

**ĐẠI HỌC QUỐC GIA TP.HCM**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

□□&□□

****

ĐỒ ÁN MÔN HỌC

**KHO DỮ LIỆU VÀ OLAB**

Đề tài:

**XÂY DỰNG KHO DỮ LIỆU VỀ KẾT QUẢ HUY CHƯƠNG**

**TRONG OLYMPIC TỪ 1896 - 2016**

*Lớp****:***

*GVHD:* **Cô Nguyễn Thị Kim Phụng**

*Sinh viên****:***  **Thân Trung Hiếu - 19521513**

TP.HCM, Tháng 11 - 2022

**NHẬN XÉT GVHD**

**MỤC LỤC**

Mục Lục

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 4](#_Toc120628674)

[1.1 Lý do chọn đề tài 4](#_Toc120628675)

[1.2 Nội dung đề tài 4](#_Toc120628676)

[1.3 Xác định bài toán 4](#_Toc120628677)

[1.4 Mô tả dữ liệu gốc 5](#_Toc120628678)

[CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU VÀ TRỰC QUAN HÓA DỮ LIỆU LÊN ĐỒ THỊ 7](#_Toc120628679)

[2.1 Tiền xử lý dữ liệu 7](#_Toc120628681)

[2.2 Tạo mạng liên kết 9](#_Toc120628682)

[2.3 Trực quan hóa mạng liên kết bằng code python 9](#_Toc120628683)

[2.4 Trực quan hóa đồ thị 1 phía trên gephi 13](#_Toc120628684)

[CHƯƠNG 3: PHÂN TÍCH VÀ TRỰC QUAN HÓA BẰNG CÁC THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG 15](#_Toc120628685)

[3.1 Thuật toán phát hiện cộng đồng Louvain 15](#_Toc120628686)

[3.2 Thuật toán Girvan Newman 19](#_Toc120628687)

[3.2 Thuật toán K-means 22](#_Toc120628688)

[CHƯƠNG 4: PHÂN TÍCH VÀ TRỰC QUAN HÓA BẰNG CÁC ĐỘ ĐO MẠNG XÃ HỘI 28](#_Toc120628689)

[4.1 Độ đo Closeness centrality 28](#_Toc120628690)

[4.2 Độ đo Betweeness centrality 30](#_Toc120628691)

[4.2.1 Độ đo Node Betweeness 30](#_Toc120628696)

[4.2.2 Độ đo Edge Betweeness 32](#_Toc120628697)

[4.3 Thuật toán PageRank 33](#_Toc120628698)

[4.4 Eigenvector và Eigenvalue 35](#_Toc120628699)

[4.4.1 Eigenvector centrality 35](#_Toc120628700)

[4.4.2 Eigenvalue 37](#_Toc120628701)

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## Lý do chọn đề tài

Thế vận hội Olympics có một lịch sử phong phú với hơn 120 năm, gắn liền với sự phát triển của lịch sử các nước nói riêng và thế giới nói chung. Trong nhiều năm qua, Thế vận hội Olympics đã được tổ chức trên toàn thế giới với sự tham gia của các các vận động viên đến từ khắp nơi thi đấu cạnh tranh để thành công trong các môn thể thao khác nhau. Để hiểu rõ và trả lời được các câu hỏi như thế mạnh của từng quốc gia ở các bộ môn thể thao trong các kì Olympics, Quốc gia nào dẫn đầu về vận động viên nam hoặc nữ trong môn thể thao nào hay số lượng huy chương của từng quốc gia… Nhóm đã quyết định chọn đề tài phân tích dữ liệu về các vận động viên và kết quả huy chương trong các kì Olympics từ năm 1896 đến năm 2016.

## Nội dung đề tài

* Tên dataset: 120 years of Olympic history: athletes and results
* Link dataset: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/metadata
* Mô tả: Đây là tập dữ liệu bao gồm tất cả các Thế vận hội từ Athens 1896 đến Rio 2016
* Dữ liệu có 271.117 dòng dữ liệu và 15 cột thuộc tính.
* Dữ liệu đã qua xử lý

## Xác định bài toán

Input: Tập dữ liệu ban đầu trên nguồn dữ liệu Kaggle đã qua tiền xử lý dữ liệu

Ouput: Đưa ra độ đo, đưa ra cộng đồng phục vụ cho việc phân tích mạng xã hội “120 years of Olympic history: athletes and results.”

