**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

A picture containing text, device, gauge

Description automatically generated

**TRẦN HOÀNG NAM**

**19127481**

**BÁO CÁO ĐỒ ÁN**

**VISUALIZATION**

**Chuyên ngành: Khoa Học Dữ Liệu**

Thành phố Hồ Chí Minh – 2021

1. **MỨC ĐỘ HOÀN THÀNH**

Thu thập dữ liệu (50/50): Đã hoàn toàn thu thập dữ liệu thành công trong khoảng ngày 28/11 – 9/12

Visualization

|  |  |
| --- | --- |
| Mô tả xu hướng | Sử dụng line chart(5 biểu đồ) |
| So sánh số lượng người chết ở các khu vực | Sử dụng bar chart (1 biểu đồ) |
| Xem xét tỉ lệ Deaths/Total case các khu vực | Sử dụng pie chart (7 biểu đồ) |
| Xem xét mối quan hệ nhân quả của tỉ lệ test và tỉ lệ người chết | Sử dụng bar chart(3 biểu đồ) |

1. **THU THẬP DỮ LIỆU**

Chúng ta sẽ thu thập dữ liệu từ trang web <https://www.worldometers.info/coronavirus/>, dữ liệu chúng ta thu thập sẽ là bảng Reported Cases and Deaths by Country or Territory bằng file PARSE\_HTML notebook

Graphical user interface, application

Description automatically generated

* Dữ liệu chúng ta thu thập được sẽ được lưu vào file theo định dạng ‘ngày-tháng-năm.csv’
* Chúng ta sẽ tiến hành thu thập dữ liệu ngày hôm trước của ngày hiện tại, nguyên nhân là số liệu hiện tại trong ngày có thể bị thay đổi theo gian, do đó, để đảm bảo dữ liệu không thay đổi theo thời gian, chúng ta sẽ tiến hành thu thập dữ liệu của ngày trước đó.

Ví dụ hôm nay là ngày 10/12/2021 thì chúng ta sẽ thu thập dữ liệu của ngày 9/12/2021

* Kết quả dữ liệu thu thập được sẽ nằm trong thư mục Data

Sau khi thu thập dữ liệu bằng file PARSE\_HTML, chúng ta sẽ dùng file PHANTICH để visualize

1. **VISUALIZE**
2. **Xu hướng của số lượng ca bệnh**

Đầu tiên, chúng ta sẽ bắt đầu xem xu hướng của số ca active hiện nay, tổng số lượng ca mới trên thế giới

Do chúng ta cần thể hiện xu hướng của số lượng ca bệnh nên chúng ta sẽ dùng biểu đồ đường (line chart) để thể hiện

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Như chúng ta thấy ở trên, tổng số ca active hiện tại trên thế giới có xu hướng tăng và không có dấu hiệu nào sẽ giảm, tuy nhiên, tổng số lượng ca mới mỗi ngày thì lại có thời điểm tăng giảm không rõ ràng.

Để tìm hiểu kỹ hơn, chúng ta sẽ xem xét riêng từng khu vực bao gồm EU,Asia,Africa,North America, South America,Australia/Oceania, ở đây do một số nước không cập nhật khu vực trên web, nên chúng ta sẽ để những nước đó ở một khu vực riêng là unknow, chúng ta tạm thời không xét đến các nước này do thiếu hụt thông tin về khu vực

A picture containing graphical user interface

Description automatically generated

Bảng số liệu (nghìn ca)

Table

Description automatically generated

Dựa theo biểu đồ và bảng kết quả trên thì số lượng ở các khu vực ít ca như south America và Australia đang có xu hướng giảm nhẹ, tuy nhiên các khu vực có nhiều ca như EU,NA lại có xu hướng tăng cực kỳ nhanh, khá bất ngờ là Asia lại giảm mạnh số ca active

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

* Chúng ta sẽ xem xét đến số ca mới mỗi ngày và các ca hồi phục mỗi ngày ở các khu vực

Như chúng ta thấy trên hình, số ca mới nhiễm mỗi ngày có thể lên tới khoảng 400 nghìn ca, trong khi đó, số ca hồi phục mỗi ngày tối đa chỉ khoảng 300 nghìn

* có thể nói covid 19 có tốc độ lây lan rất khủng khiếp, không lạ khi tổng số ca active luôn có xu hướng tăng chứ không hề có xu hướng giảm, dù cho số lượng ca mới mỗi ngày có thể tăng giảm nhưng luôn vượt trội hơn số ca hồi phục mỗi ngày

1. **So sánh số lượng người chết giữa các khu vực**

Vì đây là so sánh giữa các category nên chúng ta sẽ dùng biểu đồ cột, bar chart để dễ so sánh

Graphical user interface, background pattern

Description automatically generated with medium confidence

Tính tới thời điểm hiện tại số lượng người chết ở khu vực EU là cao nhất và khá cách biệt so với phần còn lại. Tuy nhiên số ca chết ở South America, North America và Asia là khá gần nhau

Chúng ta sẽ dùng biểu đồ tròn (pie chart) để xem xét tỉ lệ số ca chết/tổng số vì biểu đồ tròn thể hiện tỉ lệ

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Graphical user interface

Description automatically generated with low confidenceTổng kết

* Rõ ràng EU dù đang có số lượng ca chết do nhiễm bệnh cao nhất thế giới, thế nhưng tỉ lệ chết lại đứng thứ 5, chỉ cao hơn Australia và Asia
* Ngược lại với điều đó, Africa tuy có số lượng nhiễm bệnh thấp, nhưng tỉ lệ chết lại rất cao
* Liệu có thể kết luận rằng thể trạng người châu Phi rất khoẻ, do đó mà người miễn dịch thì sẽ không mắc bệnh, nhưng người mắc bệnh thì là do hệ miễn dịch không đủ khoẻ không?

South America có số lượng người chết cao thứ 4, chênh lệnh không đáng kể với NA và Asia, nhưng lại có tỉ lệ chết cao nhất trong số các khu vực

1. **Mối quan hệ nhân quả**

Chúng ta sẽ nhìn sơ qua về cột tests/1m pop, cột tot/1m pop, death/1m pop để kiểm tra xem, liệu có mối liên hệ nào giữ tỉ lệ test và tỉ lệ nhiễm bệnh, tỉ lệ chết do bệnh hay không, do số lượng nước khá nhiều(trên 200 nước) do đó chúng ta chỉ lấy những nước có số lượng người chết >10000 để xem xét

A picture containing text, measuring stick

Description automatically generated

Như chúng ta thấy trên hình, dường như nước nào có tỉ lệ tests/1M thì dù cho số ca nhiễm bệnh Total/1M cao thì số ca chết lại thấp, trong khi ngược lại, số lượt test/1M càng thấp thì tỉ lệ Deaths/1M lại càng cao dù cho số ca nhiễm bệnh thấp.