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

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÁO CÁO CUỐI KỲ**Môn học : Dữ liệu lớn

Học kỳ I (2022-2023)

**ĐỀ TÀI:**

**ỨNG DỤNG KỸ THUẬT HỌC MÁY XỬ LÝ BÀI TOÁN TRÊN TẬP DỮ LIỆU NHỮNG CUỘC KHỦNG BỐ Ở THỔ NHĨ KỲ**

Sinh viên thực hiện:

Trần Nhật Tân MSSV: 19522177

Bùi Ngọc Thành MSSV: 1952

Huỳnh Quốc Khánh MSSV: 1952

Lê Thế Tiệm MSSV: 1952

GVHD: ThS. Nguyễn Hồ Duy Tri

Lớp: IS405.N11.HTCL

**Thành phố Hồ Chí Minh, tháng 12 năm 2022**

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

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÁO CÁO CUỐI KỲ**Môn học : Dữ liệu lớn

Học kỳ I (2022-2023)

**ĐỀ TÀI:**

**ỨNG DỤNG KỸ THUẬT HỌC MÁY XỬ LÝ BÀI TOÁN TRÊN TẬP DỮ LIỆU NHỮNG CUỘC KHỦNG BỐ Ở THỔ NHĨ KỲ**

Sinh viên thực hiện:

Trần Nhật Tân MSSV: 19522177

Bùi Ngọc Thành MSSV: 1952

Huỳnh Quốc Khánh MSSV: 1952

Lê Thế Tiệm MSSV: 1952

GVHD: ThS. Nguyễn Hồ Duy Tri

Lớp: IS405.N11.HTCL

**Thành phố Hồ Chí Minh, tháng 12 năm 2022**

Mục lục

[**LỜI CẢM ƠN** 3](#_Toc120797475)

[**NHẬN XÉT CỦA GIẢNG VIÊN** 4](#_Toc120797476)

[**I. TỔNG QUAN** 5](#_Toc120797477)

[**1.Giới thiệu dataset** 5](#_Toc120797478)

[**2. Xác định bài toán** 5](#_Toc120797479)

[**3. Mô tả dữ liệu** 5](#_Toc120797480)

[**II. XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU** 7](#_Toc120797481)

[**1. Các thư viện được sử dụng** 7](#_Toc120797482)

[**2. Làm sạch dữ liệu** 8](#_Toc120797483)

[**3. Chuyển đổi dataframe thành đồ thị** 11](#_Toc120797484)

[**III. CÁC ĐỘ ĐO VÀ KẾT QUẢ** 16](#_Toc120797485)

[**1. Degree Centrality** 17](#_Toc120797486)

[**2. Betweenness Centrality** 18](#_Toc120797487)

[**3. Closeness Centrality** 19](#_Toc120797488)

[**4. Page Rank** 21](#_Toc120797489)

[**IV.THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG** 23](#_Toc120797490)

[**1. Thuật toán Girvan Newman** 23](#_Toc120797491)

[**2. Thuật toán Louvain** 28](#_Toc120797492)

# **LỜI CẢM ƠN**

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Với lòng biết ơn sâu sắc nhất, đầu tiên em xin gửi lời cảm ơn chân thành đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM và quý Thầy Cô khoa Hệ thống thông tin đã giúp cho nhóm có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này.

Đặc biệt nhóm chúng em xin gửi lời cảm ơn chân thành tới thầy Nguyễn Hồ Duy Tri – giảng viên lý thuyết và thực hành môn Dữ liệu lớn đã tận tình giúp đỡ, trực tiếp chỉ bảo, hướng dẫn em trong suốt quá trình làm đề tài. Nhờ đó, chúng em đã tiếp thu được nhiều kiến thức bổ ích trong việc vận dụng cũng như kỹ năng làm đề tài. Nếu không có những lời hướng dẫn, dạy bảo của thầy thì em nghĩ đề tài này của nhóm rất khó có thể hoàn thiện được. Một lần nữa, em xin chân thành cảm ơn thầy.

Cuối cùng, các thành viên trong nhóm đã làm việc hết công suất để hoàn thành tốt đề tài của mình. Xin chân thành cảm ơn!