## Mô tả dữ liệu gốc

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *STT* | *Thuộc tính* | *Kiểu dữ liệu* | *Mô tả* | *Các giá trị* |
| *1* | ID | Int | Mã định danh duy nhất của vận động viên | ID |
| *2* | Name | Varchar | Tên vận động viên | Name |
| *3* | Sex | Varchar | Giới tính của vận động viên Phân biệt:  1. M: Nam giới  2. F: Nữ giới | Sex |
| *4* | Age | Int | Tuổi của vận động viên (tính theo năm) | Age |
| *5* | Height | Int | Chiều cao (tính bằng cm) | Height |
| *6* | Weight | Int | Cân năng (tính bằng kg) | Weight |
| *7* | Team | Varchar | Tên đội mà vận động viên đã thi đấu | Team |
| *8* | NOC | Varchar | Mã gồm 3 chữ cái của Quốc gia tham dự Olympics | NOC |
| *9* | Games | Varchar | Năm và mùa của Thế vận hội mà vận động viên thi đấu  (Định dạng: YYYY Season) | Games |
| *10* | Year | Int | Năm Thế vận hội mà vận động viên tham gia | Year |
| *11* | Season | Varchar | Mùa của Thế vận hội mà vận động viên đã thi đấu | Season |
| *12* | City | Varchar | Thành phố đăng cai Thế vận hội mà vận động viên đã thi đấu | City |
| *13* | Sport | Varchar | Môn thể thao mà vận động viên đã thi đấu | Sport |
| *14* | Event | Varchar | Sự kiện mà vận động viên thi đấu | Event |
| *15* | Medal | Varchar | Huy chương do vận động viên đạt được; phân loại:   * Vàng * Bạc * Đồng   NA: nếu không giành được huy chương. | Medal |
| *16* | Region | Varchar | Tên quốc gia thi đấu | Region |

# CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU VÀ TRỰC QUAN HÓA DỮ LIỆU LÊN ĐỒ THỊ



## 2.1 Tiền xử lý dữ liệu

Thêm thư viện cần thiết

Text

Description automatically generated

Dữ liệu sẽ được đọc từ file csv đưa vào dataframe và lấy 2 cột Year và Sport. Graphical user interface

Description automatically generated

Dữ liệu ban đầu gồm 2711116 dòng.

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

Table

Description automatically generated

Kết quả nhận được là một bộ dữ liệu 906 dòng và 2 cột

## 2.2 Tạo mạng liên kết

* *Đồ thị 2 phía*

*Node: là số môn thể thao được tổ chức(Sport) và năm tổ chức sự kiện (Year),*

*Edge: mối quan hệ giữa việc năm tổ chức sự kiện và những môn thể thao được tổ chức trong năm đó*

Đưa dữ liệu từ dataframe vào đồ thị vô hướng:

Text

Description automatically generated

Có thể thấy có 35 năm tổ chức sự kiện

Số môn thể thao là 66

Số cạnh là 906

## 2.3 Trực quan hóa mạng liên kết bằng code python

* ***Đồ thị 2 phía***
* Tạo đồ thị 2 phía

Text

Description automatically generated with medium confidence

Diagram

Description automatically generated with medium confidence

* ***Nhận xét:*** Nhìn vào đồ thị có thể thầy một năm có thể tổ chức được nhiều môn thể thao và một môn thể thao có thể được tổ chức ở nhiều năm
* ***Trực quan hóa đồ thị 1 phía trên python***
* *- Đồ thị 1 phía theo môn thể thao (Sport)*
* *Node: là môn thể thao được tổ chức*
* *Edge: Hai môn thể thao được tổ chức cùng một năm sẽ được nối với nhau để tạo thành cạnh, ý nghĩa nói lên là khả năng xuất hiện các môn thể thao trong cùng năm*
* *Weight : Trọng số là số vị trí trùng nhau mà 2 môn xảy ra trùng năm*

Tạo đồ thị 1 phía:

* Text

  Description automatically generated

*A picture containing diagram

Description automatically generated*

## 2.4 Trực quan hóa đồ thị 1 phía trên gephi

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

Text

Description automatically generated

* Đồ thị 1 phía trên Gephi

Chart, radar chart

Description automatically generated

# CHƯƠNG 3: PHÂN TÍCH VÀ TRỰC QUAN HÓA BẰNG CÁC THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG

## 3.1 Thuật toán phát hiện cộng đồng Louvain

* ***Thực thi thuật toán Louvain bằng Gephi***

Chart, radar chart

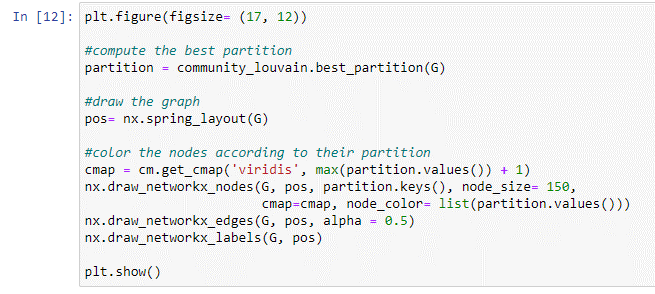
Description automatically generated

* ***Nhận xét:*** Đồ thị được phân thành 3 cụm.

