhằm tự động khai thác những thông tin, những tri thức tiềm ẩn, hữu ích từ những cơ sở dữ liệu lớn cho các đơn vị, tổ chức, doanh nghiệp,... từ đó thúc đẩy khả năng sản xuất và cạnh tranh của các đơn vị, tổ chức này. Để hiểu rõ hơn, chúng ta đi vào đề tài “*Thu thập và phân cụm giá cổ phiếu từ Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh (HOSE) trong năm 2021*” nhằm có cái nhìn tổng quan về lĩnh vực này.

Trong CHƯƠNG 1 – MÔ TẢ ĐỀ TÀI sẽ trình bày cụ thể các yêu cầu của đề tài. Các thư viện, module được sử dụng,... sẽ được trình bày trong CHƯƠNG 2 – CƠ SỞ LÝ THUYẾT. CHƯƠNG 3 – THỰC HIỆN CHƯƠNG TRÌNH là phần mã nguồn nhóm đã thực hiện theo các yêu cầu của đề tài. Phần cuối cùng, CHƯƠNG 4 – KẾT LUẬN là những nhận xét, đánh giá của nhóm về đề tài này.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc91271686)

[BÁO CÁO CUỐI KỲ ĐƯỢC HOÀN THÀNH ii](#_Toc91271687)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc91271688)

[TÓM TẮT iv](#_Toc91271689)

[MỤC LỤC 1](#_Toc91271690)

[DANH MỤC CÁC CHỮ VIẾT TẮT 2](#_Toc91271691)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 3](#_Toc91271692)

[CHƯƠNG 1 – MÔ TẢ ĐỀ TÀI 5](#_Toc91271693)

[CHƯƠNG 2 – CƠ SỞ LÝ THUYẾT 7](#_Toc91271694)

[2.1 Thư viện Selenium WebDriver 7](#_Toc91271695)

[2.2 Thư viện BeautifulSoup 8](#_Toc91271696)

[2.3 Hàm sleep() trong module time 8](#_Toc91271697)

[2.4 Gói tslearn 8](#_Toc91271698)

[CHƯƠNG 3 – THỰC HIỆN CHƯƠNG TRÌNH 10](#_Toc91271699)

[3.1 Nhiệm vụ 1 10](#_Toc91271700)

[3.2 Nhiệm vụ 2 13](#_Toc91271701)

[3.3 Nhiệm vụ 3 15](#_Toc91271702)

[CHƯƠNG 4 – KẾT LUẬN 23](#_Toc91271703)

[TÀI LIỆU THAM KHẢO 24](#_Toc91271704)

DANH MỤC CÁC CHỮ VIẾT TẮT

HOSE Ho Chi Minh Stock Exchange (Sở Giao dịch Chứng khoán Tp. HCM)

API Application Programming Interface (Giao diện lập trình ứng dụng)

HTML HyperText Markup Language (Ngôn ngữ đánh dấu siêu văn bản)

DDoS Distributed Denial of Service (Tấn công từ chối dịch vụ phân tán)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

[Hình 1.1 Ví dụ kết quả cần hiển thị ở nhiệm vụ 3 6](#_Toc91271705)

[Hình 3.1 Import thư viện, module cần thiết cho nhiệm vụ 1 10](#_Toc91271706)

[Hình 3.2 Danh sách 60 mã chứng khoán 10](#_Toc91271707)

[Hình 3.3 Danh sách rỗng để lưu giá đóng cửa sau mỗi vòng lặp 10](#_Toc91271708)

[Hình 3.4 Sử dụng Selenium WebDriver để mở trình duyệt web 11](#_Toc91271709)

[Hình 3.5 Vòng lặp for để lấy dữ liệu (1) 11](#_Toc91271710)

[Hình 3.6 Vòng lặp for để lấy dữ liệu (2) 12](#_Toc91271711)

[Hình 3.7 Đảo ngược mảng dữ liệu để hiện thị đúng theo dòng thời gian 12](#_Toc91271712)

[Hình 3.8 Một phần dữ liệu trong file data.csv 13](#_Toc91271713)

[Hình 3.9 Import thư viện, module cần thiết cho nhiệm vụ 2 13](#_Toc91271714)

[Hình 3.10 Vòng lặp scale dữ liệu của 60 mã chứng khoán 13](#_Toc91271715)

[Hình 3.11 Đoạn code vẽ 5 mã đầu tiên trước khi dữ liệu được scale 14](#_Toc91271716)

[Hình 3.12 Dữ liệu trước khi scale 14](#_Toc91271717)

[Hình 3.13 Đoạn code vẽ 5 mã đầu tiên sau khi dữ liệu được scale 15](#_Toc91271718)

[Hình 3.14 Dữ liệu sau khi scale 15](#_Toc91271719)

[Hình 3.15 Khởi tạo mảng tập dữ liệu chuỗi thời gian 16](#_Toc91271720)

[Hình 3.16 Code phương pháp khuỷu tay tìm k tốt nhất với biến thể Euclidean k-means 16](#_Toc91271721)

[Hình 3.17 Biểu đồ thể hiện các biến dạng k của biến thể Euclidean k-means 16](#_Toc91271722)

[Hình 3.18 Code phương pháp khuỷu tay tìm k tốt nhất với biến thể DBA k-means 17](#_Toc91271723)

[Hình 3.19 Biểu đồ thể hiện các biến dạng k của biến thể DBA k-means 17](#_Toc91271724)

[Hình 3.20 Code phương pháp khuỷu tay tìm k tốt nhất với biến thể Soft-DTW k-means 18](#_Toc91271725)

[Hình 3.21 Biểu đồ thể hiện các biến dạng k của biến thể Soft-DTW k-means 18](#_Toc91271726)

[Hình 3.22 Vòng lặp vẽ hình và in mã chứng khoán của từng cụm biến thể Euclidean k-means 19](#_Toc91271727)

[Hình 3.23 Kết quả phân cụm dữ liệu chứng khoán với biến thể Euclidean k-means 20](#_Toc91271728)

[Hình 3.24 Vòng lặp vẽ hình và in mã chứng khoán của từng cụm biến thể DBA k-means 21](#_Toc91271729)

[Hình 3.25 Kết quả phân cụm dữ liệu chứng khoán với biến thể DBA k-means 21](#_Toc91271730)

