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

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN**

---------



ĐỒ ÁN MÔN HỌC

**MẠNG XÃ HỘI**

**IS353.N22.HTCL**

**< DATASET BOOKING HOTEL >**

**Sinh viên thực hiện:**

Lê Hoàng Huy 20521392

Huỳnh Thị Thanh Ngân 20521645

**Giảng viên:**

Nguyễn Thị Kim Phụng

Thành phố Hồ Chí Minh, tháng 5 năm 2023

**NHẬN XÉT GIÁO VIÊN**

......................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

.............................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

**Mục Lục**

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 2](#_Toc136480344)

[1.Lý do chọn đề tài: 2](#_Toc136480345)

[2.Nội dung đề tài 3](#_Toc136480346)

[3. Xác định bài toán 3](#_Toc136480347)

[4. Mô tả dữ liệu gốc 3](#_Toc136480348)

[CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU VÀ TRỰC QUAN HÓA DỮ LIỆU 4](#_Toc136480349)

[1. Tiền xử lý dữ liệu: 4](#_Toc136480350)

[2. Tạo mạng liên kết: 5](#_Toc136480351)

[3. Trực quan hóa mạng liên kết bằng code python 5](#_Toc136480352)

[4. Trực quan hóa đồ thị 1 phía trên gephi 8](#_Toc136480353)

[CHƯƠNG 3: PHÂN TÍCH VÀ TRỰC QUAN HÓA BẰNG CÁC THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG 11](#_Toc136480354)

[1. Thuật toán phát hiện cộng đồng Louvain 11](#_Toc136480355)

[2. Thuật toán Girvan Newman 16](#_Toc136480356)

[CHƯƠNG 4: PHÂN TÍCH VÀ TRỰC QUAN HÓA BẰNG CÁC THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG 20](#_Toc136480357)

[1. Đo độ Closeness centrality. 20](#_Toc136480358)

[2.Đo độ Betweeness centrality. 23](#_Toc136480359)

[3. Thuật toán PageRank 26](#_Toc136480360)

[4. Eigenvector và Eigenvalue 28](#_Toc136480361)

[4.1. Eigenvector centrality 28](#_Toc136480362)

[4.2. Eigenvalue 30](#_Toc136480363)

[TÀI LIỆU THAM KHẢO 31](#_Toc136480364)

[[1] Link Code Python (Google Colab) 31](#_Toc136480365)

[[2] CodePython\_Sampleupdate.doc 31](#_Toc136480366)

[[3] Link Dataset 31](#_Toc136480367)

# [CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI](#_Toc56371969)

## Lý do chọn đề tài:

Đề tài về dataset booking hotel là một lựa chọn tuyệt vời để chạy các thuật toán centrality. Lý do vì đây là một lĩnh vực đang phát triển mạnh mẽ trong ngành du lịch và khách sạn, với một lượng dữ liệu lớn và đa dạng. Dữ liệu booking hotel bao gồm thông tin về các lượt đặt phòng, vị trí khách sạn, giá cả, độ phổ biến của các dịch vụ, đánh giá của khách hàng về chất lượng dịch vụ và các tiện ích khác.

Việc chạy các thuật toán centrality sẽ giúp phân tích và tìm ra các yếu tố quan trọng nhất trong việc đặt phòng khách sạn. Ví dụ như thuật toán degree centrality có thể giúp xác định các khách sạn đang được quan tâm nhiều nhất hoặc thuật toán betweenness centrality giúp phân tích các địa điểm vận chuyển quan trọng nhất để tiện cho khách hàng đến khách sạn.

Kết quả phân tích của các thuật toán centrality giúp cho các nhà quản lý khách sạn có thông tin để tối ưu hóa chất lượng dịch vụ và quản lý thị trường, từ đó cải thiện năng suất kinh doanh và tăng lợi nhuận. Việc chạy các thuật toán centrality trên dataset booking hotel còn có thể giúp đưa ra các đề xuất về phân khúc khách hàng, nhằm tạo ra các gói dịch vụ phù hợp với các nhu cầu khác nhau của khách hàng. Chẳng hạn, các thuật toán clustering có thể phân cụm các loại khách hàng khác nhau và đưa ra các gợi ý dịch vụ tương ứng để thu hút khách hàng đến với khách sạn.

