**VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**INFORMATION SYSTEM FACULTY**

A picture containing logo, graphics, clipart, symbol

Description automatically generated

**MÔN HỌC: MẠNG XÃ HỘI**

**Đề tài: Phân tích dữ liệu thể thao Olympic and Medals**

**GVHD** : Nguyễn Thị Kim Phụng

**Lớp**: IS353.N22.HTCL

**Sinh viên thực hiện:**

Nguyễn Thanh Sơn – 20521847

**HO CHI MINH CITY,** **1 JUNE 2023**

Mục lục

[I. GIỚI THIỆU 3](#_Toc138094066)

[II. DỮ LIỆU 4](#_Toc138094067)

[1. Giới thiệu nguồn dữ liệu 4](#_Toc138094068)

[2. Xử lý và phân tích dữ liệu 4](#_Toc138094069)

[2.1. Làm sạch dữ liệu 4](#_Toc138094070)

[2.2. Chuyển đổi dataframe thành đồ thị 5](#_Toc138094071)

[III. THUỘC TOÁN PHÁT HIỆN CỘNG ĐỒNG 10](#_Toc138094072)

[1. Thực thi thuật toán Louvain bằng Gephi 10](#_Toc138094073)

[2. Thực thi bằng Python 10](#_Toc138094074)

[3. Một số thuật toán khác 14](#_Toc138094075)

[IV. XẾP HẠNG 22](#_Toc138094076)

[1. Thuật toán Pagerank (Thực thi bằng Gephi) 22](#_Toc138094077)

[2. Thuật toán Closeness\_centrality (Thực thi bằng Gephi) 23](#_Toc138094078)

[3. Thuật toán betweenness\_centrality (Thực thi bằng Gephi) 24](#_Toc138094079)

[V. TÀI LIỆU THAM KHẢO 26](#_Toc138094080)

# GIỚI THIỆU

Hiện nay, đại hội Olympic trở thành một sự kiện được ưa thích và quan tâm rộng rãi trên toàn cầu. Việc giao lưu thể thao không chỉ mang lại lợi ích về sức khỏe cho các vận động viên mà còn có tầm quan trọng về việc thắt chặt mối quan hệ giữa các quốc gia, khẳng định vị thế tinh thần của một quốc gia và tạo dựng danh tiếng cho các liên đoàn và đội tuyển thể thao. Việc phân tích và đánh giá các bên tham gia là một yếu tố thiết yếu để hiểu rõ hơn về sự kiện này. Trong môn học Mạng xã hội, chúng ta có thể sử dụng các phương pháp đo lường và khai thác dữ liệu từ mạng xã hội để phân tích các thông tin có trong đó và tạo ra những cụm thông tin liên quan.

# DỮ LIỆU

## Giới thiệu nguồn dữ liệu

* Dataset về Olympic Sports and Medals, 1896-2014 được lấy từ Kaggle.com.
* Dataset gồm 9 cột và 31165 dòng
* Link dataset: [Olympic Sports and Medals, 1896-2014 | Kaggle](https://www.kaggle.com/datasets/the-guardian/olympic-games)

|  |  |  |
| --- | --- | --- |
| STT | THUỘC TÍNH | Ý NGHĨA |
| 1 | Year | Đại diện cho năm diễn ra kỳ Olympic |
| 2 | City | Thành phố nơi tổ chức kỳ Olympic |
| 3 | Sport | Môn thể thao tham gia trong kỳ Olympic |
| 4 | Discipline | Cụ thể hơn về loại hình thể thao trong mỗi môn |
| 5 | Athlete | Tên của vận động viên tham gia Olympic |
| 6 | Country | Quốc tịch của vận động viên |
| 7 | Gender | Giới tính của vận động viên |
| 8 | Event | Sự kiện thi đấu trong mỗi môn thể thao |
| 9 | Medal | Loại huy chương mà vận động viên giành được. |

## Xử lý và phân tích dữ liệu

### 2.1. Làm sạch dữ liệu

Bộ dữ liệu nhìn chung có vài thuộc tính bị thiếu dữ liệu,cần loại bỏ khỏi

dataset. Dữ liệu sẽ được đọc từ file csv đưa vào dataframe. Và vì dataset khá lớn nên em chỉ lấy dữ liệu của năm 2012.