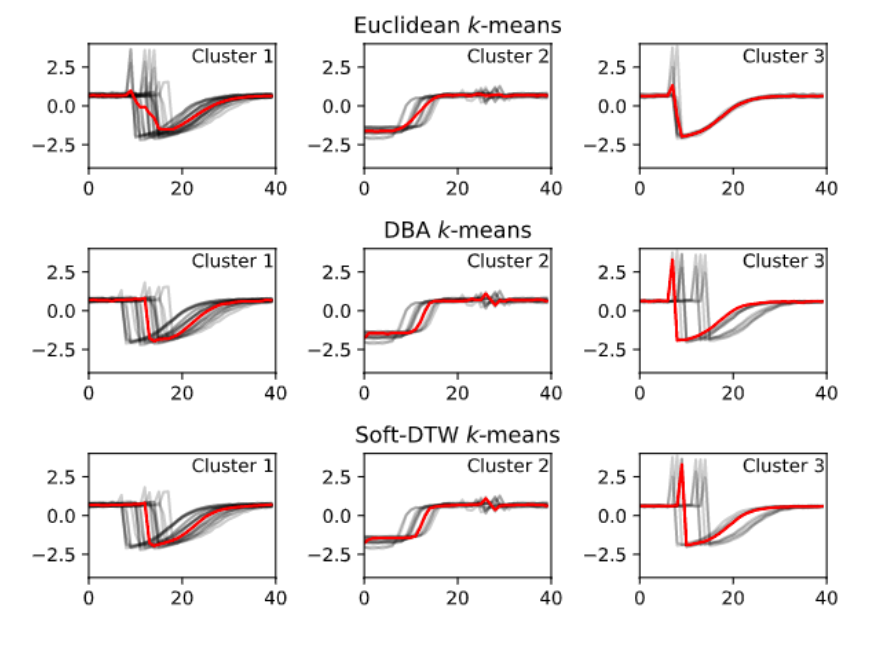
[Hình 3.26 Vòng lặp vẽ hình và in mã chứng khoán của từng cụm biến thể Soft-DTW k-means 22](#_Toc91271731)

[Hình 3.27 Kết quả phân cụm dữ liệu chứng khoán với biến thể Soft-DTW k-means 22](#_Toc91271732)

# – MÔ TẢ ĐỀ TÀI

Đề tài “*Thu thập và phân cụm giá cổ phiếu từ Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh (HOSE) trong năm 2021*” có 3 nhiệm vụ chính như sau:

* Nhiệm vụ 1:
  + Chọn 60 mã cổ phiếu bất kỳ từ HOSE.
  + Thu thập giá cổ phiếu hàng ngày (giá đóng cửa) của 60 mã cổ phiếu này từ bất kỳ trang web nào.
* Nhiệm vụ 2:
* Chuẩn hóa dữ liệu (giá cổ phiếu thu thập được ở nhiệm vụ 1) bằng *TimeSeriesScalerMeanVariance*. Sau đó đưa ra một bức ảnh giá cổ phiếu trước khi chuẩn hóa và một bức ảnh giá cổ phiếu sau khi chuẩn hóa của 5 mã cổ phiếu bất kỳ.
* Nhiệm vụ 3:
* Sử dụng phương pháp khuỷu tay (Elbow) để chọn *k* tốt nhất cho mỗi biến thể:
* Standard Euclidean *k*-means
* DBA *k*-means
* Soft-DTW *k*-means
* Sau đó sử dụng chuỗi thời gian *k*-means với ba biến thể trên để phân cụm giá cổ phiếu.
* Trình bày kết quả như *Hình 1.1*, với:
* Đường màu xám là đường giá cổ phiếu chuyển động trong từng cụm.
* Đường màu đỏ là đường trung bình của các đường màu xám trong từng cụm.



Hình . Ví dụ kết quả cần hiển thị ở nhiệm vụ 3

Ngoài ra, ở nhiệm vụ 2 ta có thể áp dụng một số phương pháp chuẩn hóa cho dữ liệu chuỗi thời gian để thu được kết quả tốt hơn. Sau đó đưa ra so sánh giữa các phương pháp này với TimeSeriesScalerMeanVariance ở nhiệm vụ 3.

# – CƠ SỞ LÝ THUYẾT

Để thực hiện từng nhiệm vụ trong đề tài này, ta cần hình dung một bố cục tổng quát như sau:

* Đầu tiên, ta cần thu thập dữ liệu về từ một trang web chứng khoán bất kỳ. Dĩ nhiên với một lượng lớn dữ liệu thì ta không thể làm bằng cách thủ công được. Chính vì vậy chúng ta cần thu thập dữ liệu một cách tự động và để làm được điều đó ta phải sử dụng đến WebDriver và cụ thể là Selenium WebDriver. Các công cụ này sẽ hỗ trợ tương tác với các thành phần của trang web như nhấp (click), cuộn (scroll), đóng/mở tab, điền biểu mẫu,…
* Một thư viện không thể thiếu trong việc thu thập dữ liệu từ một trang web đó chính là BeautifulSoup. Thư viện này dùng để trích xuất các thông tin từ các thẻ trong tệp HTML, đây là cái mà chúng ta cần để thực hiện nhiệm vụ 1 (thu thập giá đóng cửa của các mã cổ phiếu).
* Việc thực hiện liên tục các thao tác với web một cách tự động có thể làm chương trình của chúng ta bị chặn bởi cơ chế chống DDoS. Vì vậy ta cần sử dụng hàm sleep() để delay quá trình thực thi tránh bị từ chối truy cập trong quá trình thu thập dữ liệu tự động.
* Tiếp theo, ta sử dụng TimeSeriesScalerMeanVariance để chuẩn hóa dữ liệu vừa thu thập được cho nhiệm vụ 2 và TimeSeriesKMeans để phân cụm dữ liệu trong nhiệm vụ 3. Và tslearn là gói cung cấp cho chúng ta hai công cụ này.
* Ngoài ra, ta còn sử dụng thư viện Matplotlib để thực hiện vẽ các biểu đồ theo yêu cầu của đề tài.

Dưới đây là giới thiệu một cách ngắn gọn và cách cài đặt các gói, thư viện, module vừa nêu.

## Thư viện Selenium WebDriver

Selenium là một trong những công cụ kiểm thử phần mềm tự động mã nguồn mở phổ biến nhất hiện nay trong việc kiểm thử các ứng dụng Web. Nó có thể chạy được trên hầu hết các trình duyệt web như Chrome, FireFox, Safari, Opera và hầu hết các hệ điều hành như Windows, Mac OS, Linux. Selenium cho phép người dùng lập trình trên các ngôn ngữ khác nhau như Python, Java, Rupy, C#,…

Thông qua Selenium Python API, ta có thể truy cập tất cả các chức năng của Selenium WebDriver một cách trực quan.

Để sử dụng thư viện này ta phải cài đặt gói selenium bằng câu lệnh:

*pip install selenium*

## Thư viện BeautifulSoup

BeautifulSoup là một thư viện dùng để trích xuất dữ liệu từ các tập tin HTML. Nó hoạt động cùng với các trình phân tích cú pháp (parser) cung cấp cho người các để điều hướng, tìm kiếm và chỉnh sửa trong cây phân tích cú pháp (tree parser). Ta có thể sử dụng các hàm như find(), find\_all(),… để truy xuất đến các thẻ trong toàn bộ tập tin HTML.

Để cài đặt và sử dụng thư viện này ta thực hiện các câu lệnh sau:

*pip install beatifulsoup4*

*from bs4 import BeautifulSoup*

## Hàm sleep() trong module time

Module time dùng để xử lý các tác vụ liên quan đến thời gian trong python. Trong module này có nhiều hàm khác nhau, nhưng trong đề tài này ta sử dụng hàm sleep(). Hàm này không trả về giá trị nào mà chỉ delay trình thực thi với thời gian delay bằng giá trị truyền vào.

