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

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ**

**MÔN MẠNG XÃ HỘI**

**Giảng viên hướng dẫn:**  Ths Nguyễn Thị Kim Phụng

**Sinh viên thực hiện :**  Lê Tiến Vinh - 19522521

TP.HCM Tháng 12 Năm 2022

MỤC LỤC

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 3](#_Toc123331452)

[1.1 Giới thiệu. 3](#_Toc123331453)

[1.2 Tập dữ liệu. 3](#_Toc123331454)

[1.2.1 Tập dữ liệu: 3](#_Toc123331455)

[1.2.2 Giải thích ý nghĩa về tập dữ liệu: 3](#_Toc123331456)

[CHƯƠNG 2: XỬ LÝ VÀ PHÂN THÍCH DỮ LIỆU. 5](#_Toc123331457)

[2.1 Làm sạch dữ liệu. 5](#_Toc123331458)

[2.2 Chuyển đổi dữ liệu trên thành đồ thị. 6](#_Toc123331459)

[CHƯƠNG 3: CÁC THUẬT TOÁN XẾP HẠNG. 10](#_Toc123331460)

[3.1 Closeness Centrality 10](#_Toc123331461)

[3.1.1 Sử dụng Python. 10](#_Toc123331462)

[3.1.2 Sử dụng Gephi. 12](#_Toc123331463)

[3.2: Page Rank. 13](#_Toc123331464)

[3.2.1 Sử dụng Python 14](#_Toc123331465)

[3.2.2 Sử dụng Gephi. 15](#_Toc123331466)

[3.3 Betweenness Centrality. 16](#_Toc123331467)

[3.3.1 Sử dụng Python. 16](#_Toc123331468)

[3.3.2 Sử dụng Gephi. 18](#_Toc123331469)

[CHƯƠNG 4: CÁC THUẬT TOÁN GOM CỤM. 20](#_Toc123331470)

[4.1 Thuật toán Louvain. 20](#_Toc123331471)

[4.1.1 Sử dụng Python. 20](#_Toc123331472)

[4.1.2. Visualize với Pivot Table. 20](#_Toc123331473)

[CHƯƠNG 5: BÀI TẬP TÍNH TAY CÁC ĐỘ ĐO, GOM CỤM TRÊN 10 NODE NGẪY NHIÊN 22](#_Toc123331474)

[CHƯƠNG 6: KẾT LUẬN. 25](#_Toc123331475)

# **CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI**

* 1. Giới thiệu.

Hiện nay ngành giải trí cụ thể là ngành video game trên thế giới có độ phổ biến rất cao, vì thế đã có rất nhiều nhà phát hành game và cũng nhiều nền tảng game ra đời. Bởi vì càng nhiều nhà phát hành và nền tảng game ra đời như vậy nên sự canh tranh trong ngành phát triển video game giải trí cũng sẽ tăng cao.

Thông qua môn mạng xã hội em muốn phân tích về vấn đề cạnh tranh của các nhà phát hành hiện tại trên những nền tảng hiện tại, mục đích là có thể giúp hạn chế đi sự trùng lặp khi một nhà phát hành game mới có ý định phát triển game trên một nền tảng nào đó đã có quá nhiều sự cạnh tranh dẫn đến khó phát triển, hoặc là gây nhàm chán với người sử dụng.

Input: Tập dữ liệu về thông tin của những Video Game.

Output: Đưa ra những thông tin về số lượng nhà phát hành cùng phát hành game trên những nền tảng hay là nhà phát hành nào phát hành game trên nhiều nền tảng nhất.