A picture containing rectangle

Description automatically generated

* ***Thực thi thuật toán Louvain bằng python***



Đồ thị phân cụm sử dụng thuật toán Louvain:

Diagram

Description automatically generated

Danh sách các cụm



* ***Tổng số cụm:*** 3
* ***Thành phần node của từng cụm***
* ***Cụm thứ 0***: *Speed Skating, Cross Country Skiing, Ice Hockey, Biathlon, Alpine Skiing, Luge, Bobsleigh, Figure Skating, Nordic Combined, Freestyle Skiing, Ski Jumping, Curling, Snowboarding, Short Track Speed Skating, Skeleton*

***Ý nghĩa:*** *Trong cụm này, những môn thể thao được tổ chức nhiều nhất là :*

***Pivot Cụm thứ 0:***

* ***Cụm thứ 1:*** *Football, Tug-Of-War, Athletics, Swimming, Sailing, Gymnastics, Art Competitions, Weightlifting, Wrestling, Water Polo, Hockey, Rowing, Fencing, Equestrianism, Shooting, Boxing, Cycling, Diving, Tennis, Modern Pentathlon, Golf, Rugby, Lacrosse, Polo, Cricket, Racquets, Motorboating, Military Ski Patrol, Croquet, Jeu De Paume, Roque, Alpinism, Basque Pelota, Aeronautics*

***Ý nghĩa:*** *Trong cụm này, những môn thể thao được tổ chức nhiều nhất là :*

***Pivot Cụm thứ 1:***

* ***Cụm thứ 2:*** *Basketball, Judo, Badminton, Handball, Taekwondo, Canoeing, Softball, Archery, Volleyball, Synchronized Swimming, Table Tennis, Baseball, Rhythmic Gymnastics, Rugby Sevens, Trampolining, Beach Volleyball, Triathlon*

***Ý nghĩa:*** *Trong cụm này, những môn thể thao được tổ chức nhiều nhất là :*

***Pivot Cụm thứ 2:***

## 3.2 Thuật toán Girvan Newman

Thực thi Girvan Newman bằng code Python

Graphical user interface, text, application

Description automatically generated

Đồ thị Girvan newman

Diagram

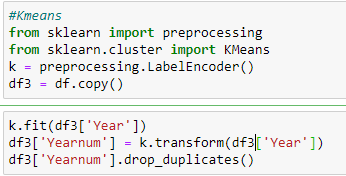
Description automatically generated

Từ đồ thị, ta thấy có node có màu sắc khác với những node còn lại

## 3.2 Thuật toán K-means

Code python

Thêm thư viện và chuyển dữ liệu qua dạng số:



Text

Description automatically generated

Sử dụng phương pháp Elbow để chọn số cụm

A picture containing graphical user interface

Description automatically generated

Từ đồ thị Elbow, ta thấy đoạn dữ liệu thay đổi khi K bắt đầu bằng 2, để chuẩn xác, ta lấy k = 3 trở đi.