Để sử dụng hàm này ta thực hiện cài đặt như sau:

*from time import sleep*

## Gói tslearn

tslearn là một gói Python cung cấp các công cụ học máy để phân tích chuỗi thời gian. Gói này được xây dựng trên thư viện scikit-learn, numpy và scipy do đó nó sẽ phụ thuộc vào các thư viện này. Gói này cũng cung cấp các module TimeSeriesScalerMeanVariance và TimeSeriesKMeans để thực hiện các yêu cầu trong nhiệm vụ 2 và 3 của đề tài.

Để cài đặt gói này ta thực hiện câu lệnh sau:

*pip install tslearn*

Để sử dụng TimeSeriesScalerMeanVariance cho nhiệm vụ 2 ta thực hiện câu lệnh:

*from tslearn.preprocessing import TimeSeriesScalerMeanVariance*

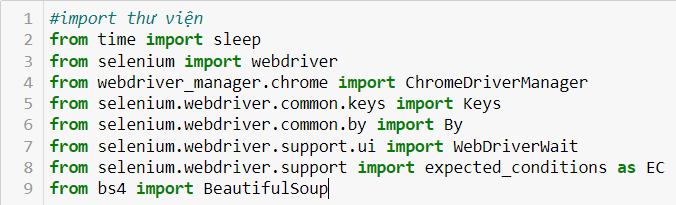
Để sử dụng TimeSeriesKMeans cho nhiệm vụ 3 ta thực hiện câu lệnh:

*from tslearn.clustering import TimeSeriesKMeans*

# – THỰC HIỆN CHƯƠNG TRÌNH

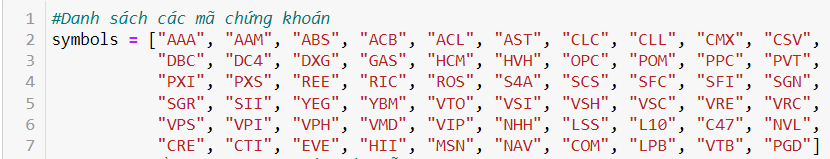
## Nhiệm vụ 1

Đầu tiên, thực hiện import các thư viện và module cần thiết.



Hình . Import thư viện, module cần thiết cho nhiệm vụ 1

Sau đó tạo một danh sách 60 mã chứng khoán để thực hiện thu thập dữ liệu. Các mã chứng khoán này được chọn thủ công và gán vào danh sách.



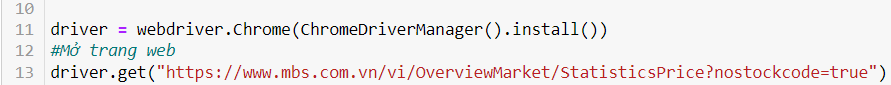
Hình . Danh sách 60 mã chứng khoán

Tạo một list rỗng để lưu giá đóng cửa của mỗi mã chứng khoán sau khi thực hiện vòng lặp.



Hình . Danh sách rỗng để lưu giá đóng cửa sau mỗi vòng lặp

Sử dụng Selenium WebDriver để tự động mở trình duyệt Chrome và truy cập vào trang web Công ty Cổ phần Chứng khoán MB (MBS) với đường dẫn: “<https://www.mbs.com.vn/vi/OverviewMarket/StatisticsPrice?nostockcode=true>”.



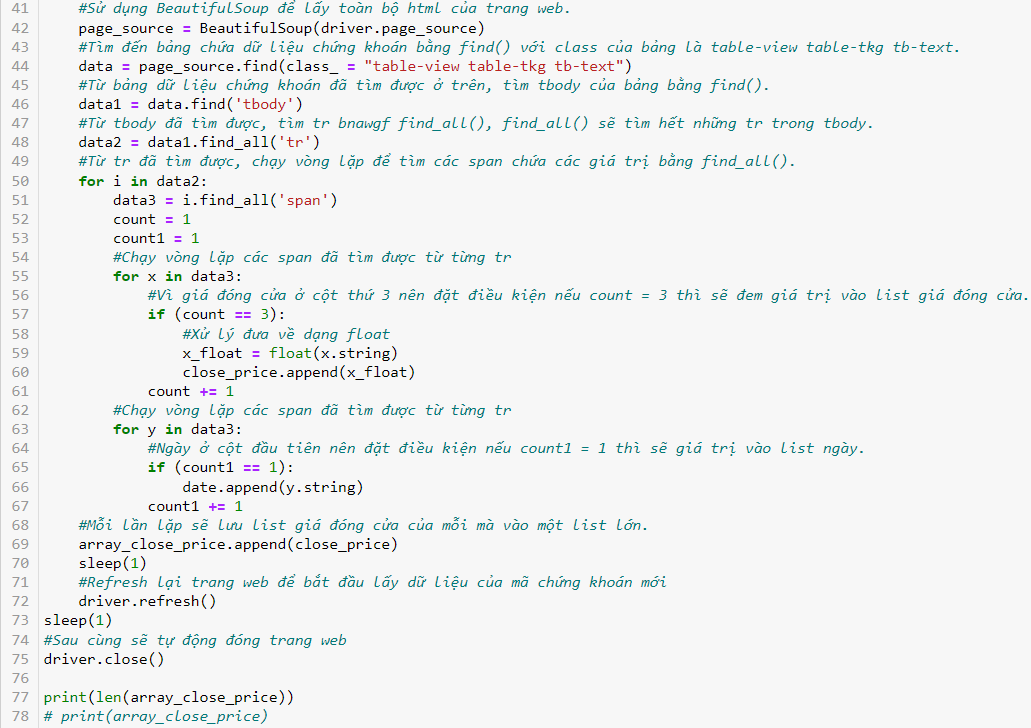
Hình . Sử dụng Selenium WebDriver để mở trình duyệt web

Sử dụng vòng lặp for để thu thập toàn bộ dữ liệu từ 60 mã chứng khoán với ngày bắt đầu là 01/01/2021 và ngày kết thúc là 12/12/2021.



Hình . Vòng lặp for để lấy dữ liệu (1)

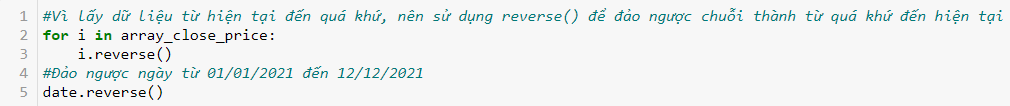
Ta cũng cần sử dụng thư viện BeautifulSoup để lấy dữ liệu ở bước này.



