ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN THỰC HÀNH

|Đề tài|

THỰC HIỆN QUY TRÌNH KHOA HỌC DỮ LIỆU VÀ MÔ HÌNH HÓA DỮ LIỆU

|Trợ giảng phụ trách đồ án|

Trần Đại Chí
Nguyễn Bảo Long
Lê Nhựt Nam
Nguyễn Thái Vũ

Môn học: Nhập môn Khoa học dữ liệu

Thành phố Hồ Chí Minh - 2022

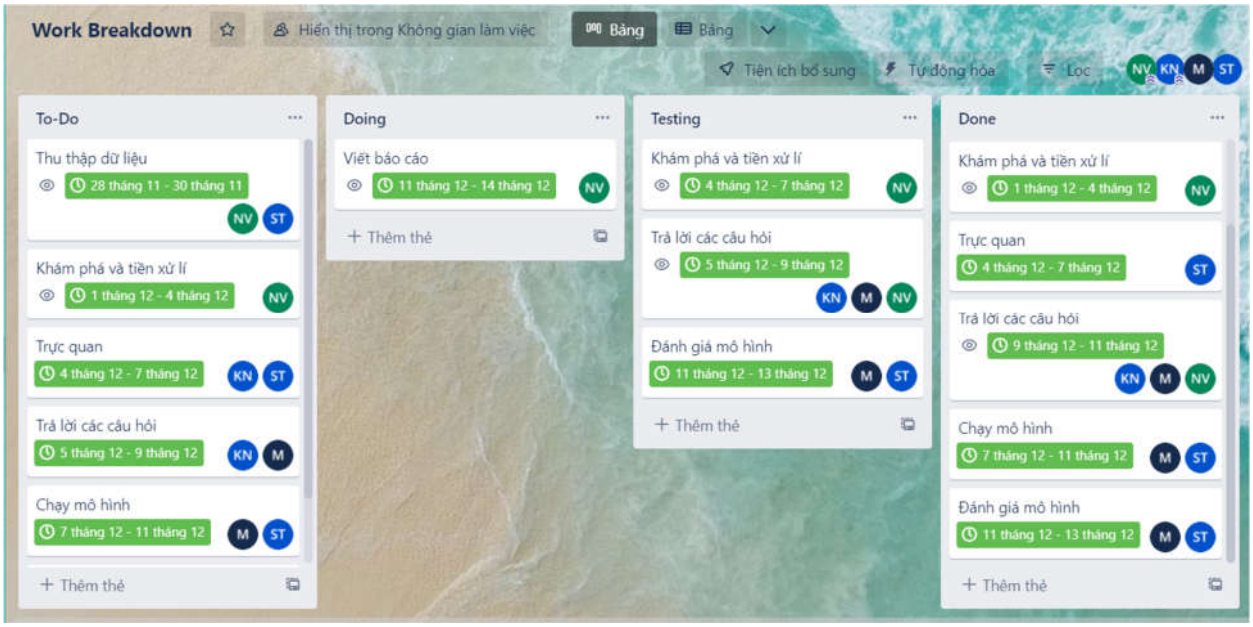
THÔNG TIN THÀNH VIÊN

MSSV	Họ tên
20120357	Nguyễn Đức Minh Quân
20120402	Nguyễn Hoàng Việt
20120120	Nguyễn Việt Khoa
20120565	Nguyễn Tấn Sơn