A screenshot of a computer

Description automatically generated

### 2.2. Chuyển đổi dataframe thành đồ thị

* Đồ thị 2 phía:
* Node: quốc tịch của vận động viên (Country) và loại hình thể thao (Discipline)
* Edge: mối quan hệ giữa quốc tịch vận động viên và loại hình thể thao.

A picture containing text, font, screenshot

Description automatically generated

Có thể thấy số quốc tịch là 85, số loại hình thể thao là 39 và số cạnh là 1945.

* Code đồ thị 2 phía:

A close-up of a computer code

Description automatically generated with low confidenceA picture containing sketch, drawing, line, child art

Description automatically generated

* Nhìn vào đồ thị ta có thể thấy mỗi một quốc tịch đều tham dự nhiều môn thể thao và mỗi một môn thể thao đều được nhiều quốc tịch thi đấu => đồ thị vô hướng.
* Code đồ thị 1 phía:

A picture containing text, screenshot, font

Description automatically generatedA picture containing symmetry

Description automatically generated

* Xuất dữ liệu đồ thị 1 phía ra file csv để thực hiện trên Gephi:

A picture containing text, font, screenshot, white

Description automatically generated

* Đồ thị 1 phía trên Gephi:

A picture containing sketch, drawing, art

Description automatically generated

# THUỘC TOÁN PHÁT HIỆN CỘNG ĐỒNG

## Thực thi thuật toán Louvain bằng Gephi

A picture containing colorfulness, drawing

Description automatically generated

* Khi nhìn vào đồ thị ta có thể thấy thuật toán cho kết quả 3 cụm.

## Thực thi bằng Python

A computer code on a white background

Description automatically generated with low confidence

A picture containing art

Description automatically generated with medium confidenceA picture containing text, screenshot, font

Description automatically generated

* Số cụm là 3:
  + Cụm 0: 'CHN','MAS','RUS','CAN','HUN','JPN','LTU','KOR','BRA','UKR', 'CUB','KAZ','COL','IND','AZE','THA','MGL','TJK','UZB','BUL', 'VEN','EGY','ROU','PRK','GEO','SGP','TPE','INA','CMR','ARM'

\*Ý nghĩa: Với mỗi quốc tịch trong cụm đều có vận động viên tham gia

A picture containing text, screenshot

Description automatically generated

A picture containing text, screenshot, font, plot

Description automatically generated

* + Cụm 1:

'TUN','RSA','ETH','KEN','JAM','ALG','MAR','TUR','BRN','GUA', 'GRN','DOM','TTO','PUR','BAH','IRL','BOT','IRI','EST','SLO', 'POL','QAT','FIN','UGA'

\*Ý nghĩa: Với mỗi quốc tịch trong cụm đều có vận động viên tham gia

A picture containing text, screenshot, rectangle

Description automatically generated

A picture containing text, screenshot, font, plot

Description automatically generated

* + Cụm 2:

'USA','GBR','AUS','MEX','GER','ITA','NED','BLR','ESP','FRA', 'CRO','SRB','CZE','NZL','DEN','SVK','NOR','POR','LAT','HKG', 'SUI','SWE','KSA','MNE','ARG','BEL','GRE','CYP','KUW','GAB', 'AFG'

\*Ý nghĩa: Với mỗi quốc tịch trong cụm đều có vận động viên tham gia A screenshot of a computer

Description automatically generated with low confidence

A picture containing text, screenshot, plot, font

Description automatically generated

## Một số thuật toán khác

1. Girvan Newman

A screenshot of a computer program

Description automatically generated with medium confidence

A picture containing art

Description automatically generated with low confidence

* Số cụm là 2:

Cụm 0:

'POL', 'USA', 'IRL', 'AZE', 'GBR', 'KOR', 'MGL', 'BRA', 'CAN', 'JPN', 'MNE', 'GER', 'RUS', 'NOR', 'COL', 'BOT', 'UGA', 'LAT', 'SVK', 'CHN', 'BUL', 'TTO', 'DEN', 'BAH', 'KUW', 'SWE', 'MAS', 'CMR', 'IRI', 'TUR', 'CZE', 'BLR', 'EGY', 'IND', 'QAT', 'TUN', 'HKG', 'ARG', 'HUN', 'GEO', 'GUA', 'CYP', 'KAZ', 'INA', 'ALG', 'MEX', 'SUI', 'THA', 'FIN', 'ETH', 'SLO', 'JAM', 'BEL', 'LTU', 'UKR', 'RSA', 'BRN', 'TPE', 'PUR', 'ESP', 'AUS', 'EST', 'DOM', 'GRN', 'GRE', 'CRO', 'ARM', 'PRK', 'VEN', 'FRA', 'ROU', 'ITA', 'MAR', 'NZL', 'KEN', 'SRB', 'SGP', 'NED', 'GAB', 'AFG', 'TJK', 'CUB', 'UZB', 'POR'

\*Ý nghĩa: Với mỗi quốc tịch trong cụm đều có vận động viên tham gia

A screenshot of a table

Description automatically generated with low confidence A picture containing text, screenshot, plot, line

Description automatically generated

* + Cụm 1: 'KSA'

\*Ý nghĩa: Với mỗi quốc tịch trong cụm đều có vận động viên tham gia

A screenshot of a computer

Description automatically generated with medium confidence A picture containing text, screenshot, number, line

Description automatically generated

1. K means

A computer code on a white background

Description automatically generated with low confidence

A picture containing text, line, diagram, plot

Description automatically generated A picture containing text, font, screenshot

Description automatically generated A yellow and purple dots

Description automatically generated with low confidence

A computer code on a white background

Description automatically generated with low confidence

* Khi nhìn vào biểu đồ ta thấy số cụm là 2 và khi in ra ta được

Cụm 0: USA, MAS, RUS, MEX, TUN, RSA, LTU, NED, SRB, MAR, TUR, TTO, PUR, UKR, SLO, POL, QAT, UGA, NZL, THA, MGL, TJK, UZB, SVK, NOR, POR, LAT, SUI, SWE, KSA, VEN, ROU, MNE, PRK, KUW, SGP, TPE

\*Ý nghĩa: có thể thấy các quốc tịch AUS, CHN, ESP, FRA, GBR, GER, ITA, JPN tham gia rất đông và có dự phân chia rõ rệt giữa các quốc tịch tham gia nhiều và tham gia ít.

* + Cụm 1: CHN, GBR, AUS, CAN, GER, HUN, ITA, JPN, BLR, ESP, FRA, KOR, BRA, CRO, ETH, KEN, JAM, ALG, BRN, GUA, GRN, DOM, CZE, BAH, IRL, BOT, CUB, IRI, EST, FIN, KAZ, COL, DEN, IND, AZE, BUL, HKG, EGY, ARG, BEL, GEO, GRE, CYP, GAB, AFG, INA, CMR, ARM

\*Ý nghĩa: ta có thể thấy rõ dù cụm 1 mật độ tham gia của các quốc tịch đều hơn so với cụm 0 và không có quốc tịch nào có chỉ số cao hơn 100.

# XẾP HẠNG

## Thuật toán Pagerank (Thực thi bằng Gephi)

Ý nghĩa: quốc tịch CHN có hệ số Pagerank cao nhất hay nói cách khác là quốc tịch tham gia rất nhiều

A screenshot of a computer

Description automatically generated with medium confidence

Thực thi bằng Python:

A screenshot of a computer

Description automatically generated with medium confidence

## Thuật toán Closeness\_centrality (Thực thi bằng Gephi)

Ý nghĩa: quốc tịch USA có độ đo lớn nhất tương ứng có nhiều vận động viên tham dự nhất

A picture containing text, screenshot, number

Description automatically generated

Thực thi bằng Python:

A screenshot of a computer

Description automatically generated

## Thuật toán betweenness\_centrality (Thực thi bằng Gephi)

A picture containing text, screenshot, number

Description automatically generated

Thực thi bằng Python:

A screenshot of a computer

Description automatically generated with medium confidence

# TÀI LIỆU THAM KHẢO

[1] “Olympic Sports and Medals, 1896-2014” [Olympic Sports and Medals, 1896-2014 | Kaggle](https://www.kaggle.com/datasets/the-guardian/olympic-games) (accessed May 31, 2023).

[2] N. T. Kim Phụng, “Slide bài giảng và hướng dẫn thực hành do giảng viên cung cấp” ThS. Nguyễn Thị Kim Phụng