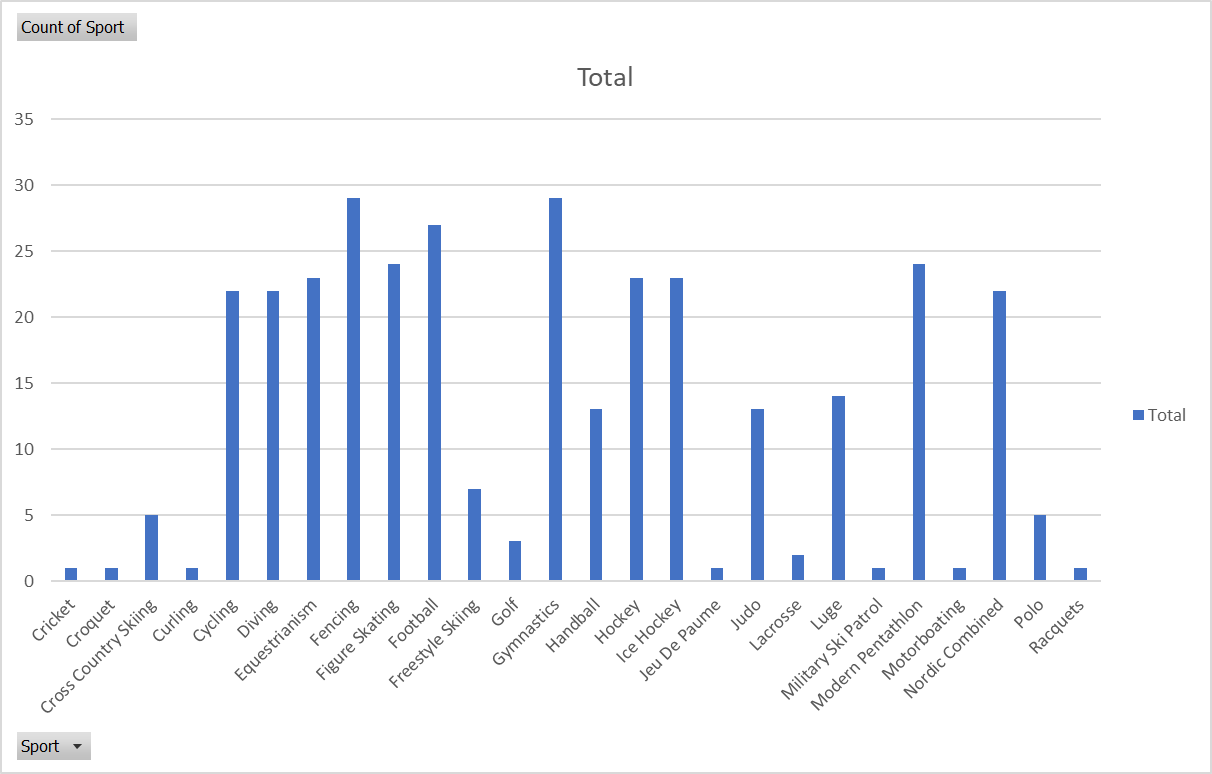
Phân cụm theo k =3:

Graphical user interface

Description automatically generated with medium confidence

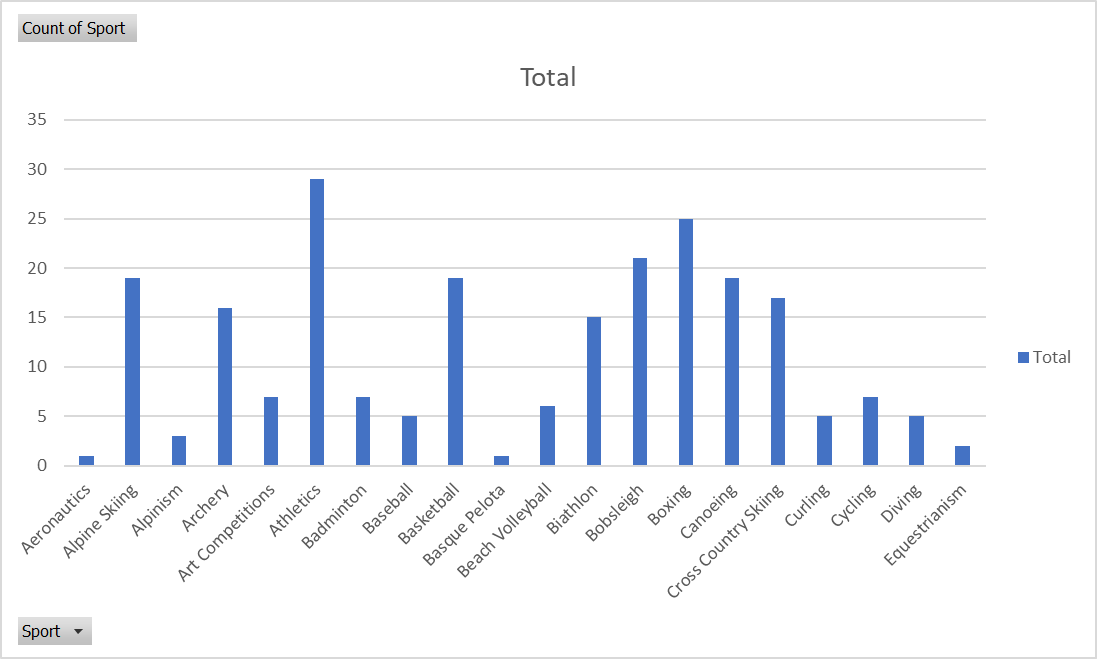
Từ đây ta được 3 cụm:

Cụm 1: *Judo, Football, Ice Hockey, Gymnastics, Handball, Luge, Hockey, Fencing, Equestrianism, Cycling, Modern Pentathlon, Diving, Figure Skating, Golf, Nordic Combined, Freestyle Skiing, Cross Country Skiing, Curling, Lacrosse, Polo, Cricket, Racquets, Motorboating, Military Ski Patrol, Croquet, Jeu De Paume*



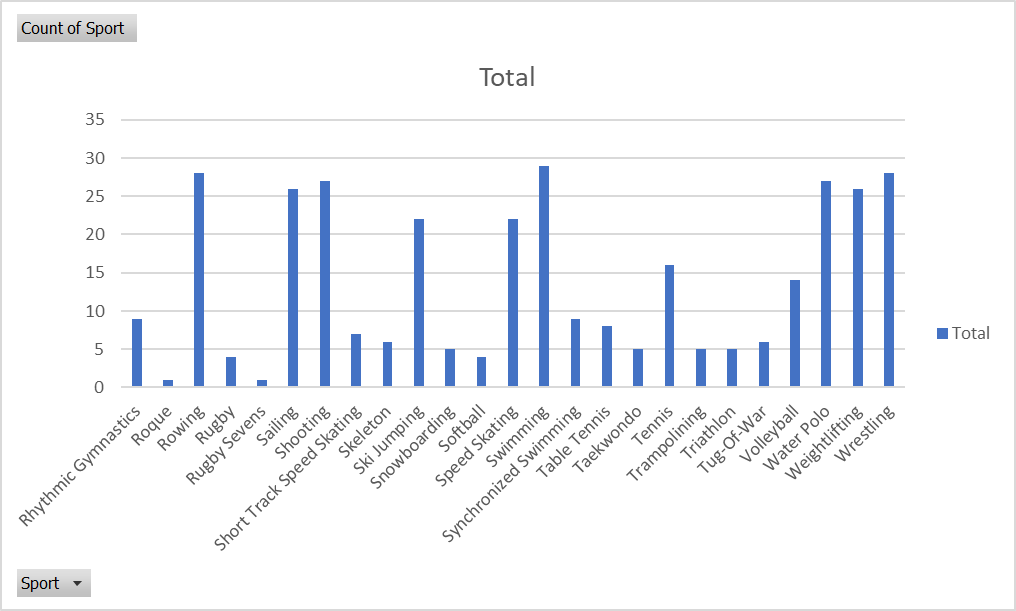
Ý nghĩa: Trong cụm này, những môn thể thao được tổ chức nhiều nhất là : *Fencing, Gymnastics và Football*

Cụm 2: *Basketball, Cross Country Skiing, Athletics, Badminton, Biathlon, Art Competitions, Alpine Skiing, Bobsleigh, Boxing, Diving, Canoeing, Cycling, Equestrianism, Archery, Baseball, Beach Volleyball, Curling, Alpinism, Basque Pelota, Aeronautics*



Ý nghĩa: Trong cụm này, những môn thể thao được tổ chức nhiều nhất là : *Athletics và Boxing*

Cụm 3: *Tug-Of-War, Speed Skating, Swimming, Sailing, Weightlifting, Wrestling, Water Polo, Rowing, Shooting, Taekwondo, Tennis, Softball, Volleyball, Synchronized Swimming, Table Tennis, Rhythmic Gymnastics, Rugby Sevens, Trampolining, Triathlon, Ski Jumping, Snowboarding, Rugby, Short Track Speed Skating, Skeleton, Roque*

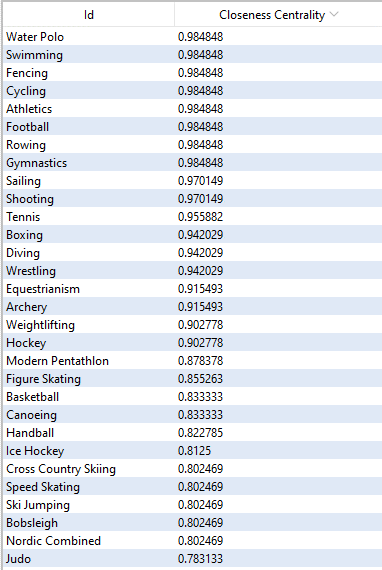


Ý nghĩa: Trong cụm này, những môn thể thao được tổ chức nhiều nhất là : *Swimming, Wrestling, Rowing, Water Polo, Shooting, Sailing và Weightlifting*