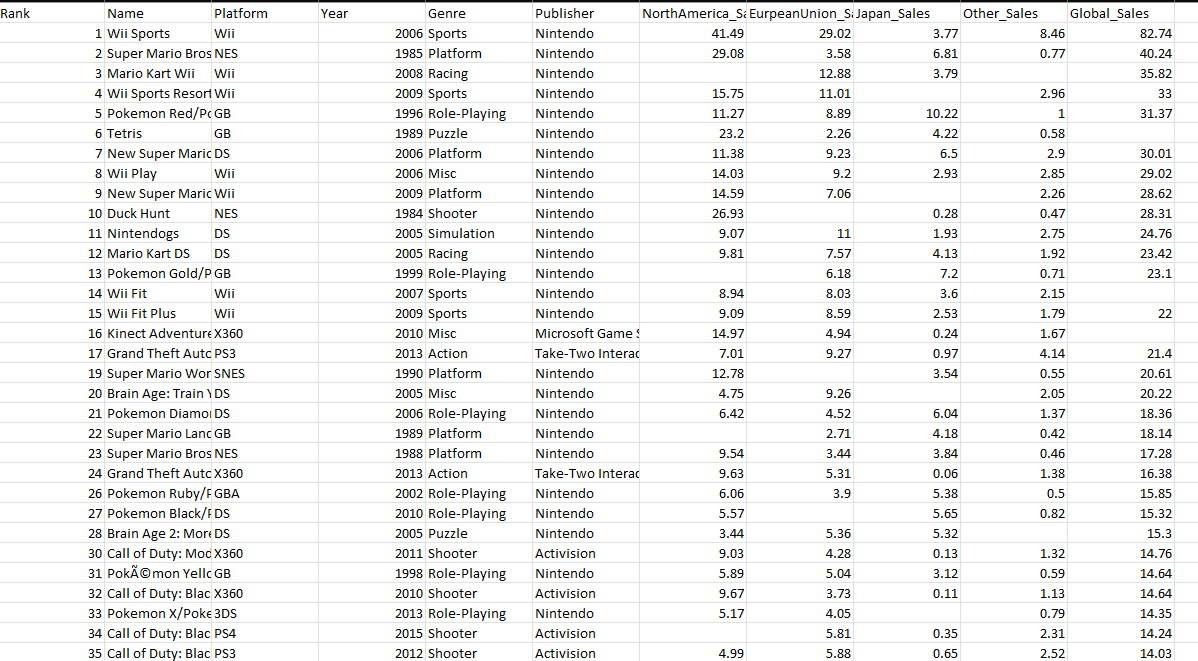
* 1. Tập dữ liệu.

1.2.1 Tập dữ liệu:

Thông tin về video game được lấy từ link sau

[https://www.kaggle.com/datasets/atharvaingle/video-games-dataset9](https://www.businessinsider.com/top-20-best-airlines-in-world-none-in-us-qatar-2022-9)

1.2.2 Giải thích ý nghĩa về tập dữ liệu:

Sau khi tải về dataset dưới dạng file csv từ đường link trên ta sẽ có được data như sau:

Như hình trên có thể thấy mỗi dòng là một tựa video game trong đó:

|  |  |
| --- | --- |
| Tên Cột | Ý Nghĩa |
| Rank | Xếp Hạng Của Game |
| Name | Tên Của Game |
| Platform | Nền Tảng Phát Hành Game |
| Year | Name Phát Hành Game |
| Genre | Thể Loại Game |
| Publisher | Nhà Phát Hành Game |
| NortAmerica\_Sales | Doanh Thu Ở Thị Trường Bắc Mỹ |
| EuropeanUnion\_Sales | Doanh Thu Ở Thị Trường Châu Âu |
| Japan\_Sales | Doanh Thu Ở Thị Trường Nhật Bản |
| Other\_Sales | Doanh Thu Ở Các Thị Trường Khác |
| Global\_Sales | Tổng Doanh Thu Toàn Cầu |

Bài toán cần giải quyết: Dựa trên dataset trên, chọn ra các đỉnh, cạnh và áp dụng các thuật toán mạng xã hội để tìm ra những nền tảng nào là được các nhà phát hành game lưu ý và phát hành trên nó nhiều nhất.