# **NHẬN XÉT CỦA GIẢNG VIÊN**

# **I. TỔNG QUAN ĐỀ TÀI**

## **1.1 Lý do chọn đề tài**

Khủng bố là hoạt động phá hoại, đe dọa bằng lời nói, hình ảnh hoặc video giết người do cá nhân hoặc tổ chức thực hiện làm thiệt mạng người, đặc biệt là thường dân, hoặc gây tổn thất cho xã hội và cộng đồng để tác động vào tâm lý đối phương gây hoang mang khiếp sợ, nhằm mục đích chính trị hoặc tôn giáo. Một đặc tính thống nhất của khủng bố là việc sử dụng bừa bãi bạo lực đối với những người không có khả năng chống cự với mục đích là sự nổi tiếng cho một nhóm, một phong trào, một cá nhân hoặc gây áp lực lên đối thủ chính trị buộc họ phải chấp nhận một giải pháp chính trị có lợi cho mình. Các tổ chức khủng bố có thể khai thác nỗi sợ hãi của con người để hỗ trợ đạt được những mục tiêu này. Đối tượng bị khủng bố gây thiệt hại có thể là tính mạng, sức khỏe, danh dự nhân phẩm, tài sản (của cá nhân, tổ chức hay của nhà nước) hoặc sự vững mạnh của một chính quyền nhà nước.

Khủng bố là mối quan tâm lớn đối với các quốc gia trên toàn thế giới, vì nó thường báo hiệu mở đầu cho chiến tranh và dân thường bị mắc kẹt trong cuộc chiến. Chúng ta không thể thay đổi thực tế mà chúng ta không hiểu đầy đủ, vì vậy đây là một số dữ liệu liên quan về chủ đề này Nhận thấy tầm quan trọng ấy, nhóm tập trung nghiên cứu ứng dụng giải thuật K-Means để phân cụm những cuộc tấn công ở các thành phố trong quốc gia Thổ Nhĩ Kỳ từ đó đề ra những chiến lược phù hợp giúp chống lại những cuộc khủng bố, góp phần mang lại an toàn cho người dân và xã hội.

## **1.2. Mô tả dữ liệu**

**a. Nguồn dữ liệu**

Tên dữ liệu:

Nguồn cung cấp:

Link dữ liệu: [terrorism-in-turkey | Kaggle](https://www.kaggle.com/datasets/senakaradeniz/terrorisminturkey).

Mô tả nguồn dữ liệu:

**b. Các thuộc tính của tập dữ liệu**

**-** Mỗi dòng trong tập dữ liệu là một vụ khủng bố tại Thổ Nhĩ Kỳ.

- Dữ liệu gồm 547 dòng và 14 thuộc tính.

|  |  |  |
| --- | --- | --- |
| Tên cột | Kiểu dữ liệu | Ý nghĩa |
| Id | Int | Id của dữ liệu |
| Year | Int | Năm xảy ra vụ khủng bố |
| Country | String | Quốc gia (Thổ Nhĩ Kỳ) |
| Region | String | Vùng của quốc gia (Trung Đông) |
| city | String | Thành phố xảy ra vụ khủng bố |
| AttackType | String | Kiểu tấn công |
| Killed | Int | Số người tử vong |
| Wounded | Int | Số người bị thương |
| Target | String | Mục tiêu của cuộc tấn công khủng bố |
| Summary | String | Diễn biến chi tiết của cuộc khủng bố |
| Group | String | Tổ chức khủng bố |
| Target\_type | String | Loại mục tiêu của cuộc tấn công khủng bố |
| Weapon\_type | String | Loại vũ khí dùng để khủng bố |
| Motive | String | Động cơ khủng bố |

- Bảng dữ liệu:

Graphical user interface, table

Description automatically generated

1.3 Mô tả bài toán

1.4 Kỹ thuật tiền xử lý dữ liệu được lựa chọn

## **1.5 Thuật toán khai thác dữ liệu được lựa chọn**

1.6

1.7

1.8

**II. TIỀN XỬ LÝ DỮ LIỆU**

## **2.1. Các thư viện được sử dụng**

Text

Description automatically generated

## **2.2 Đọc dữ liệu đầu vào**