

ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Hệ thống thu thập, lưu trữ, xử lý và dự báo giá
chứng khoán Việt Nam

TRẦN NGUYỄN ANH TUẤN

tuan.tna200565@sis.hust.edu.vn

Ngành Khoa học máy tính

Giảng viên hướng dẫn: TS. Trần Việt Trung

Chữ kí GVHD

Khoa: Khoa học máy tính

Trường: Công nghệ Thông tin và Truyền thông

HÀ NỘI, 01/2025

LỜI CAM KẾT

Họ và tên sinh viên:

Điện thoại liên lạc:

Email:

Lớp:

Hệ đào tạo:

Tôi – *Trần Nguyễn Anh Tuấn* – cam kết Đồ án Tốt nghiệp (ĐATN) là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của *TS. Trần Việt Trung*. Các kết quả nêu trong ĐATN là trung thực, là thành quả của riêng tôi, không sao chép theo bất kỳ công trình nào khác. Tất cả những tham khảo trong ĐATN – bao gồm hình ảnh, bảng biểu, số liệu, và các câu từ trích dẫn – đều được ghi rõ ràng và đầy đủ nguồn gốc trong danh mục tài liệu tham khảo. Tôi xin hoàn toàn chịu trách nhiệm với dù chỉ một sao chép vi phạm quy chế của nhà trường.

Hà Nội, ngày tháng năm

Tác giả ĐATN

Họ và tên sinh viên

LỜI CẢM ƠN

Quá trình học tập hơn 4 năm tại Đại học Bách Khoa Hà Nội là một quãng thời gian khó quên với em cùng bao thăng trầm trong học tập và cuộc sống. Trong quá trình đó, các thầy cô của các Khoa, Viện và Trường đã tạo tiền đề rèn luyện về mặt kiến thức và kỹ năng để em có khả năng thực hiện đồ án này.

Em xin gửi lời cảm ơn đến TS. Trần Việt Trung, người hướng dẫn em hoàn thành đồ án với những gợi ý, những lời nhận xét và đánh giá quá trình hoàn thiện. Từ đó, em nhận thức được và thúc đẩy bản thân hoàn thiện hết sức với khả năng của mình.

Em xin gửi lời cảm ơn đến anh, chị tại Tổng Công Ty Truyền Thông (VNPT MEDIA) đã tạo điều kiện cho em quan sát, khơi gợi một số ý tưởng nhỏ trong quá trình thực hiện đồ án.

Em xin gửi lời cảm ơn đến bố mẹ, dù bố mẹ không áp lực đến tiến độ nhưng là động lực to lớn để động viên những lúc em muốn từ bỏ.

Quá trình thực hiện đồ án đã giúp em nhận thức nhiều bài học về lập kế hoạch-đánh giá, về phân tích thiết kế, về lập trình, về trình bày-báo cáo để em nắm rõ hơn năng lực bản thân cho định hướng sau này.

Em xin chân thành cảm ơn!

TÓM TẮT NỘI DUNG ĐỒ ÁN

Thị trường tài chính hiện nay được biết là môi trường năng động và phức tạp, nơi việc dự báo giá cổ phiếu đóng vai trò quan trọng trong việc hỗ trợ các quyết định đầu tư hay quản lý rủi ro. Tuy nhiên, việc dự báo giá cổ phiếu gặp nhiều thách thức do các yếu tố như lượng dữ liệu khổng lồ sản sinh theo thời gian từ báo cáo tài chính, giá cổ phiếu lịch sử, và phức tạp hơn là tin tức thị trường. Như vậy, với xu hướng dữ liệu ngày càng lớn, đa dạng và phức tạp cần được khai thác và quản lý hiệu quả với hệ thống chuyên biệt.

