TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

KHOA CÔNG NGHỆ THÔNG TIN

MÔN: LẬP TRÌNH CHO KHOA HỌC DỮ LIỆU

LỚP: 21KHDL1

KẾ HOẠCH THỰC HIỆN ĐỒ ÁN CUỐI KỲ

NHÓM: 10 - ORIGINI

DANH SÁCH THÀNH VIÊN

HỌ VÀ TÊN	MSSV
TRẦN ĐÌNH QUANG	21127406
VŨ MINH PHÁT	21127739

Link GitHub

Link Google Drive

MỤC LỤC

I. THỜI GIAN THỰC HIỆN

- Thời gian bắt đầu làm đồ án: ngày 16/10/2023.
- Thời gian nộp đồ án: ngày 26/12/2023.

II. MÔ TẢ CHI TIẾT TỪNG GIAI ĐOẠN VÀ PHÂN CHIA CÔNG VIỆC CHO CÁC THÀNH VIÊN

1. **Giai đoạn 1:** Kéo dài 2 tuần từ ngày 16/10 - 29/10

Do việc thành lập nhóm diễn ra khá trễ nên các tập dữ liệu phổ biến (như: âm nhạc, phim ảnh, v.v.) đều có các nhóm khác lựa chọn.

Khi này, mỗi thành viên sẽ tìm kiếm các tập dữ liệu mà mình hứng thú với điều kiện:

- Không được nằm trong "danh sách cấm" mà giảng viên lý thuyết đã đưa ra.
- Không được trùng chủ đề với các nhóm khác.

Kết quả: Cả nhóm nhất trí lựa chọn tập dữ liệu "Google Analytics Customer Revenue Prediction" được cung cấp bởi Kaggle.

2. **Giai đoạn 2:** Kéo dài 4 tuần từ ngày 30/10 – 26/11

Đầu tiên, mỗi thành viên cần tự tìm hiểu về dữ liệu và viết một notebook riêng để phác thảo quy trình khám phá và tiền xử lý dữ liệu đã chọn. Sau đó, cả nhóm sẽ dựa trên bản phác thảo này để phát triển các công đoạn tiếp theo.

Về bản phác thảo quy trình khám phá và tiền xử lý dữ liệu, mỗi thành viên có thể tùy ý lựa chọn, sắp xếp thứ tự các bước mình muốn thực hiện nhưng BẮT BUỘC phải trả lời được đầy đủ các câu hỏi đã cập trong slide "Final Project.pdf" mà giảng viên cung cấp.

Mỗi thành viên có nên dành thêm thời gian để tìm hiểu các notebook hay ở trên mạng từ GitHub, Kaggle, v.v. để có thể hiểu thêm về các điểm "đặc biệt" trong tập dữ liệu mà đòi hỏi các bước xử lý chuyên biệt. Đồng thời có thể tham khảo ý tưởng tiền xử lý của họ để biến đổi và áp dụng cho bài làm của bản thân mình.

Sau khi mỗi thành viên hoàn tất bản phác thảo của mình thì cả nhóm có thể xem xét nên giữ ý tưởng nào, nên bổ sung cái gì để hoàn thiện bài làm chung của cả nhóm. Khi này cũng yêu cầu phải tạo ra một repo trên GitHub để các thành viên có thể làm việc chung với nhau.

Kết quả:

Ngày 23/11, sau cuộc họp, các ý tưởng phác thảo về quy trình khám phá và tiền xử lý dữ liệu của thành viên Vũ Minh Phát được cả nhóm thông qua. Tuy chỉ là một bản phác thảo nhưng các đoạn code trong file cũng khá đầy đủ về xử lý, trực quan hóa dữ liệu.

Tuy nhiên, notebook còn thiếu rất nhiều cell markdown để mô tả đoạn code cần làm gì, hay nhận xét về kết quả trả về, các biểu đồ có hơi "xấu" nên cần phải chỉnh sửa thêm để người dùng có thể dễ dàng nắm bắt các thông điệp được truyền tải.

Bên cạnh đó, cả nhóm cũng quyết định dùng repo sẵn có của thành viên Vũ Minh Phát làm repo chính của cả nhóm vì nó có bộ khung (skeleton) khá đầy đủ. Khi này, từ một nhánh main, repo xuất hiện thêm hai nhánh phụ chứa công việc của hai thành viên, mở ra một giai đoạn 3 với hai luồng công việc song song chia đều cho hai thành viên.

3. **Giai đoạn 3:** Kéo dài 2 tuần từ ngày 27/11 – 10/12

Thành viên	Nội dung công việc	Kết quả
Vũ Minh Phát	- Hoàn thiện, chỉnh sửa lại notebook 02 theo những góp ý từ thành viên Trần Đình Quang. - Sau đó, viết notebook 01 để trình bày về quá trình "thu thập dữ liệu", bao gồm: mô tả về bài toán trên Kaggle và sơ lược về tập dữ liệu.	- Hoàn thiện khá tốt notebook 02 theo những góp ý của thành viên trong nhóm. - Notebook 01 chưa đề cập đến cách tác giả thu thập dữ liệu và chưa kiểm tra "data license".
Trần Đình Quang	- Đặt từ 3 – 4 câu hỏi để thực hiện phân tích khám phá dữ liệu (EDA).	- Đặt ra 3 câu hỏi để hiểu thêm về dữ liệu. Sau đó tiền xử lý và phân tích dữ liệu để trả lời cho 3 câu hỏi đó. - Câu hỏi số 2 có vẻ hơi đơn giản nên cần chỉnh sửa lại một ít.

4. **Giai đoạn 4:** Kéo dài 1 tuần từ ngày 11/12 – 17/12

Thành viên	Nội dung công việc	Kết quả
Vũ Minh Phát	- Đặt thêm từ 1-2 câu hỏi trong notebook 03.	- Đặt thêm 2 câu hỏi và phân tích dữ liệu để trả lời cho 2 câu hỏi này.
Trần Đình Quang	- Chỉnh sửa lại các câu hỏi đã đặt ra trước đó theo góp ý của nhóm. - Tìm hiểu và trả lời các thông tin còn thiếu trong notebook 01 – thu thập dữ liệu.	- Đặt lại câu hỏi #2 và phân tích dữ liệu để trả lời câu hỏi đó. - Tìm thấy "data license" từ Kaggle và bổ sung vào phần "Thu thập dữ liệu".

5. **Giai đoạn 5:** Kéo dài 1 tuần từ ngày 18/12 – 24/12

Đây là tuần cuối cùng để thực hiện đồ án, đòi hỏi các thành viên nêu cao tính tự giác để hoàn thành các công việc còn thiếu. Khi này các thành viên có trách nhiệm xem xét lại toàn bộ bài làm của mình và chỉnh sửa nếu cần.

Sau đó, các thành viên sẽ tổng kết lại đồ án theo như mục "5. Reflection" trong slide "Final Project.pdf". Và bổ sung các tài liệu tham khảo vào phần "References".

Thành viên	Nội dung công việc
Vũ Minh Phát	- Viết TOC cho các notebook. - Hoàn thiện các report. - Tổng kết đồ án theo như mục "5. Reflection". - Bổ sung các tài liệu tham khảo vào phần "References".
Trần Đình Quang	- Hoàn thiện file README.md - Chỉnh sửa lại các đoạn code để nó chạy nhanh hơn. - Tổng kết đồ án theo như mục "5. Reflection". - Bổ sung các tài liệu tham khảo vào phần "References".