Hình . Vòng lặp for để lấy dữ liệu (2)

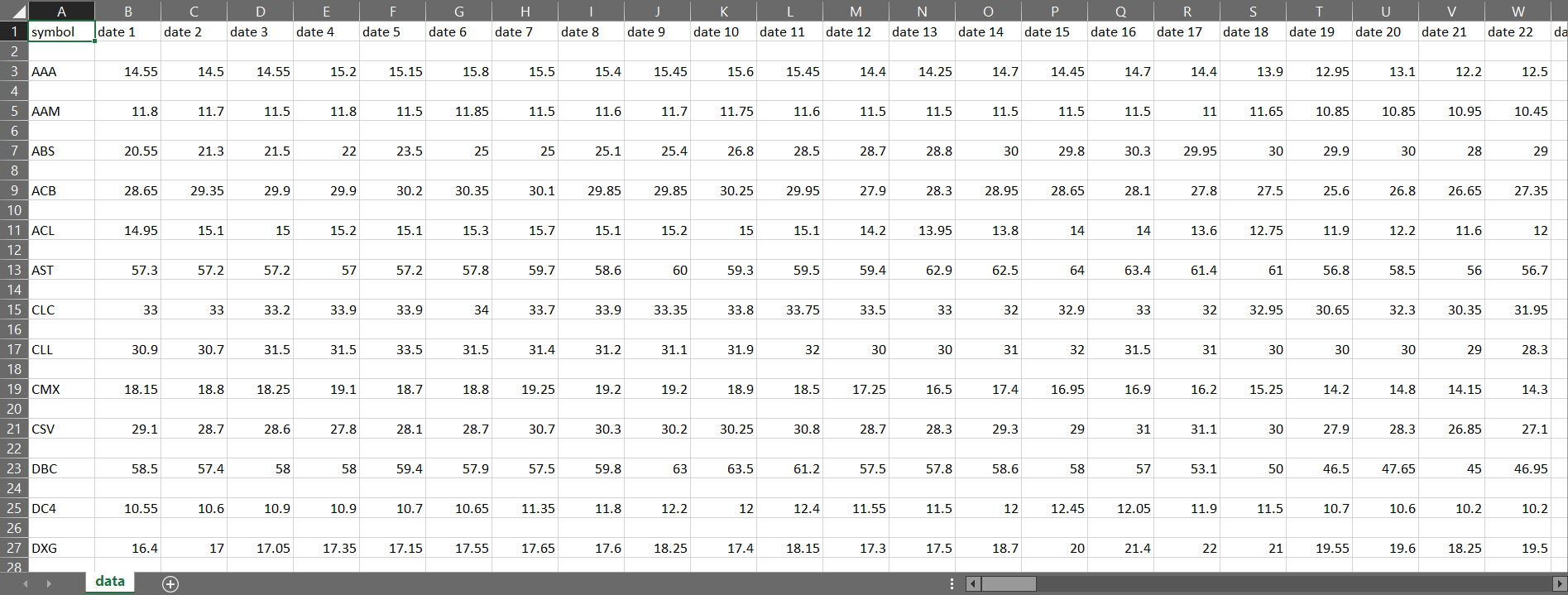
Kết thúc vòng lặp for ta có được danh sách mảng giá đóng cửa của toàn bộ 60 mã chứng khoán bao gồm ngày và giá của từng ngày cụ thể.

Vì vòng lặp thực hiện thu thập dữ liệu từ ngày 12/12/2021 đến ngày 01/01/2021 nên ta dùng hàm reverse() đảo ngược mảng dữ liệu để thu được mảng mới có chiều ngược lại – 01/01/2021 đến 12/12/2021.



Hình . Đảo ngược mảng dữ liệu để hiện thị đúng theo dòng thời gian

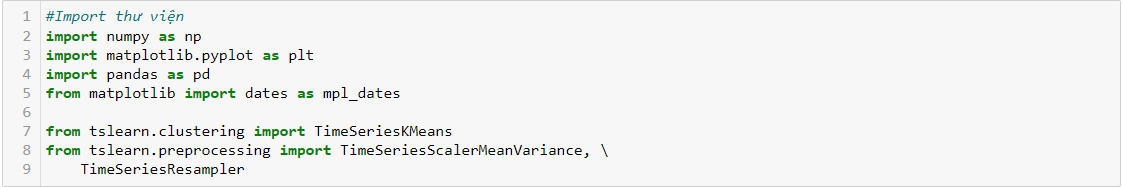
Toàn bộ dữ liệu thu thập được đều được lưu trong file data.csv



Hình . Một phần dữ liệu trong file data.csv

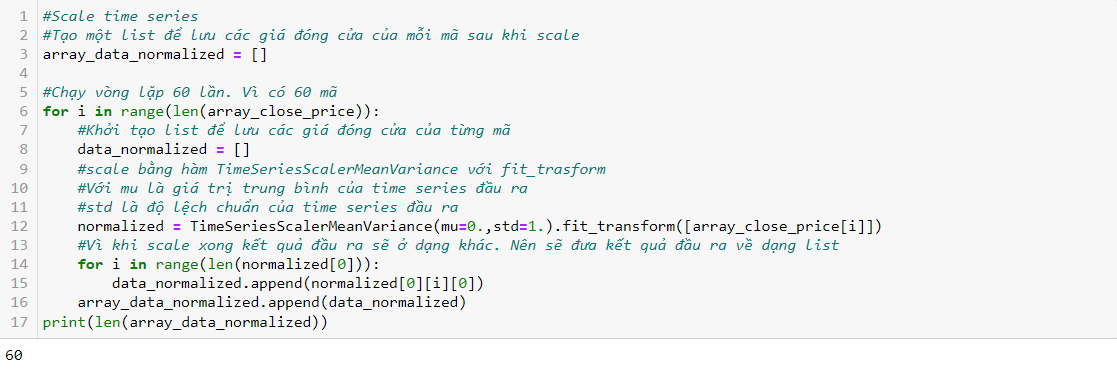
## Nhiệm vụ 2

Thực hiện import các thư viện và module cần thiết.



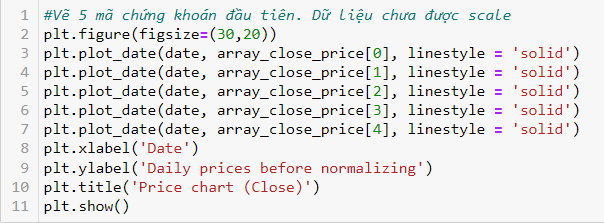
Hình . Import thư viện, module cần thiết cho nhiệm vụ 2

Tạo một danh sách rỗng để lưu giá đóng cửa của mỗi mã chứng khoán sau khi scale và thực hiện vòng lặp để scale.