MỤC LỤC

Nội dung

PHÂN CHIA CÔNG VIỆC NHÓM



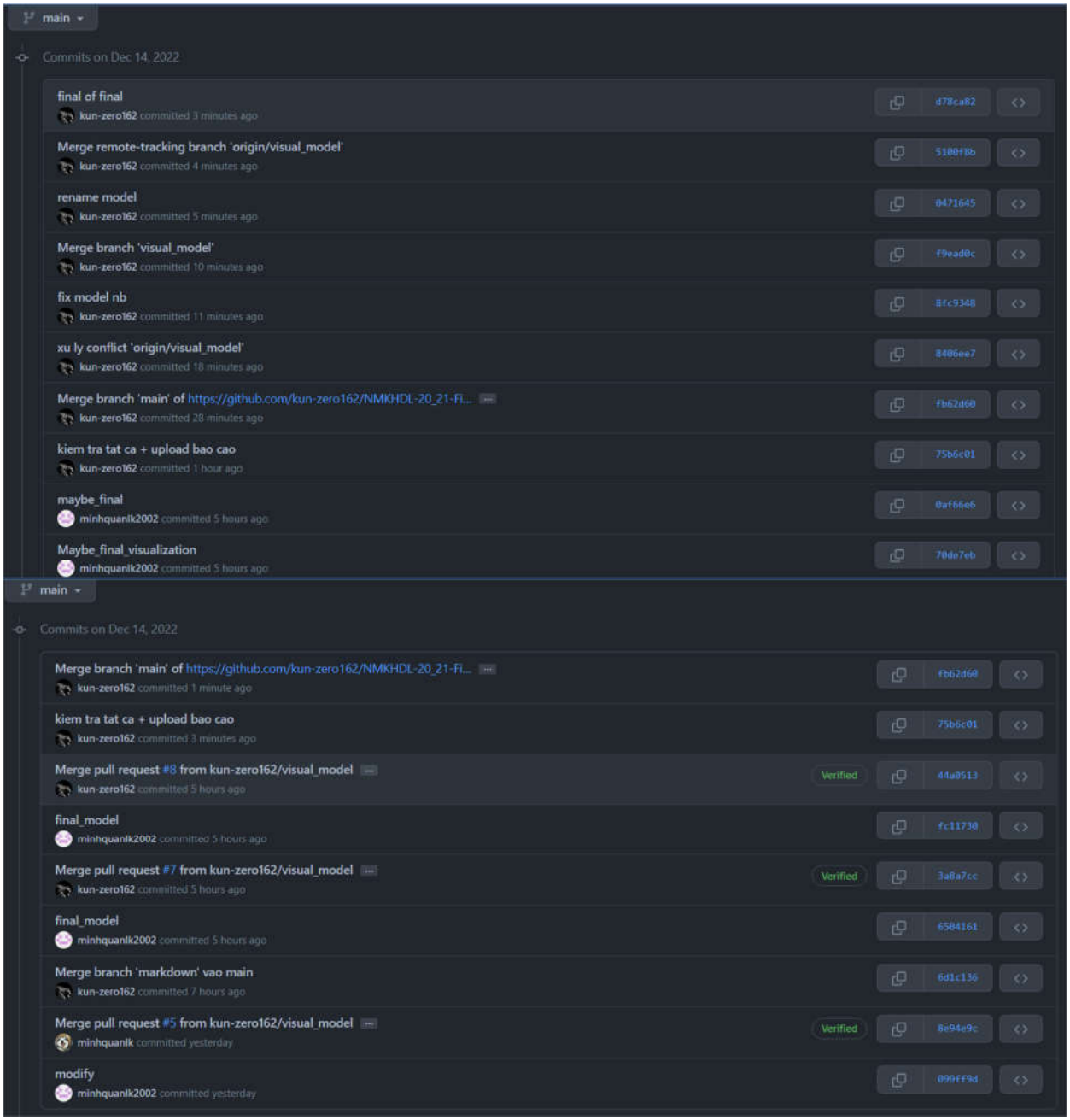
Việc xác định dữ liệu để thu thập và khám phá do cả nhóm bỏ phiếu và chọn ra. Công việc cụ thể:

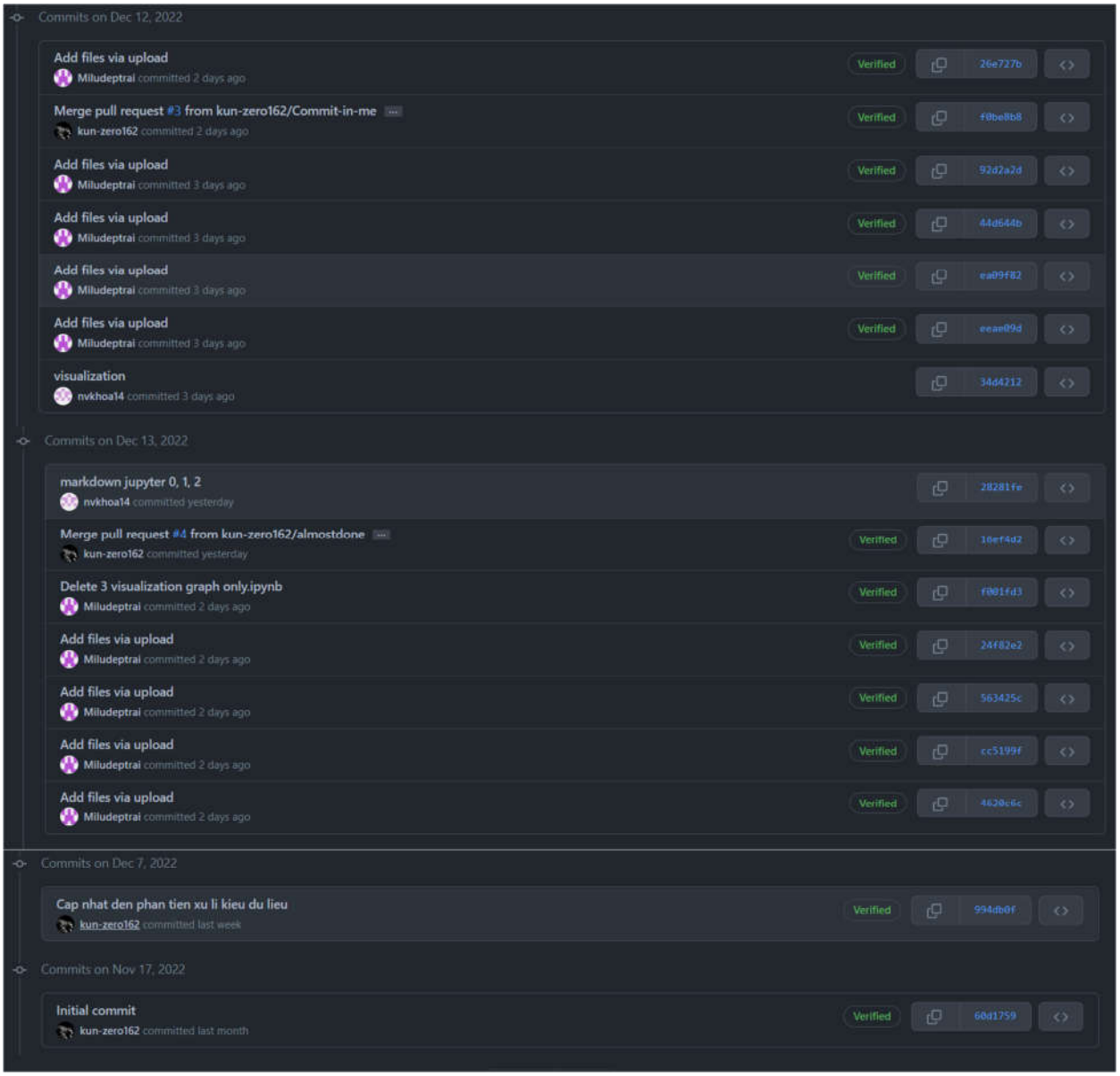
- Nguyễn Hoàng Việt: thu thập dữ liệu, khám phá và tiền xử lí; kiểm tra tính đúng đắn cho các câu trả lời của phần đặt câu hỏi; phân chia công việc cho nhóm trên Trello và viết báo cáo.
- Nguyễn Tấn Sơn: thu thập dữ liệu, trực quan hóa dữ liệu và chạy mô hình.
- Nguyễn Việt Khoa: tham gia trực quan hóa dữ liệu, đặt ra các câu hỏi và trả lời các câu hỏi đó.
- Nguyễn Đức Minh Quân: Đặt ra các câu hỏi, chạy mô hình và đề xuất phương hướng chạy mô hình. Trả lời các câu hỏi, trực quan hoá và đánh giá mô hình.

LỊCH SỬ LÀM VIỆC TRÊN GITHUB

Link repository Github: https://github.com/kun-zero162/NMKHDL-20_21-Final-Project

Lịch sử commit:





BÁO CÁO TỪNG THÀNH VIÊN

Họ và tên: Nguyễn Hoàng Việt

MSSV: 20120402

Những khó khăn gặp phải khi làm đồ án:

- Gặp khó khăn trong việc tìm kiếm ý tưởng cho các câu hỏi.
- Có ý tưởng về câu hỏi nhưng không thể tìm được câu trả lời "hợp lí" từ dữ liệu.

