KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện thoại: (08) 38.354.266 – Fax: (08) 38.350.096

LAB

THỐNG KÊ VÀ TRỰC QUAN HÓA DỮ LIỆU

Biên soạn:
Lê Ngọc Thành

1. Nội dung

Thống kê và trực quan hóa dữ liệu để tìm mối quan hệ giữa các trường dữ liệu thực tế.

2. Yêu cầu

Thời gian và cách thức nộp, xem trên Moodle.

Nội dung cần nộp:

- Báo cáo trình bày trong file .doc/.docx/pdf chứa:
  - Mức độ hoàn thành tổng thể của mỗi yêu cầu.
  - Chi tiết thuật toán, chạy ví dụ, nhận xét.
    Khuyến khích trình bày đơn giản, có hình minh họa.
- Souce code kèm hướng dẫn chạy nếu thực hiện trong môi trường khác Jupyter Notebook hoặc python gốc.
- Dataset được lấy gốc theo từng ngày, nếu có modify thì tạo file riêng.
- Ngôn ngữ lập trình bắt buộc: Python
  - Cho phép sử dụng các thư viện đã được giới thiệu trong lý thuyết.

3. Yêu cầu chi tiết

Report coronavirus cases

Now Yesterday Search:

All Europe North America Asia South America Africa Oceania

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	4,885,970	+86,704	319,878	+3,358	1,903,201	2,662,891	44,754	627	41.0			
1	USA	1,549,359	+21,695	91,955	+977	354,340	1,103,064	16,857	4,684	278	12,255,697	37,052	330,769,370
2	Russia	290,678	+8,926	2,722	+91	70,209	217,747	2,300	1,992	19	7,147,014	48,977	145,927,122
3	Spain	278,188	+469	27,709	+59	196,958	53,521	1,152	5,950	593	3,037,840	64,977	46,752,654
4	Brazil	254,220	+13,140	16,792	+674	100,459	136,969	8,318	1,197	79	735,224	3,462	212,376,810
5	UK	246,406	+2,711	34,796	+160	N/A	N/A	1,559	3,632	513	2,682,716	39,543	67,843,268
6	Italy	225,886	+451	32,007	+99	127,326	66,553	749	3,735	529	3,041,366	50,294	60,472,166
7	France	179,927	+358	28,239	+131	61,728	89,960	1,998	2,757	433	1,384,633	21,218	65,256,433
8	Germany	177,289	+638	8,123	+74	154,600	14,566	1,133	2,117	97	3,147,771	37,584	83,752,125
9	Turkey	150,593	+1,158	4,171	+31	111,577	34,845	903	1,788	50	1,650,135	19,591	84,227,597
10	Iran	122,492	+2,294	7,057	+69	95,661	19,774	2,294	1,461	84	701,640	8,367	83,859,705
11	India	100,340	+4,642	3,156	+131	39,233	57,951		73	2	2,302,792	1,671	1,378,344,732
12	Peru	94,933	+2,660	2,789	+141	30,306	61,838	866	2,884	85	661,132	20,086	32,914,644

Hình 1. Dữ liệu thống kê từng ngày ca nhiễm virus Covid-19 từ tổ chức Worldometer

Từ khoảng cuối năm 2019 đến nay, một bệnh dịch hạch lan tràn khủng khiếp trên toàn thế giới. Mỗi ngày có hàng ngàn người bị nhiễm và hàng chục đến hàng trăm người chết. Tổ chức Worldometer (www.worldometers.info) đã thu thập dữ liệu thống kê từ nhiều nguồn và từ nhiều quốc gia báo cáo hàng ngày để tổng hợp thành một bảng trong Hình 1. Trong trang web, tổ chức Worldmeter cũng thực hiện vẽ biểu đồ để cho thấy sự thay đổi trực quan tình hình diễn biến dịch bệnh. Tuy nhiên chúng ta tạm thời không sử dụng nó.

Bạn được quốc gia giao trọng trách để tìm hiểu dữ liệu này như giữa các trường dữ liệu có mối quan hệ gì không, liệu có bất thường trong dữ liệu hay không như báo cáo quốc gia khác với dữ liệu tổng hợp, sự bất bình thường trong việc nhảy số liệu, …. Hãy vận dụng các kiến thức về thống kê và trực quan hóa dữ liệu để hiểu về dữ liệu đang có.

Cụ thể trong lab này, bạn được yêu cầu thực hiện các nhiệm vụ sau:

- Thu thập số liệu thống kê từng ngày từ trang Worldmeter.
    - Bạn có thể chọn làm trên 1 ngày xác định. Do trang Worldometer chỉ thể hiện ngày hôm nay và ngày hôm qua nên bạn nên thu thập nhiều ngày để có thể có được thống kê, trực quan tốt hơn.
    - Bạn có thể thủ công để chép dữ liệu và lưu trữ vào định dạng chuẩn .CSV hoặc sử dụng code để lấy dữ liệu (khuyến khích)
    - Bạn có thể tiền xử lý dữ liệu trước khi chuyển sang pha tiếp theo nhưng cần báo cáo vấn đề này trong mục *Tiền xử lý dữ liệu*. Dữ liệu gốc và dữ liệu đã điều chỉnh cần lưu lại và nộp kèm trong bài nộp.
- Sử dụng nhận xét, code/thuật toán để thể hiện thống kê, trực quan các mối quan hệ giữa các trường dữ liệu
    - Các thông kê về dữ liệu cần mô tả ý nghĩa, điều gì đã rút ra được từ đó.
    - Thảo luận và chọn ra các trường dữ liệu để thể hiện trực quan bằng các loại biểu đồ đã học.
    - Việc chọn biểu đồ cần giải thích tính phù hợp với tính chất trường dữ liệu. Có thể sử dụng nhiều hơn 1 loại biểu đồ cho trường dữ liệu nhưng cần giải thích lí do.
    - Việc thể hiện quan hệ phải tích hợp dần dần nghĩa là từ đơn giản đến phức tạp, từ một trường đơn đến quan hệ giữa nhiều trường, …
    - Ngoài quan hệ độc lập, cần xem xét liệu trong dữ liệu có quan hệ nhân quả không (cause-effect). Ví dụ: liệu có thể có mối quan hệ giữa tỉ lệ ca nhiễm tăng với số ca chết không, … Cần chứng minh thông qua các phép trực quan dữ liệu.
    - Số lượng tối thiểu biểu đồ trực quan là 6 biểu đồ. Lưu ý đây là số tối thiểu và đánh giá có thể chỉ dừng ở mức 60%.

4. Qui định

- Bài không có báo cáo sẽ không chấm.
- Các nguồn tài liệu tham khảo (nếu có) cần ghi đầy đủ trong báo cáo ở mục *Tài liệu tham khảo*. Lưu ý cần phân biệt giữa tham khảo và đạo văn.
- Nếu kích thước dữ liệu >20MB thì upload lên server ngoài như Google Drive, …, nộp link và giữ link public ít nhất trong 2 năm. Báo cáo cần để trên Moodle.
- Bài giống nhau sẽ 0 điểm môn học.

5. Liên hệ

Mọi thắc mắc trong quá trình thực hiện vui lòng gửi mail về lnthanh@fit.hcmus.edu.vn