Đề xuất mô hình xử lý:

Lấy nhà phát hành làm Đỉnh

Cạnh là đường nối giữa hai nhà phát hành mà có phát hành game trên cùng 1 nền tảng

# **CHƯƠNG 2: XỬ LÝ VÀ PHÂN THÍCH DỮ LIỆU.**

## 2.1 Làm sạch dữ liệu.

Link Source Code:

Các thư viện được sử dụng:

Text

Description automatically generated

Hình 1: Import thư viện vào file source code

Tiến hành đọc file csv:

Text

Description automatically generated

Hình 2: Đọc file csv

Graphical user interface, text

Description automatically generated

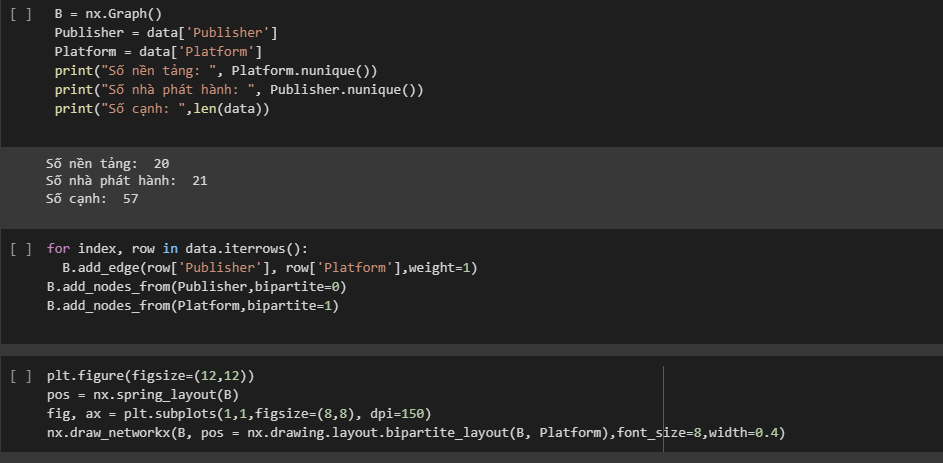
Hình 3: Check null, duplicates và xử lí.

## 2.2 Chuyển đổi dữ liệu trên thành đồ thị.

Đồ thị hai phía.

Node: Là tên của nhà phát hành và nền tảng phát hành.

Edge: Cứ mỗi nhà phát hành phát hành game trên 1 nền tảng sẽ tạo ra 1 cạnh.



Hình 3: Đưa dữ liệu đã tiền xử lí vào đồ thị vô hướng

Hiển thị đồ thị hai phía .

Diagram

Description automatically generated

Hình 5: Đồ thị hai phía.

Nhìn vào đồ thị có thể thấy rất nhiều nhà phát hành có phát hành game ở 2 nền tảng trở lên.

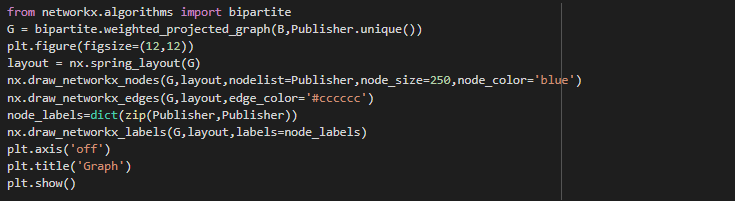
Đồ thị một phía.

Node: là các nhà phát hành.

Edge: hai nhà phát hành cùng phát hành trên cùng 1 nền tảng sẽ tạo nên 1 cạnh.

Điều này nói lên mỗi nhà phát hành có phát hành trên nhiều nền tảng khác nhau và tại mỗi nền tảng sẽ có nhiều nhà phát hành cùng cạnh tranh phát triển.

Weight: Trọng số là số nhà phát hành cùng phát hành trên 1 nền tảng .



Hình 6: Code hiển thị đồ thị 1 phía.

