

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN



NHÓM 3

BÁO CÁO PHÂN CÔNG CÔNG VIỆC VÀ
ĐÁNH GIÁ THÀNH VIÊN

NGÀNH: KHOA HỌC DỮ LIỆU

Thành phố Hồ Chí Minh – 2022

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN



NHÓM 3

**BÁO CÁO PHÂN CÔNG CÔNG VIỆC VÀ
ĐÁNH GIÁ THÀNH VIÊN**

| Đề tài |

**THỰC HIỆN MỘT QUY TRÌNH KHOA HỌC DỮ LIỆU VỚI TẬP
DỮ LIỆU VỀ VIỆC LÀM IT Ở VIỆT NAM**

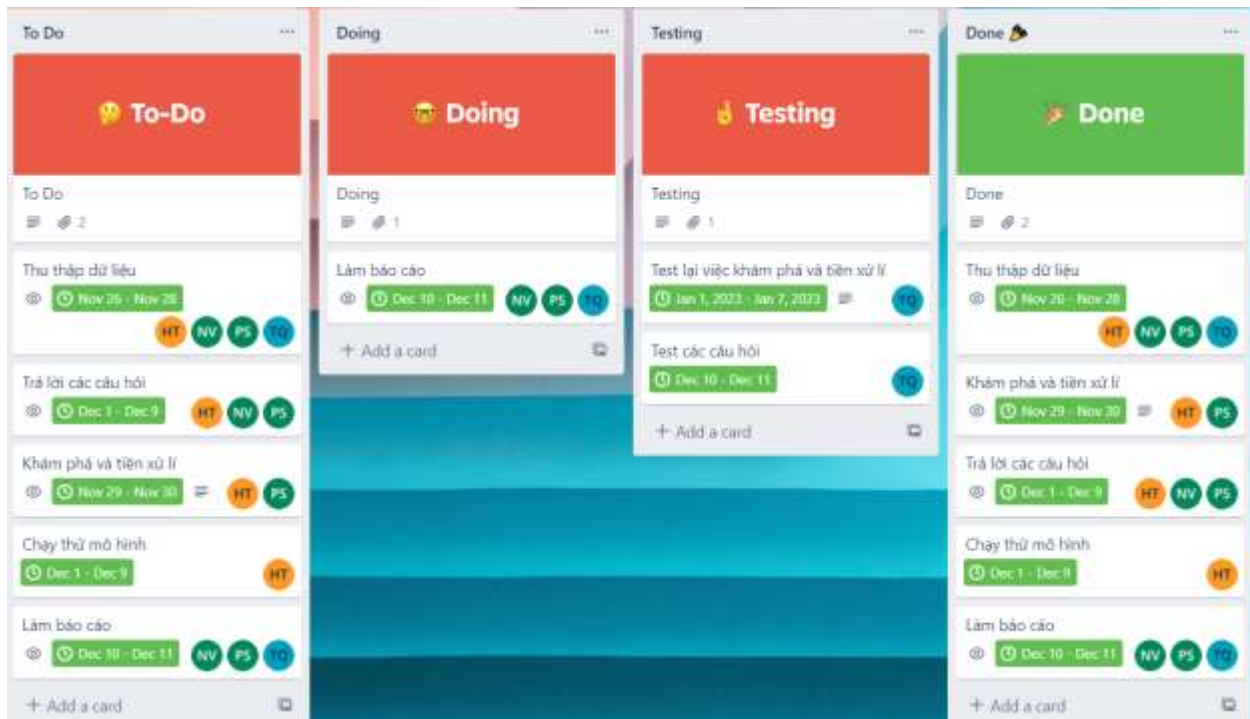
NGÀNH: KHOA HỌC DỮ LIỆU

Thành phố Hồ Chí Minh - 2020

CÁC THÀNH VIÊN TRONG NHÓM:

- PHẠM PHÚ HOÀNG SƠN – 20120366
- HÀ XUÂN TRƯỜNG - 20120391
- NGUYỄN HOÀNG VIỆT – 20120402
- TRẦN MINH QUANG – 20120559

PHÂN CHIA CÔNG VIỆC CỦA NHÓM



Simple Poll APP 8:55 AM
2 chủ đề cho LTKHDL

1 <https://www.kaggle.com/datasets/psycon/bnbusdt-2017-to-2022-historical-dataset?resource=download> 1

