NHẬP MÔN KHOA HỌC DỮ LIỆU – 22KHDL2

PROJECT PROPOSAL

1. Thông tin nhóm

Student ID	Full Name	Email
21127072	Nguyễn Hữu Khánh	nhkhanh21@clc.fitus.edu.vn
21127160	Nguyễn Thanh Sơn	ntson21@clc.fitus.edu.vn
21127246	Lê Minh Đức	nhkhanh21@clc.fitus.edu.vn

2. Chủ đề chính

Dự báo thời tiết là một lĩnh vực không mới nhưng luôn luôn cần thiết trong đời sống thường nhật và các hoạt động sản xuất, kinh doanh, văn hoá, văn nghệ. Chúng ta luôn cần dự báo thời tiết để ra quyết định bắt đầu hay trì hoãn một hoạt động nào đó. Từ việc ra khơi đánh bắt cá, gieo hạt cho mùa vụ, tưới tiêu, cho đến những sự kiện lớn như những concert, thế vận hội… Ngày nay, nhu cầu phân tích dữ liệu thời tiết không chỉ dừng lại ở việc dự đoán thời tiết tốt hay xấu, nắng nóng hay mưa bão mà ta còn có nhu cầu phân tích dữ liệu để đánh giá mức ảnh hưởng của thời tiết đến sức khoẻ con người. Trước thực trạng môi trường càng ngày càng ô nhiễm do tác động của con người, dẫn đến hệ luỵ biến đổi khí hậu thì việc phân tích tác động của thời tiết gây ảnh hưởng đến sức khoẻ con người và đưa ra các cảnh báo là vô cùng cần thiết.

Trong đồ án lần này, nhóm sẽ tiến hành phân tích dữ liệu thời tiết được thu tập tại trạm khí tượng thuỷ văn Toronto City, Canada.

3. Mô tả dữ liệu

Tập dữ liệu có hơn 83.000 dòng được thu thập bởi trạm khí tượng thuỷ văn thành phố Tonronto Canada. Mỗi dòng trong tập dữ liệu sẽ chứa giá trị đo đạc được tại trạm trong một khung giờ của một ngày cụ thể *(Ví dụ: dòng đầu của tập dữ liệu sẽ chứa các giá trị đo đạc được tại trạm lúc 0:00 ngày 1/1/2015)*. Và với hơn 83.000 dòng, nhóm đã tiến hành thu thập dữ liệu từ ngày 1/1/2015 đến 31/5/2024. Chi tiết trong các đường link bên dưới:

- Website gốc dữ liệu.
- Tập dữ liệu.
- Source code thu thập dữ liệu từ trang web (trích xuất từ html).

4. Mục tiêu của dự án

Về tổng quan, dự án này là cơ hội tốt nhất để nhóm có thể áp dụng các kiến thức lý thuyết được học trong môn Nhập môn Khoa học dữ liệu vào thực tiễn. Qua đó, việc thực hiện dự án một cách nghiêm túc sẽ giúp nhóm có thể hiểu, vận dụng và cải thiện các kiến thức, kỹ năng xoay quanh quá trình phân tích dữ liệu, bao gồm hiểu biết về quy trình thực hiện phân tích dữ liệu; các kỹ thuật thu thập dữ liệu từ trang web thực tế với Selenium, BeautifulSoup4…; các cách thức tiếp cận, khám phá dữ liệu trong đó có các kiến thức về trực quan hoá dữ liệu; kỹ thuật tiền xử lí dữ liệu; kỹ thuật mô hình hoá dữ liệu để giúp phân loại, dự đoán… Song song đó, việc lựa chọn chủ đề về khí tượng thuỷ văn cũng sẽ giúp cho các thành viên của nhóm có dịp tự học, tự tìm hiểu thêm các kiến thức về khí tượng thuỷ văn (bổ sung domain knowledge).

Về đầu ra sản phẩm làm việc nhóm, nhóm ít nhất phải thực hiện được toàn bộ các mục được giảng viên đề ra và ở mức độ chấp nhận được. Cụ thể, với tập dữ liệu thu thập được, sản phẩm cuối cùng mà nhóm cần hoàn thành phải bao gồm:

- Định hình được đúng và đủ các vấn đề (các meaningful questions) cần giải quyết và có thể được giải quyết bởi dữ liệu.
- Tập dữ liệu đã được tiền xử lí.
- Các mô hình phục vụ giải quyết các bài toán được đặt ra (ví dụ: phân loại thời tiết, dự báo thời tiết, phân tích mức độ ảnh hưởng đến sức khoẻ con người,…). Các mô hình này cần được hoàn thiện kỹ lượng và phải có mức chính xác cao (~80% trở lên), sai số thấp nhất có thể.
- Báo cáo tổng kết.

5. Kế hoạch thực hiện đồ án.

Tuần	Ngày	Tiến độ cần hoàn thành
Tuần 2	6/10 – 13/10	- Chọn đề tài, tìm kiếm nguồn dữ liệu
Tuần 3	13/10 – 20/10	- Nghiên cứu cấu trúc trang nguồn, thu thập dữ liệu từ nguồn với thư viện python.
Tuần 4	20/10 – 27/10	
Tuần 5	27/10 – 3/11	- Khám phá dữ liệu: o Giải thích ý nghĩa từng thuộc tính. o Xác định kiểu dữ liệu từng cột. o Lược bỏ các cột có kiểu dữ liệu không phù hợp, hoặc không mang nhiều ý nghĩa cho việc phân tích. o Trực quan hoá dữ liệu để quan sát phân phối và các đặc tính của dữ liệu. - Đặt ra các vấn đề cần thiết và có thể giải quyết bằng việc kết hợp dữ liệu được thu thập với các kiến thức ngành tương ứng, và phương pháp mô hình hoá.
Tuần 6	3/11 – 10/11	- Nghiên cứu và đưa ra các phương pháp, chiến lược phù hợp để giải quyết từng vấn đề.
Tuần 7	10/11 – 17/11	- Thực hiện giải quyết các câu hỏi được đặt ra với tập dữ liệu: o Tiền xử lí dữ liệu. o Mô hình hoá. o Giải quyết vấn đề bằng kiến thức về khí tượng thuỷ văn.
Tuần 8	17/11 – 24/11	
Tuần 9	24/11 – 1/12	- Tổng hợp, phân tích và đánh giá các mô hình đã được huấn luyện, các giải pháp đã được đề ra. - Chọn ra mô hình tốt nhất và bộ dữ liệu với chiến lược xử lí tốt nhất.
Tuần 10	1/12 – 8/12	- Thực hiện báo cáo tổng kết và nộp sản phẩm làm việc nhóm.