Chart, scatter chart

Description automatically generated

Hình 7: Đồ thị 1 phía

Xuất đồ thị 1 phía ra file csv để thực hiện trên phần mềm Gephi:

Text

Description automatically generated

*Hình 8: Code xuất file csv để thực hiện trên Gephi*

Table

Description automatically generated

Hình 9: file csv vừa xuất để thực hiện trên Gephi

Source và Target là 2 node 2 đầu của cạnh, là 2 nhà phát hành game phát hành ít nhất trên cùng một nền tảng game.

Weight: Trọng số, số nền tảng mà 2 nhà phát hành ở source và target cùng phát hành.

# **CHƯƠNG 3: CÁC THUẬT TOÁN XẾP HẠNG.**

## 3.1 Closeness Centrality

Closeness centrality là độ đo khoảng cách từ một đỉnh đến các đỉnh còn lại trong đồ thị. Đỉnh có giá trị Closenness Centrality cao nhất sẽ là đỉnh có thể truyền đạt, tiếp nhận thông tin từ các node khác trong mạng một cách nhanh nhất, ít tốn thời gian nhất.

Ý nghĩa: những nhà phát hành có closeness centrality cao là những nhà phát hành phát hành game cần phải cạnh tranh với những nhà phát triển khác trên nhiều nền tảng nhất.

### 3.1.1 Sử dụng Python.

Text

Description automatically generated

Hình 9: Code closeness centrality.

A screenshot of a computer

Description automatically generated with medium confidence

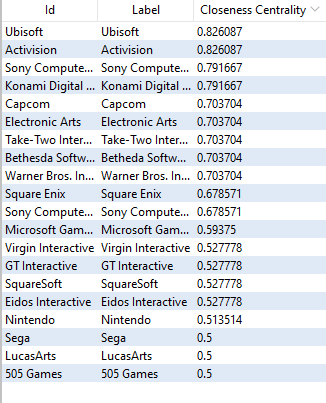
Hình 10: Top 10 Nhà phát hành game có closeness centraltiy cao nhất.

A screenshot of a computer

Description automatically generated with medium confidence

Hình 11: Top 10 Nhà phát hành game có closeness centraltiy nhỏ nhất

### 3.1.2 Sử dụng Gephi.



Hình 13: Nhà phát hành game có closeness centrality cao nhất

Table

Description automatically generated

Hình 14: Nhà phát hành game có closeness centrality thấp nhất

## 3.2: Page Rank.

PageRank là một thuật toán đo lường ảnh hưởng bắc cầu hoặc kết nối của các nút.

Nó có thể được tính toán bằng cách phân phối lặp đi lặp lại thứ hạng của một nút (ban đầu dựa trên mức độ) cho các nút lân cận hoặc bằng cách duyệt ngẫu nhiên biểu đồ và đếm tần suất chạm vào mỗi nút trong những bước này.

Thuật toán áp dụng cho toàn bộ dữ liệu cần được truy xuất, là truy vấn độc lập.

Ý nghĩa: PageRank có độ đo càng cao thì các nhà phát hành đó càng phát hành game trên nhiều nền tảng.

### 3.2.1 Sử dụng Python

*Diagram

Description automatically generated*

*Hình 15: Nhà phát hành có page rank từ cao đến thấp.*

### 3.2.2 Sử dụng Gephi.

*Table

Description automatically generated*

*Hình 16: Nhà phát hành có page rank từ cao đến thấp.*

## 3.3 Betweenness Centrality.

Betweenness centrality được định nghĩa như tổng tỷ số của các đường đi ngắn nhất từ một node tới một node khác đi qua một node cho trước. Độ đo này dung để xem xét khả năng chi phối các quan hệ giữa các nút khác trong mạng. Node có giá trị này càng lớn thì node đó sẽ có sự ảnh hưởng càng lớn đến việc phân bố cấu trúc các cụm hay nhóm trong mạng càng lớn.

Ý nghĩa: những nhà phát hành có betweenness centrality cao là những nhà phát hành phát hành game cần phải cạnh tranh với những nhà phát triển khác trên nhiều nền tảng nhất.