2 <https://www.kaggle.com/datasets/mohidabdulrehman/laptop-price-dataset> 2

3 <https://www.kaggle.com/datasets/phuc16102001/vietnam-highschool-exam-2017-to-2021> 3

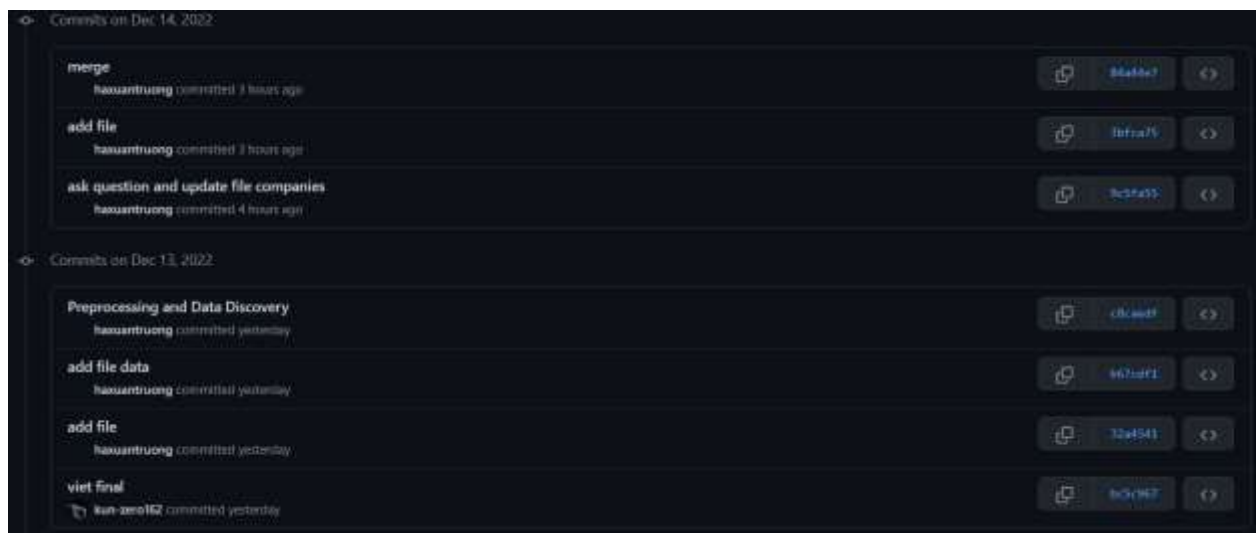
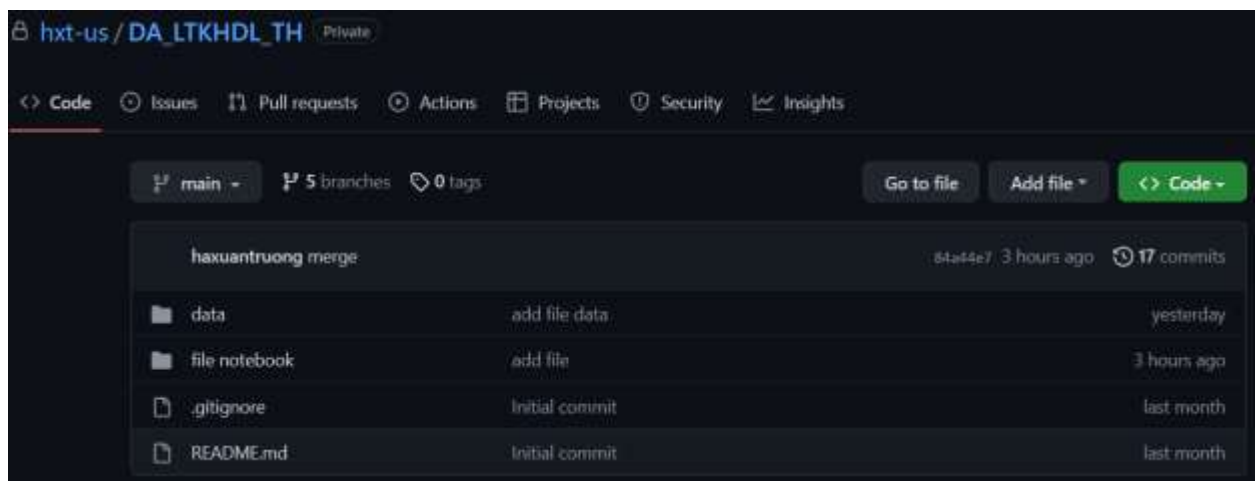
4 <https://www.kaggle.com/datasets/halhuynh/it-jobs-dataset?select=jobs.csv> 4





