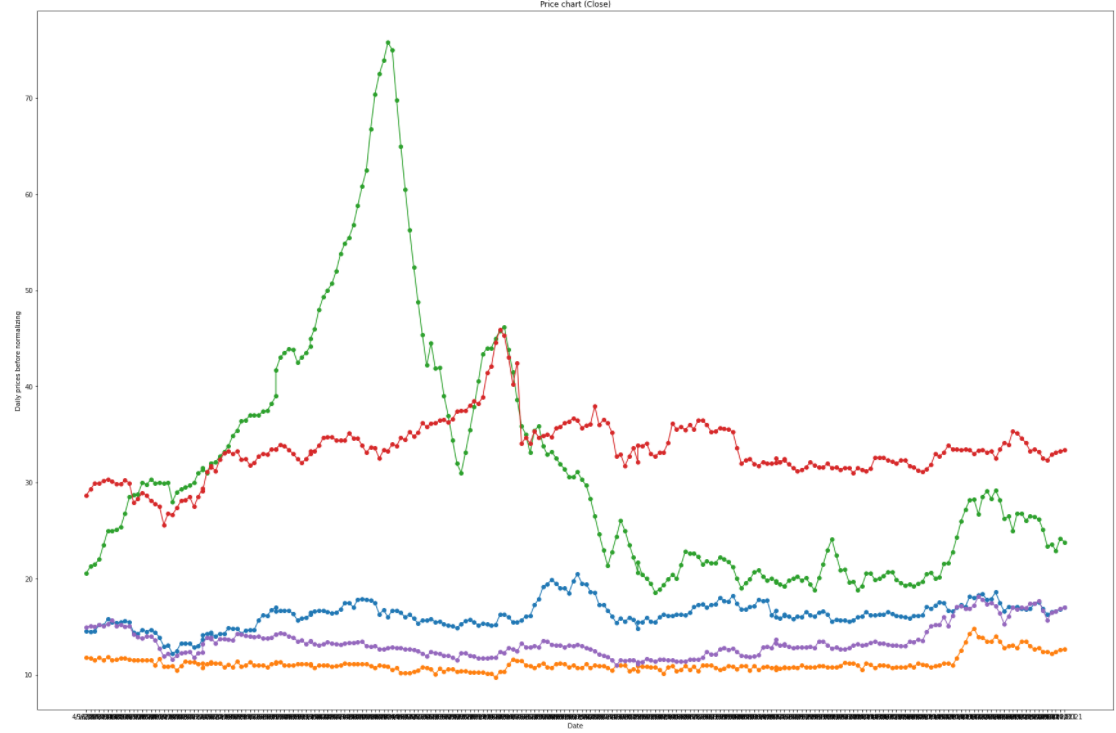


Hình . Vòng lặp scale dữ liệu của 60 mã chứng khoán

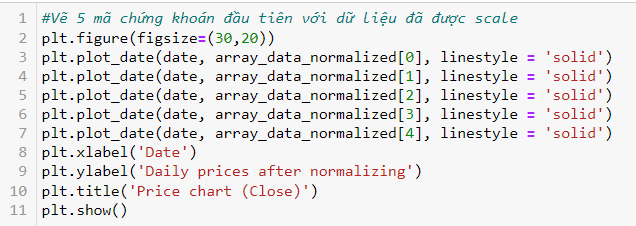
Sau đó thực hiện vẽ biểu đồ của 5 mã chứng khoán đầu tiên trước và sau khi scale.



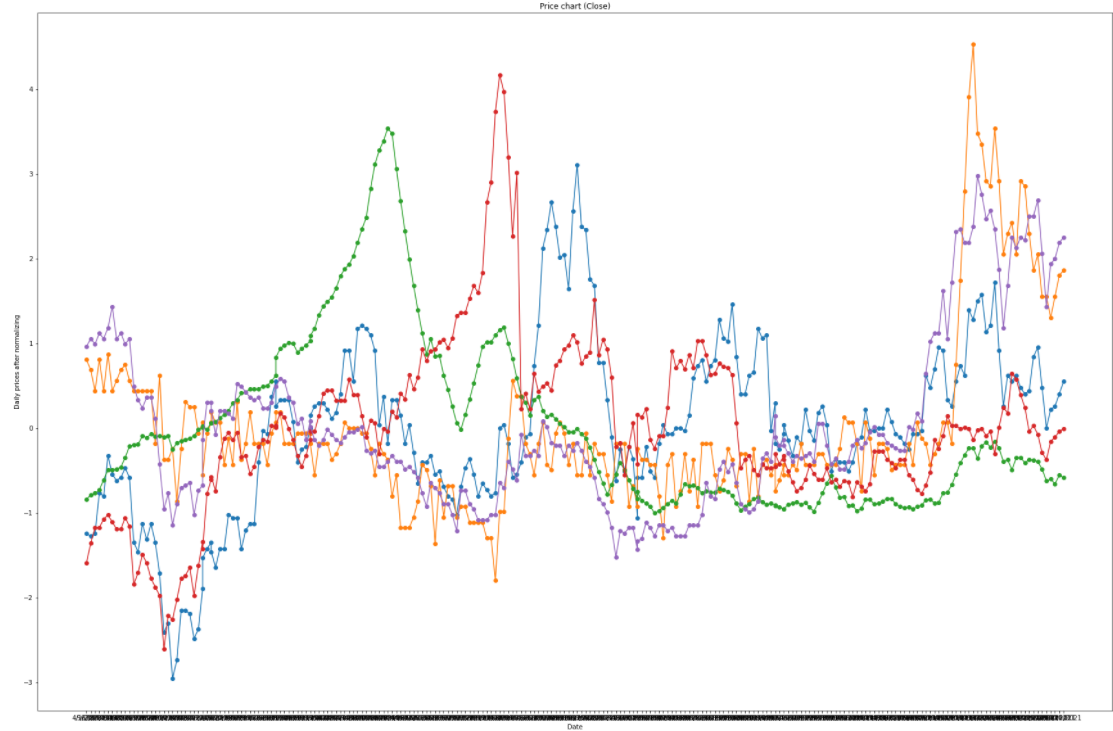
Hình . Đoạn code vẽ 5 mã đầu tiên trước khi dữ liệu được scale



Hình . Dữ liệu trước khi scale



Hình . Đoạn code vẽ 5 mã đầu tiên sau khi dữ liệu được scale



Hình . Dữ liệu sau khi scale

Sau khi thực hiện scale, dữ liệu giá đóng cửa của các mã chứng khoán ban đầu đã được scale về xung quanh giá trị 0.

## Nhiệm vụ 3

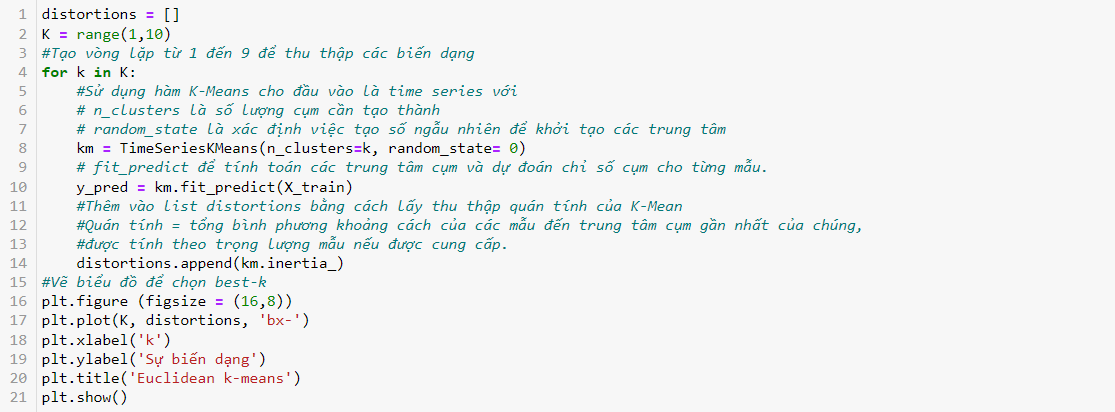
Đầu tiên, khởi tạo mảng là một tập dữ liệu chuỗi thời gian đã được chuẩn hóa.