### 3.3.1 Sử dụng Python.

Graphical user interface

Description automatically generated

*Hình 17: Kết quả giá trị Betweenness Centrality.*

Text

Description automatically generated with medium confidence

*Hình 18 Top 10 cao nhất giá trị Betweenness Centrality.*

A screenshot of a computer

Description automatically generated with medium confidence

*Hình 19: Top 10 thấp nhất theo giá trị Betweenness Centrality.*

### 3.3.2 Sử dụng Gephi.

Table

Description automatically generated

*Hình 20: Top 10 cao nhất theo giá trị Betweeness Centrality.*

Table

Description automatically generated

*Hình 21: Top 10 thấp nhất theo giá trị Betweeness Centrality.*

# **CHƯƠNG 4: CÁC THUẬT TOÁN GOM CỤM.**

## 4.1 Thuật toán Louvain.

Thuật toán Louvain là một phương pháp phân chia cộng đồng và thực hiện lặp đi lặp lại việc phân chia cộng đồng nhiều lần để có được mô đun tối đa của toàn bộ mạng.

### 4.1.1 Sử dụng Python.

Text

Description automatically generated

Hình 22: Kết quả chạy Louvain bằng python

### 4.1.2. Visualize với Pivot Table.

Chart

Description automatically generated

Hình 23: Visualize cụm 1 bằng PivotTable.

Chart

Description automatically generated

Hình 24: Visualize cụm 2 bằng PivotTable.

Chart

Description automatically generated

Hình 25: Visualize cụm 3 bằng PivotTable.

Nhận xét: dựa vào đoạn code python ta có thể thấy dataset được phân ra 3 cụm, sau khi visualize bằng Pivot Table em xin đưa ra nhận xét như sau.

Cụm 1: Là cụm mà nền tảng game PS được các nhà nhà phát hành game ở cụm đó cạnh tranh nhiều nhất.

Cụm 2: Là cụm mà nền tảng game Wii được các nhà phát hành game ở cụm đó cạnh tranh nhiều nhất, tiếp theo đó là nền tảng DS.

Cụm 3: Là cụm mà nền tảng game PS3 được các nhà phát hành game ở cụm đó cạnh tranh nhiều nhất, tiếp theo đó là nền tảng X360.

# **CHƯƠNG 5: BÀI TẬP TÍNH TAY CÁC ĐỘ ĐO, GOM CỤM TRÊN 10 NODE NGẪY NHIÊN**

Chart

Description automatically generated

Hình 26: Sơ đồ 10 nodes sau khi chọn.

Table

Description automatically generated

Hình 27: Ma trận có trọng số và Ma trận kề.

Graphical user interface

Description automatically generated with medium confidence

Hình 28: Kết quả tính Betweeness Centrality, Closeness Centrality và Node Betweeness Centrality.

Graphical user interface, application, table, Excel

Description automatically generated

Hình 29: Kết quả cuối cùng và xếp hạng khi tính bằng eigen vector – eigen value.

Table

Description automatically generated

Hình 30: Các bước tính toán khi sử dụng thuật toán pagerank.

Graphical user interface, application, table, Excel

Description automatically generated

Hình 31: Kết quả cuối cùng và xếp hạng của pagerank.

# **CHƯƠNG 6: KẾT LUẬN.**

Sau khi thông qua những phân tích trên em xin rút ra kết luận đó là với tập dataset input ta có thể thấy, nhà phát hành Ubisoft là một nhà phát hành mà phát hành game trên nhiều nền tảng nhất, vì vậy cũng có thể nói họ là nhà phát hành có nhiều đối thủ cạnh tranh nhất trên thị trường, để xác nhận việc đó thì ta có thể dùng thuật toán Page Rank và các thuật toán Centrality để xem xét chỉ số.

Bên cạnh đó nhờ các thuật toán gom cụm như louvain ta có thể xác định được những nhóm nhà phát hành có điểm chung giống nhau đó là cùng phát hành trên một số nền tảng, hoặc là xác định được một số nền tảng mà có nhiều nhà phát hành phát hành nó nhất.