Giải pháp đề xuất trong đồ án bao gồm các bước chính như sau. Thu thập dữ liệu được tự động hóa bằng cách sử dụng Selenium [1] và BeautifulSoup [2]. Dữ liệu được lưu trữ trên MongoDB [3] và MinIO [4]. Ngoài ra, việc thu thập dữ liệu mới được kiểm chứng với Evidently [5]. Quá trình xử lý dữ liệu bao gồm làm sạch, chuẩn hóa, và tích hợp dữ liệu; đồng thời phân tích cảm xúc được thực hiện bằng VADER Sentiment để định lượng tâm lý thị trường. Sau đó, mô hình dự báo XGBoost được áp dụng để phân tích và dự báo giá cổ phiếu. Mô hình sẽ được huấn luyện lại nếu phát hiện sai lệch phân phối từ dữ liệu mới. Toàn bộ hệ thống được container hóa với Docker và tự động hóa chức năng dự báo bằng Apache Airflow [6]. Kết quả các quá trình được hiển thị trên dashboard tương tác được xây dựng bằng Streamlit [7].

Đồ án mang lại các đóng góp quan trọng, bao gồm phát triển một luồng hoàn chỉnh từ thu thập, xử lý đến tự động dự báo giá cổ phiếu. Hệ thống tích hợp cảm xúc thị trường vào dữ liệu tài chính để nâng cao độ chính xác của dự báo. Ngoài ra, việc sử dụng các nền tảng lưu trữ và triển khai hiện đại như MinIO và Docker [8] giúp đảm bảo hiệu quả và khả năng mở rộng của hệ thống. Hệ thống cung cấp cách thức để kiểm tra chất lượng dữ liệu và chất lượng mô hình. Cuối cùng, dashboard được phát triển giúp cung cấp một giao diện thân thiện để trực quan hóa kết quả và hỗ trợ ra quyết định.

Sinh viên thực hiện

(Ký và ghi rõ họ tên)

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI.....	1
1.1 Đặt vấn đề.....	1
1.1.1 Sự gia tăng của dữ liệu tài chính và các yếu tố tác động liên quan ...	1
1.1.2 Hệ thống luồng giải quyết vấn đề.....	1
1.2 Mục tiêu và phạm vi đề tài.....	1
1.3 Định hướng giải pháp.....	2
1.4 Bố cục đồ án	3
CHƯƠNG 2. KHẢO SÁT VÀ PHÂN TÍCH YÊU CẦU.....	4
2.1 Khảo sát hiện trạng	4
2.1.1 Công nghệ thu thập dữ liệu	4
2.1.2 Công nghệ lưu trữ dữ liệu	4
2.1.3 Công nghệ kiểm tra chất lượng dữ liệu	4
2.1.4 Mô hình dự báo.....	5
2.2 Tổng quan chức năng	6
2.3 Yêu cầu phi chức năng	6
CHƯƠNG 3. CÔNG NGHỆ SỬ DỤNG.....	8
3.1 Selenium và XPATH.....	8
3.2 BeautifulSoup	8
3.3 Cơ sở dữ liệu phi cấu trúc - MongoDB	9
3.4 Hệ thống lưu trữ - MinIO	9
3.5 Công nghệ trực quan hóa - Streamlit.....	10
3.6 Công nghệ tự động hóa lập lịch - Apache Airflow	11
3.7 Công cụ kiểm tra chất lượng dữ liệu - Evidently	11
3.8 Công nghệ triển khai hệ thống - Docker.....	12

CHƯƠNG 4. THIẾT KẾ VÀ TRIỂN KHAI HỆ THỐNG	13
4.1 Thiết kế kiến trúc.....	13
4.1.1 Mô-đun thu thập dữ liệu	14
4.1.2 Mô-đun lưu trữ dữ liệu	25
4.1.3 Mô-đun biến đổi và xử lý	28
4.1.4 Mô-đun huấn luyện mô hình và dự báo.....	34
4.1.5 Mô-đun trực quan, đánh giá dữ liệu	39
4.2 Triển khai hệ thống.....	47
4.2.1 Các ứng dụng	47
4.2.2 Các thư viện	48
CHƯƠNG 5. KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ.....	51
5.1 Minh họa các chức năng chính.....	51
5.1.1 Thu thập dữ liệu: Trực quan hóa quá trình thu thập giá đóng của của hệ thống	51
5.1.2 Lưu trữ dữ liệu: Hệ thống lưu trữ MinIO	53
5.1.3 Xử Lý Dữ Liệu.....	54
5.1.4 Mô Hình	60
5.2 Kiểm thử.....	63
5.3 Đánh giá	64
CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	65
6.1 Kết quả đóng góp.....	65
6.1.1 Xử lý dữ liệu từ nhiều nguồn	65
6.1.2 Lưu trữ với MinIO thay vì HDFS.....	65
6.1.3 Kiểm soát đầu ra hệ thống.....	66
6.1.4 Kiểm tra hiệu quả của Evidently	66
6.2 Kết luận.....	67