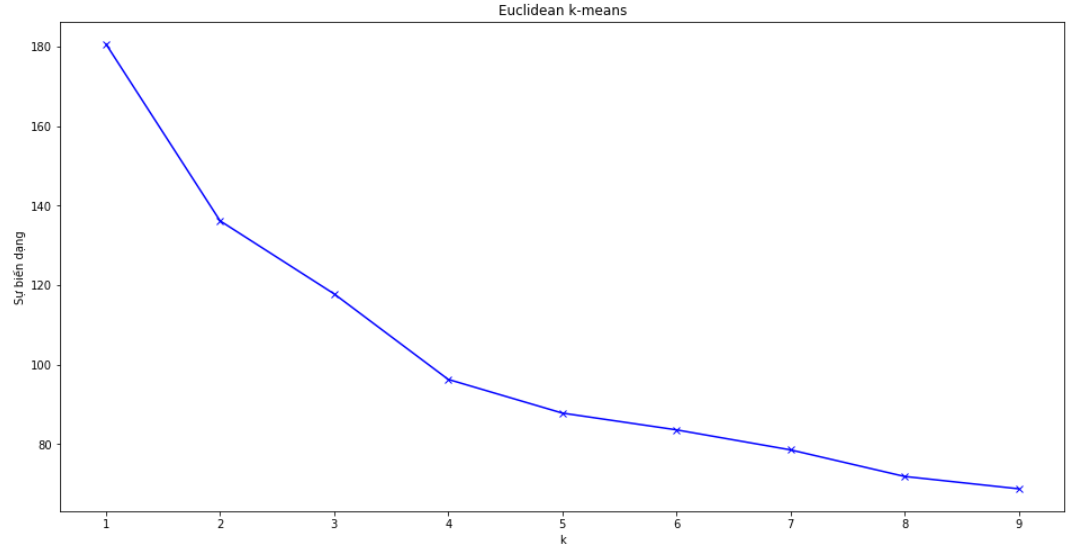
Hình . Khởi tạo mảng tập dữ liệu chuỗi thời gian

Sử dụng phương pháp khuỷu tay (Elbow) để tìm *k* tốt nhất đối với từng biến thể *k*-means:

* Euclidean *k*-means:

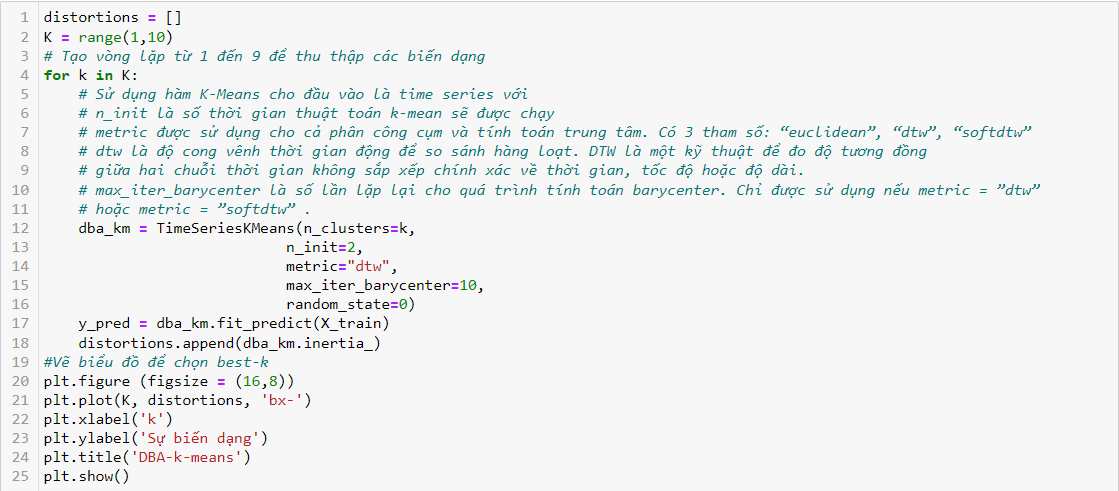


Hình . Code phương pháp khuỷu tay tìm k tốt nhất với biến thể Euclidean k-means

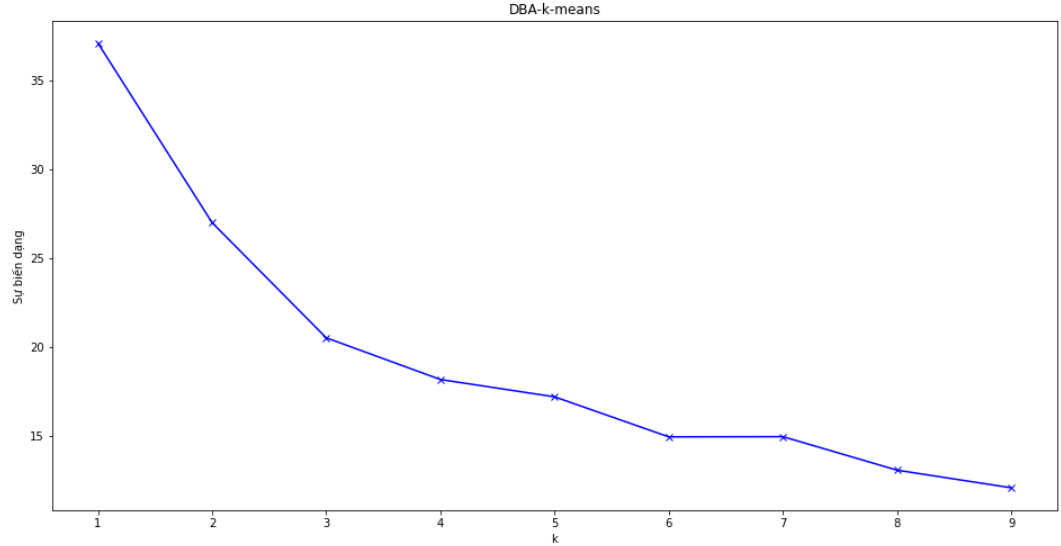


Hình . Biểu đồ thể hiện các biến dạng k của biến thể Euclidean k-means

* DBA *k*-means:

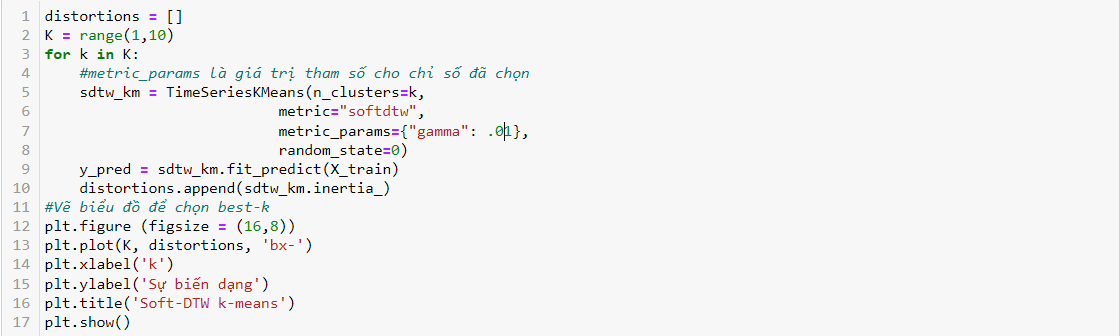


Hình . Code phương pháp khuỷu tay tìm k tốt nhất với biến thể DBA k-means

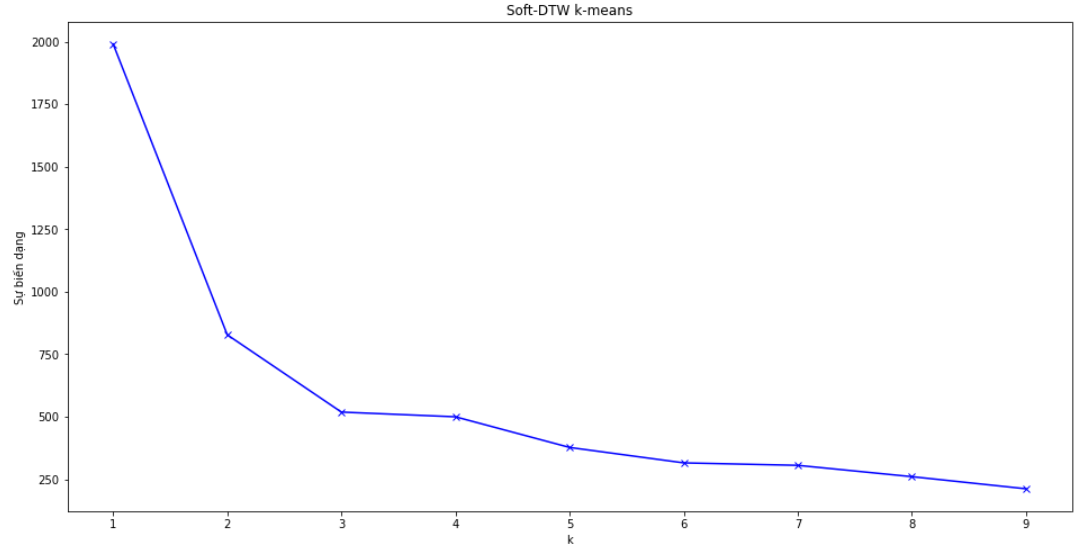


Hình . Biểu đồ thể hiện các biến dạng k của biến thể DBA k-means

* Soft-DTW *k*-means:



Hình . Code phương pháp khuỷu tay tìm k tốt nhất với biến thể Soft-DTW k-means



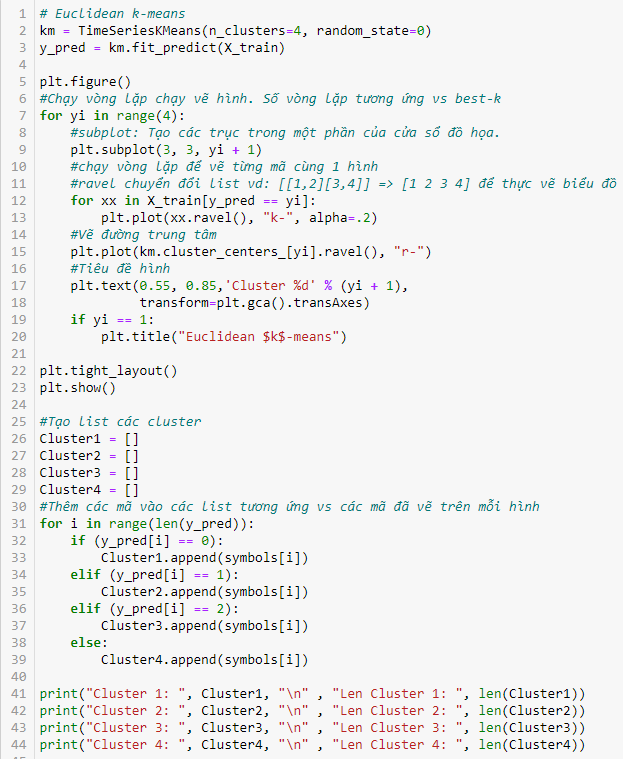
Hình . Biểu đồ thể hiện các biến dạng k của biến thể Soft-DTW k-means

Từ các biểu đồ trên ta tìm được *k* tốt nhất đối với các biến thể *k*-mean lần lượt là:

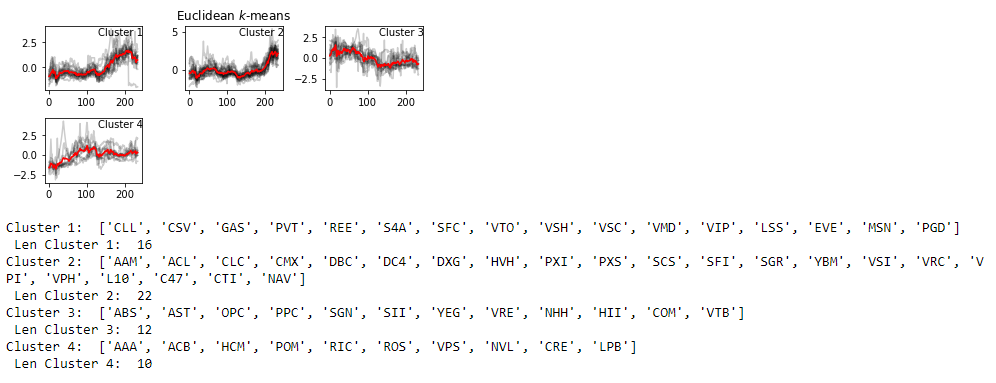
* Euclidean *k*-mean: k = 4
* DBA *k*-means: k = 3
* Soft-DTW k-means: k = 3

Sau đó, ta thực hiện vòng lặp vẽ hình với *k* tốt nhất tương ứng vừa tìm được. Đồng thời thực hiện vòng lặp để thêm các mã chứng khoán tương ứng với các mã đã vẽ trên mỗi cụm.