1

- Thu thập dữ liệu sẽ cho cả nhóm làm và bầu chọn như hình trên
- Trả lời các câu hỏi: Phạm Phú Hoàng Sơn, Nguyễn Hoàng Việt, Hà Xuân Trường
- Khám phá và tiền xử lí: Hà Xuân Trường, Phạm Phú Hoàng Sơn
- Chạy thử mô hình: Hà Xuân Trường
- Kiểm thử toàn bộ những file notebook: Trần Minh Quang
- Báo cáo: Phạm Phú Hoàng Sơn, Nguyễn Hoàng Việt, Trần Minh Quang

Lịch sử làm việc trên github

Link: https://github.com/hxt-us/DA_LTKHDL_TH



Commits on Dec 12, 2022		
merge file visualization vao main sun-zero62 committed yesterday	 6c5b681	
remove question hasuantruong committed 2 days ago	 1f63808	
del hasuantruong committed 2 days ago	 85a7d6d	
remove questions hasuantruong committed 2 days ago	 7f4b537	
Commits on Dec 7, 2022		
Merge branch 'code_PhamPhuHoangSon' into main HoangSon-123 committed last week	 895815e	
fix exploration 2nd HoangSon-123 committed last week	 8f4c32f	
Merge pull request #1 from hxt-us/code_PhamPhuHoangSon HoangSon-123 committed last week	Verified  8ba3d6d	
Commits on Dec 1, 2022		
fix exploration 1st HoangSon-123 committed last month	 6c3c5d2	
Commits on Nov 30, 2022		
add data and notebook HoangSon-123 committed last month	 14b3e85	
Commits on Nov 14, 2022		
Initial commit hxt-us committed on Nov 14	Verified  22291a6	

BÁO CÁO TỪNG THÀNH VIÊN

Họ và tên: Phạm Phú Hoàng Sơn

MSSV: 20120366

Những khó khăn gặp phải khi làm đồ án:

- Dữ liệu dạng chuỗi dài, khó tìm được ra những điểm đặc thù của dữ liệu
- Dữ liệu gồm nhiều kí tự đặc biệt để phân chia các câu với nhau, phải dùng nhiều phương thức để có thể lọc ra được
- Dữ liệu gồm cả tiếng Việt và tiếng Anh, chính vì thế để tìm ra được một số đặc trưng khá là lâu, dành nhiều thời gian dò và đọc từng ô trong tập dữ liệu

Những điều học được thông qua đồ án lần này:

- Khả năng đặt vấn đề, câu hỏi khi nhìn vào dữ liệu.
- Khả năng xử lí những điểm bất cập của dữ liệu.
- Được nâng cao kĩ năng làm việc nhóm, quản lí mã nguồn qua github, phân chia công việc qua trello
- Tìm ra lỗi sai của bản thân khi phân tích dữ liệu và khắc phục lỗi sai đó.
- Cùng cố được những kiến thức đã học về khám phá và tiền xử lí dữ liệu.

Những việc nhóm sẽ làm khi có thêm thời gian:

- Cùng cố lại bước tiền xử lí hơn, có vẻ như dữ liệu vẫn chưa phải là được xử lí một cách tối ưu nhất.
- Có thể nghĩ và tìm ra thêm những mô hình để mô hình hóa dữ liệu tốt hơn
- Đặt thêm nhiều câu hỏi về dữ liệu hơn để trả lời, từ đó rút ra được nhiều lợi ích hơn về các dữ liệu đã thu thập này

Họ và tên: Nguyễn Hoàng Việt

MSSV: 20120402

Những khó khăn gặp phải khi làm đồ án:

- Tìm kiếm các ý tưởng cho câu hỏi cần trả lời.