Những điều học được thông qua đồ án lần này:

- Có thêm kinh nghiệm trong việc khám phá và tiền xử lý dữ liệu.
- Biết thêm nhiều thư viện cũng như công cụ bổ ích trong quá trình trực quan và mô hình hóa dữ liệu.
- Phân tích và đặt ra các câu hỏi có ý nghĩa thực tiễn từ dữ liệu đã khám phá
- Nâng cao kĩ năng làm việc nhóm, quản lí 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 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: Nguyễn Việt Khoa

MSSV: 20120120

Những khó khăn gặp phải khi làm đồ án:

- Có ý tưởng về câu hỏi nhưng chưa tiếp cận thêm về dữ liệu.
- Khó khăn trong xây dựng và giải thích biểu đồ.

Những điều học được thông qua đồ án lần này:

- Tìm hiểu thêm được nhiều biểu đồ hay và hiệu quả trong trực quan hóa.
- Đào sâu phân tích thêm dữ liệu để trả lời câu hỏi.
- Nâng cao kĩ năng làm việc nhóm qua trello, quản lí mã nguồn thông qua github.

Những việc sẽ làm khi có thêm thời gian:

- 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: Nguyễn Tấn Sơn

MSSV: 20120565

Những khó khăn gặp phải khi làm đồ án:

- Hiệu chỉnh selector và xpath cho đến khi lấy được data, fix các trường hợp đặc biệt.
- Khó khăn trong việc tìm cách thay giá trị thiếu/nhiễu.

- Thư viện plotly tương đối là mới mẻ đối với nhóm và tài liệu về thư viện này cũng khó có thể tiếp cận.
- Thiếu cột giá trị binary dùng để phân lớp, phải dùng html request trên 1 trang khác.
- Các siêu tham số như Binary cross entropy loss khiến việc lập công thức khó khăn.
- Đánh giá RMSE cho cả bảng dữ liệu khá khó khăn khi các giá trị ở mỗi cột có đơn vị đo khác nhau.

Những điều học được thông qua đồ án lần này:

- Biết được cách copy selector và xpath trong browser.
- Cách dùng thư viện plotly.
- Học được nhiều kiến thức về việc lập model, thư viện sklearn, đánh giá model.

Những việc sẽ làm khi có thêm thời gian:

- 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: Nguyễn Đức Minh Quân

MSSV: 20120357

Những khó khăn khi thực hiện đồ án:

- Thường thì câu hỏi sẽ được hình thành dựa trên khả năng trả lời của dữ liệu. Chắc chắn rằng có thể trả lời thì mới đặt câu hỏi đó. Nên hạn chế trong việc đưa ra những câu hỏi mang tính "khám phá dữ liệu" thực sự, câu hỏi đề xuất còn thấy hơi đơn giản.
- Khi thực hiện đặt câu hỏi, đặc biệt là mô hình dự đoán, phải đi tìm hướng giải và bị hạn chế tầm kiến thức nên chưa tối ưu được đề xuất.
- Kiến thức về hồi quy vẫn chưa vững

Những điều học được thông qua đồ án lần này:

- Học hỏi được qua bạn bè những kiến thức hay và các kĩ thuật, các cú pháp trong python.
- Khám phá nhiều điều thú vị liên quan đến dân số thế giới như "Các quốc gia đáng sống" và các yếu tố sẽ quyết định điều đó.
- Bắt đầu làm quen được với cách đặt ra giả thuyết và tiến hành mô hình hoá dữ liệu.
- Nắm vững lại kiến thức về Hồi quy và thuật toán.

Những việc sẽ làm khi có thêm thời gian:

- Sẽ đưa ra thêm mô hình dự đoán cho dữ liệu. Ví dụ như: Liệu trong tương lai có đất nước nào soán ngôi được Trung Quốc dẫn đầu về dân số.
- Làm kỹ hơn khâu phân lớp và đưa ra những giả thuyết hay vê tình hình thế giới như sự phát triển, liên quan tới kinh tế và cả chính trị.
- Vì dân số có rất nhiều vấn đề liên quan, ví dụ như tỉ lệ nghề nghiệp, mật độ giao thông, mức độ ô nhiễm và vấn đề mật độ dân số, đất ở…