Ngoài ra, dataset booking hotel cũng cung cấp thông tin phong phú về các yếu tố liên quan đến khách sạn như địa điểm, giá cả, chất lượng dịch vụ, tiện ích, vị trí và lịch sử đặt phòng. Các thuật toán tìm kiếm tương tự (similarity searching) và thuật toán recommendation system có thể sử dụng các thông tin này để đưa ra các gợi ý khách sạn và gói dịch vụ phù hợp cho khách hàng.

Tóm lại, chọn đề tài về dataset booking hotel để chạy các thuật toán centrality là một lựa chọn hợp lý trong việc nghiên cứu và phân tích trong lĩnh vực du lịch và khách sạn. Việc áp dụng các thuật toán này sẽ giúp tối ưu hoá chất lượng dịch vụ, quản lý thị trường và tăng lợi nhuận, đồng thời cũng giúp tạo ra các gợi ý phù hợp cho khách hàng và tăng cường trải nghiệm của họ khi ở khách sạn.

## Nội dung đề tài

* Tên dataset : Booking Hotel
* Link dataset : <https://www.kaggle.com/datasets/mojtaba142/hotel-booking>
* Mô tả : Dữ liệu ban đầu được lấy từ bài viết Bộ dữ liệu nhu cầu đặt phòng khách sạn, được viết bởi Nuno Antonio, Ana Almeida và Luis Nunes cho Tóm tắt dữ liệu, Tập 22, tháng 2 năm 2019.
* Bộ dữ liệu này chứa 119390 quan sát đối với Khách sạn Thành phố và Khách sạn Nghỉ dưỡng. Mỗi quan sát đại diện cho một đặt phòng khách sạn trong khoảng thời gian từ ngày 1 tháng 7 năm 2015 đến ngày 31 tháng 8 năm 2017, bao gồm cả đặt phòng đã đến và đặt phòng đã bị hủy.
* Công cụ hỗ trợ :

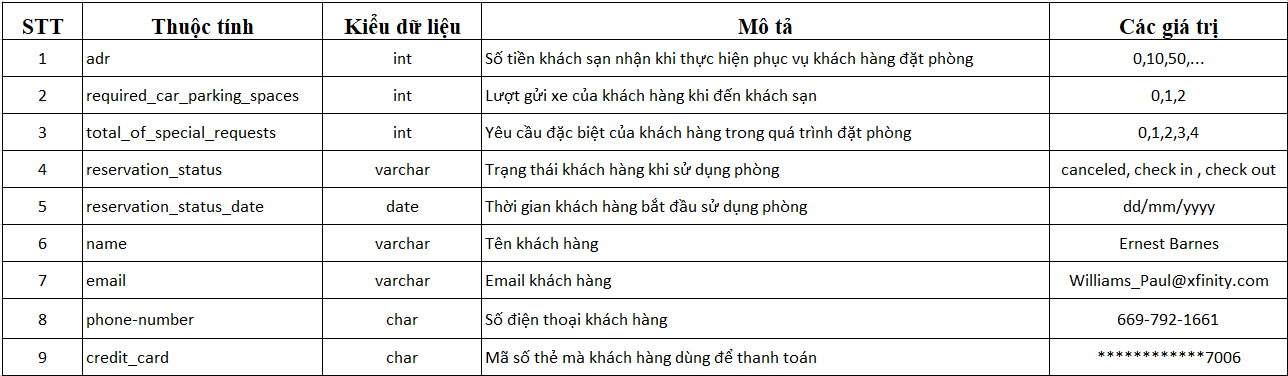
- Code python : Google Colab

- Thực hiện thuật toán và vẽ đồ thị : Gephi

## 3. Xác định bài toán

* Input : Tập dữ liệu ban đầu trên nguồn dữ liệu của kaggle được tải về để thực hiện đồ án.
* Output : Đưa ra độ đo, đưa ra cộng đồng phục vụ phân tích trực quan hóa bằng các thuật toán để giải quyết dataset : “ Booking Hotel”