# CHƯƠNG 4: PHÂN TÍCH VÀ TRỰC QUAN HÓA BẰNG CÁC ĐỘ ĐO MẠNG XÃ HỘI

## 4.1 Độ đo Closeness centrality

* ***Thực thi bằng gephi***



* ***Thực thi bằng code python***

Text

Description automatically generated

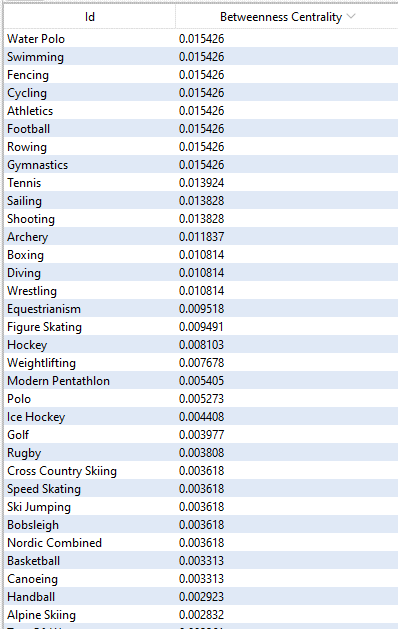
* ***Nhận xét:*** Độ đo closeness centrality càng cao thì chứng tỏ môn thể thao đó được tổ chức cùng năm với nhiều môn thể thao khác. Trong đồ thị, đường đi càng ngắn. Được tổ chức nhiều năm nhất

## 4.2 Độ đo Betweeness centrality



### Độ đo Node Betweeness

* ***Thực thi bằng gephi***



* ***Thực thi bằng python***

Text

Description automatically generated

***Nhận xét:*** Độ đo betweenness centrality càng cao thì chứng tỏ những môn thể thao này được tổ chức nhiều nhất

Đóng vai trò trung gian cho các mối quan hệ giữa các node trong mạng lưới.

### Độ đo Edge Betweeness

* ***Thực thi bằng python***

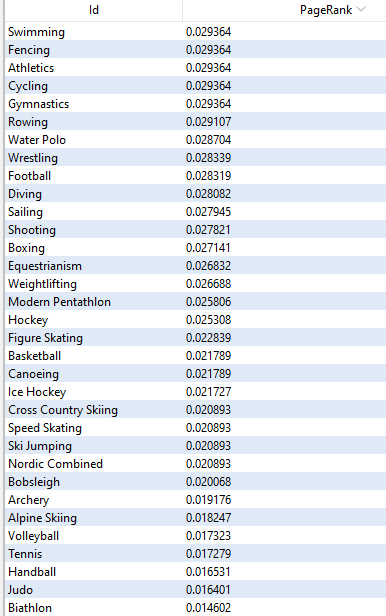
Text

Description automatically generated

* ***Nhận xét:***

## 4.3 Thuật toán PageRank

* ***Thực thi bằng gephi***



* ***Thực thi bằng python***

Text

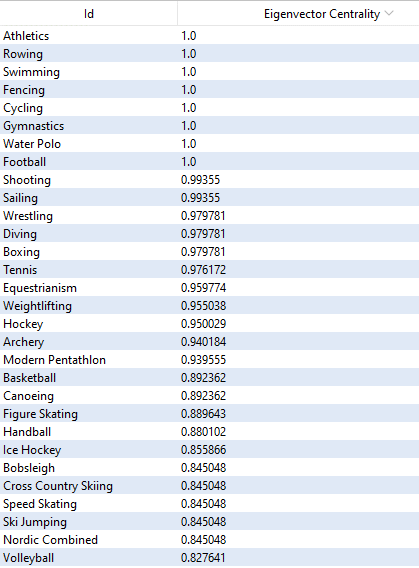
Description automatically generated

***Nhận xét:*** Độ đo pagerank càng cao thì chứng tỏ những môn thể thao này tổ chức vào nhiều năm nhất và có mức độ ảnh hưởng cao nhất.

## 4.4 Eigenvector và Eigenvalue

### 4.4.1 Eigenvector centrality

* ***Thực thi bằng gephi***



* ***Thực thi bằng python***

Text

Description automatically generated

### 4.4.2 Eigenvalue

* ***Thực thi bằng code python***

Table

Description automatically generated