6.3 Hạn chế	68
6.4 Hướng phát triển.....	68

DANH MỤC HÌNH VẼ

Hình 2.1	Minh họa hiện tượng phân phối sai lệch đối với đặc trưng X trên dữ liệu hiện tại với dữ liệu tham chiếu	5
Hình 2.2	Biểu đồ phân rã chức năng	6
Hình 4.1	Kiến trúc hệ thống	13
Hình 4.2	Bảng dữ liệu phiên giao dịch mã AAA (HTML: id="render-table-owner")	14
Hình 4.3	Bảng CĐKT của mã AAA tại CafeF (HTML: id="aNhom1") .	16
Hình 4.4	Bản cân đối kế toán của mã AAA tại VietStock trong 2 năm gần nhất, đơn vị:tỷ đồng. (HTML: class="table table-hover")	18
Hình 4.5	Giao diện Selenium Grid. Hiện tại không có trình duyệt nào đang chạy (số Session là 0)	21
Hình 4.6	Trang chứa tiêu đề và đường dẫn tin tức liên quan đến mã AAA, nửa bên phải là cấu trúc các phần tử HTML tương ứng. (HTML: class= "News_Title_Link")	22
Hình 4.7	Trang nội dung 1 bài báo liên quan đến mã AAA	23
Hình 4.8	Minh họa Biểu đồ cột	42
Hình 4.9	Minh họa Biểu đồ phân tán	42
Hình 4.10	Minh họa Biểu đồ tròn (tích hợp Pylot)	43
Hình 4.11	Sơ đồ thực thi tính năng đánh giá chất lượng và sai lệch dữ liệu	45
Hình 5.1	Thống kê quá trình thu thập (với tất cả các mã).	51
Hình 5.2	Biểu đồ tròn mô tả phần trăm tình trạng thu thập (với tất cả mã)	52
Hình 5.3	Lọc những mã không có thông tin giao dịch	52
Hình 5.4	Các tệp tin chứa thông tin cân đối kế toán của các mã thu thập từ CafeF	53
Hình 5.5	Các tệp tin chứa thông tin cân đối kế toán của các mã thu thập từ VietStock	53
Hình 5.6	Terminal log trên Docker cho biết quá trình so sánh thành công	54
Hình 5.7	"log_compare.csv" cho biết dữ liệu đã được lưu thành công: "Saved" với từng mã sau khi so sánh	54
Hình 5.8	Nội dung so sánh dữ liệu cân đối kế toán từ CafeF và Viet-Stock của mã AAA	55
Hình 5.9	Thống kê tình trạng so sánh theo cột	55
Hình 5.10	Dữ liệu cuối cùng sau khi tích hợp dùng cho huấn luyện và dự báo	56

Hình 5.11 Cập nhật điểm cảm xúc kịp thời theo mã mong muốn	56
Hình 5.12 Khôi phục dữ liệu theo mong muốn khi nhấp chuột vào "Khôi phục backup"	57
Hình 5.13 Kiểm tra sai lệch phân phối trên với tập dữ liệu demo	57
Hình 5.14 Kiểm tra sai lệch phân phối trên với tập dữ liệu thực (1)	58
Hình 5.15 Kiểm tra sai lệch phân phối trên với tập dữ liệu thực (2)	59
Hình 5.16 Scatter plot giữa giá trị thực tế và giá trị dự đoán (Đơn vị: nghìn đồng)	60
Hình 5.17 Phân phối phần dư	60
Hình 5.18 Độ quan trọng đặc trưng	61
Hình 5.19 Lập lịch tự động thao tác kiểm tra-huấn luyện-dự đoán với Airflow	62
Hình 5.20 Thư mục lưu trữ dữ liệu được phân biệt với timestamp cụ thể .	64