* Euclidean *k*-mean:



Hình . Vòng lặp vẽ hình và in mã chứng khoán của từng cụm biến thể Euclidean k-means

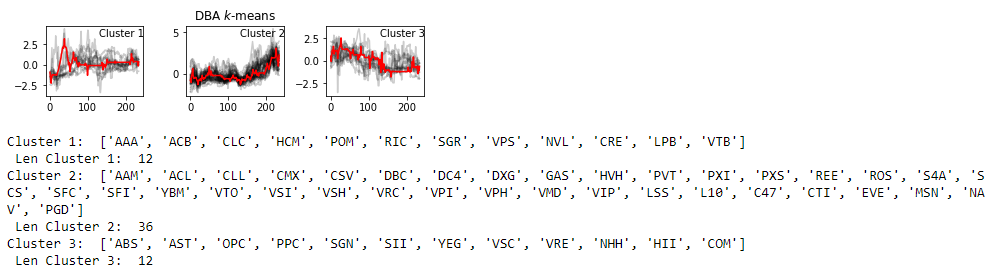


Hình . Kết quả phân cụm dữ liệu chứng khoán với biến thể Euclidean k-means

* DBA *k*-means:

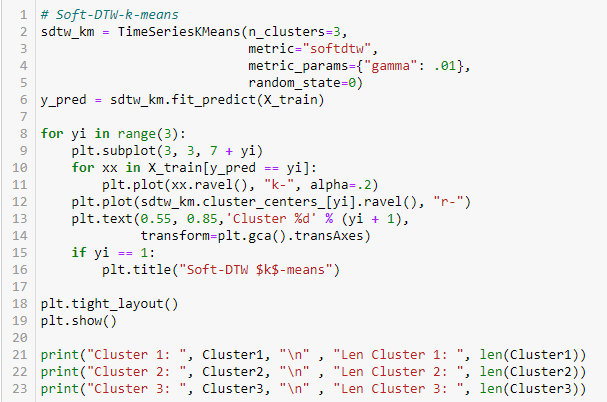


Hình . Vòng lặp vẽ hình và in mã chứng khoán của từng cụm biến thể DBA k-means

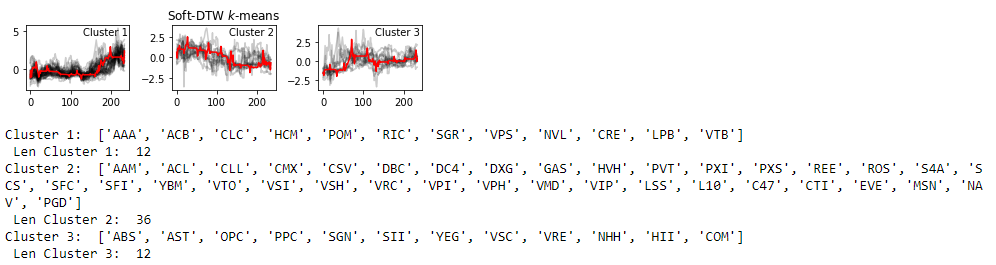


Hình . Kết quả phân cụm dữ liệu chứng khoán với biến thể DBA k-means

* Soft-DTW *k*-means:



Hình . Vòng lặp vẽ hình và in mã chứng khoán của từng cụm biến thể Soft-DTW k-means



Hình . Kết quả phân cụm dữ liệu chứng khoán với biến thể Soft-DTW k-means

# – KẾT LUẬN

Chứng khoán là một lĩnh vực đầu từ sôi động với những biến động cùng những giá trị tài sản lớn. Vì vậy, với mỗi biến động của thị trường chứng khoán dù là nhỏ cũng có thể làm nên sự thay đổi lớn trong nền kinh tế nói chung và của nhà đầu tư cá nhân, doanh nghiệp nói riêng.

Do đó, thông tin về thị trường chứng khoán là rất cần thiết đối với các nhà đầu tư. Đó phải là những thông tin chính xác, có độ tin cậy cao và được cập nhật kịp thời. Hơn nữa, thị trường chứng khoán là một thị trường thường xuyên biến động hàng giờ, hàng ngày,… Thế nên, để phản ánh nhanh chóng những biến động, thị trường cần phải tuân theo những thông số biến động của cổ phiếu như: giá cổ phiếu, khối lượng giao dịch, giá tăng/giảm. Việc phân cụm thị trường chứng khoán giúp các nhà đầu tư xác định được khả năng tài chính của mình và mục tiêu mà họ hướng đến.

Quá trình thực hiện đề tài này, nhóm chúng em đã giải quyết được các vấn đề:

* Thu thập dữ liệu giá đóng cửa của 60 mã chứng khoán trong năm 2021 từ ngày 01/01 đến ngày 12/12.
* Thực hiện chuẩn hóa dữ liệu bằng TimeSeriesScalerMeanVariance. Biểu diễn được hai biểu đồ giá cổ phiểu hàng ngày của 5 mã chứng khoán đầu tiên trước và sau khi chuẩn hóa dữ liệu.
* Thực hiện phương pháp khuỷu tay (Elbow) để tìm *k* tốt nhất và thực hiện phân cụm dữ liệu với 3 biến thể *k*-means.

Tuy nhiên, nhóm chúng em chưa tìm ra được phương pháp khác để chuẩn hóa dữ liệu chuỗi thời gian đối với dữ liệu giá cổ phiếu thu thập được ngoại trừ TimeSeriesScalerMeanVariance đã được gợi ý ở nhiệm vụ 2.

TÀI LIỆU THAM KHẢO

Tiếng Việt:

[1] Trịnh Thị Phương (2017). “Giới thiệu về Selenium Web Driver”, <https://viblo.asia/p/gioi-thieu-ve-selenium-web-driver-924lJOAb5PM>, truy cập ngày 24/11/2021.

[2] Nguyen Manh Thang (2020). “Module Time trong Python”, <https://viblo.asia/p/module-time-trong-python-07LKXeBkZV4>, truy cập ngày 25/11/2021.

Tiếng Anh:

[3] Monty Shokeen (2017). “Scrapping Webpages in Python With Beautiful Soup: The Basics”, <https://code.tutsplus.com/tutorials/scraping-webpages-in-python-with-beautiful-soup-the-basics--cms-28211>, truy cập ngày 24/11/2021.

[4] tslearn 0.5.2 documentation (2021). “tslearn’s documentation”, <https://tslearn.readthedocs.io/en/stable/>, truy cập ngày 28/11/2021.

[5] tslearn 0.5.2 documentation (2021). “tslearn.preprocessing.TimeSeriesScalerMeanVariance”, <https://tslearn.readthedocs.io/en/stable/gen_modules/preprocessing/tslearn.preprocessing.TimeSeriesScalerMeanVariance.html>, truy cập ngày 28/11/2021.

[6] tslearn 0.5.2 documentation (2021). “tslearn.clustering.TimeSeriesKMeans”, <https://tslearn.readthedocs.io/en/stable/gen_modules/clustering/tslearn.clustering.TimeSeriesKMeans.html>, truy cập ngày 28/11/2021.