- Có ý tưởng về câu hỏi nhưng không thể tìm được câu trả lời "hợp lý" từ dữ liệu.
- Dữ liệu không có dạng số nên việc đặt câu hỏi và lập model gặp khó khăn.

Những điều học được thông qua đồ án lần này:

- Có thêm kinh nghiệm trong việc khám phá và tiền xử lý dữ liệu.
- Phân tích và đặt ra các câu hỏi có ý nghĩa thực tiễn từ dữ liệu đã khám phá
- Nâng cao kỹ năng làm việc nhóm, quản lý mã nguồn qua github Tìm ra lỗi sai của bản thân khi phân tích dữ liệu và khắc phục lỗi sai đó.

Những việc nhóm sẽ làm khi có thêm thời gian:

- Tìm và thu thập dữ liệu đa dạng, phức tạp hơn, có thể thử nghiệm thu thập dữ liệu bằng cách parse HTML hoặc get API.
- Mô hình hóa dữ liệu với nhiều thuật toán khác.
- Đặt thêm nhiều câu hỏi về dữ liệu hơn để trả lời.

Họ và tên: Trần Minh Quang

MSSV:20120559

Những khó khăn gặp phải khi làm đồ án:

- Dữ liệu dạng chuỗi, không có cấu trúc đồng nhất nên khó khăn trong việc lọc ra các thành phần cần thiết cho việc khám phá dữ liệu
- Khó khăn trong việc đặt ra các câu hỏi để khám phá dữ liệu mang lại những thông tin có ích

Những điều học được thông qua đồ án lần này:

- Học thêm được nhiều thao tác để tiền xử lý dữ liệu một cách hiệu quả và khám phá chúng
- Khả năng code để xử lý dữ liệu
- Khả năng nhìn nhận và phân tích khi lần đầu làm việc với dữ liệu lớn
- Khả năng làm việc nhóm, sử dụng github và trello

Những việc nhóm sẽ làm khi có thêm thời gian:

- Xử lý các dữ liệu dạng chuỗi chi tiết hơn để dễ dàng cho việc lọc dữ liệu và mô hình hóa chúng một cách trực quan.
- Nghiên cứu dữ liệu để có thêm các câu hỏi mang lại những thông tin có ích và giải quyết những câu hỏi đó
- Khám phá thêm nhiều dữ liệu khác để học thêm các thao tác xử lý dữ liệu và trực quan hóa dữ liệu.

Họ và tên: Hà Xuân Trường

MSSV: 20120391

Những khó khăn gặp phải khi làm đồ án:

- Dữ liệu gồm nhiều file.
- Dữ liệu gồm quá nhiều file dạng chuỗi, khó xử lý hơn dạng số
- Chưa có nhiều kinh nghiệm khi viết các mô tả và nhận xét về câu hỏi. Gặp vấn đề trong việc thực hiện model do toàn bộ là chuỗi
- Các dữ liệu dài, khó đọc hết được bằng mắt thường để kiểm tra dữ liệu trước

Những điều học được thông qua đồ án lần này:

- Trải nghiệm được làm việc theo quy trình khoa học dữ liệu một cách chặt chẽ
- Có nhiều kinh nghiệm hơn khi làm việc với git/github
- Khả năng đặt câu hỏi cho dữ liệu thành thạo hơn.
- Củng cố kiến thức matplotlib và pandas.
- Nâng cao khả năng làm việc nhóm

Những việc nhóm sẽ làm khi có thêm thời gian:

- Thực hiện chạy model dữ liệu, do là dạng văn bản kí tự nên việc xử lý dính tới xử lý ngôn ngữ tự nhiên, điều này vượt qua năng lực xử lý của nhóm trong thời gian này.
- Tối đa phân khai thác dữ liệu và khám phá một cách cặn kẽ dữ liệu hơn, tệp dữ liệu còn nhiều thứ khai thác được
- Đặt thêm câu hỏi để giải quyết.