## 4. Mô tả dữ liệu gốc

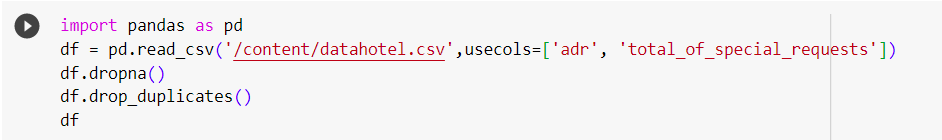


*Hình 1. Mô tả dữ liệu*

[CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU VÀ TRỰC QUAN HÓA DỮ LIỆU](#_Toc56371969)

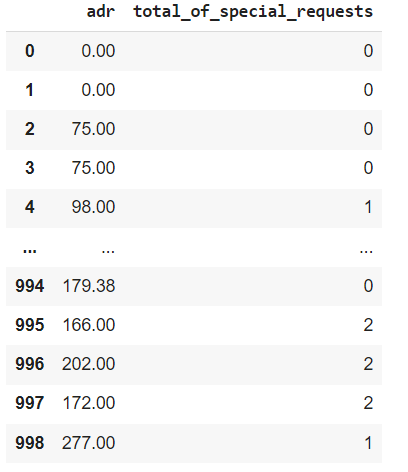
## 1. Tiền xử lý dữ liệu:

**\*** Làm sạch dữ liệu : Bộ dữ liệu nhìn chung có vài thuộc tính bị thiếu dữ liệu,cần loại bỏ khỏi dataset.



*Hình 2. Code python làm sạch và tách dữ liệu*

\* Kết quả sau khi tách 2 cột ra khỏi dataset



*Hình 3. Làm sạch và tách dữ liệu*

## 2. Tạo mạng liên kết:

**-** Sử dụng đoạn code python để tính số nút và số cạnh của đồ thị từ dataset đã tách



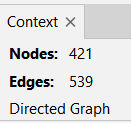
*Hình 4. Đếm số cạnh và nút của đồ thị bằng code python*

Có thể thấy sau khi tách và làm sạch dữ liệu để chọn 2 cột đối lập nhau thì số cạnh và số nút của dữ liệu đã giảm xuống và hoàn toàn khớp với dữ liệu bên Gephi

\* Google colab:

* Số cạnh của đồ thị là: 539
* Số nút của đồ thị là : 421

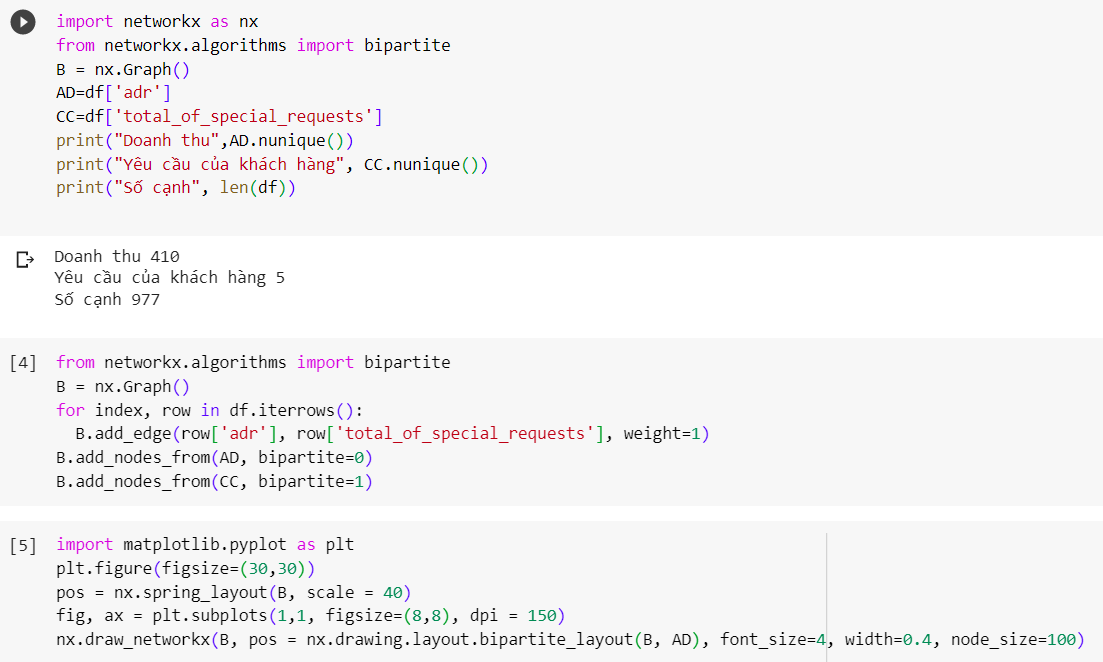
\* Gephi:



*Hình 5. Nút và cạnh của đồ thị khi thêm dữ liệu vào gephi*

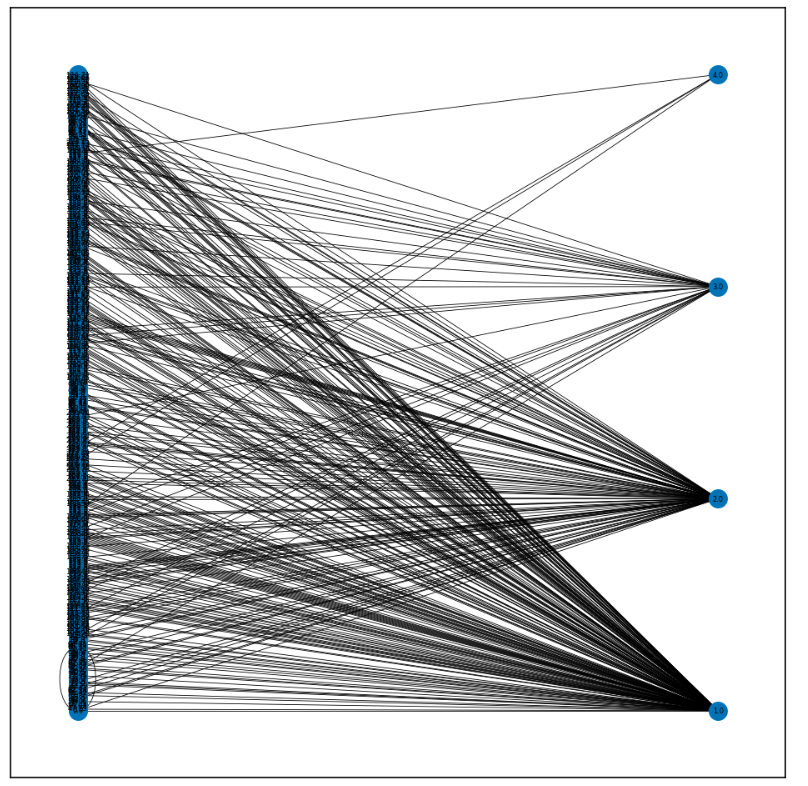
## 3. Trực quan hóa mạng liên kết bằng code python

**\*** Code đồ thị 2 phía trên Google colab :



*Hình 6. Code python vẽ đồ thị 2 phía*

**\*** Kết quả đồ thị 2 phía :

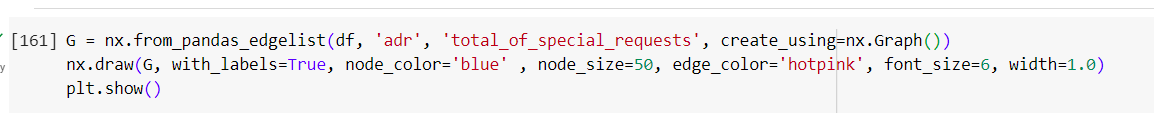


*Hình 7. Đồ thị 2 phía*

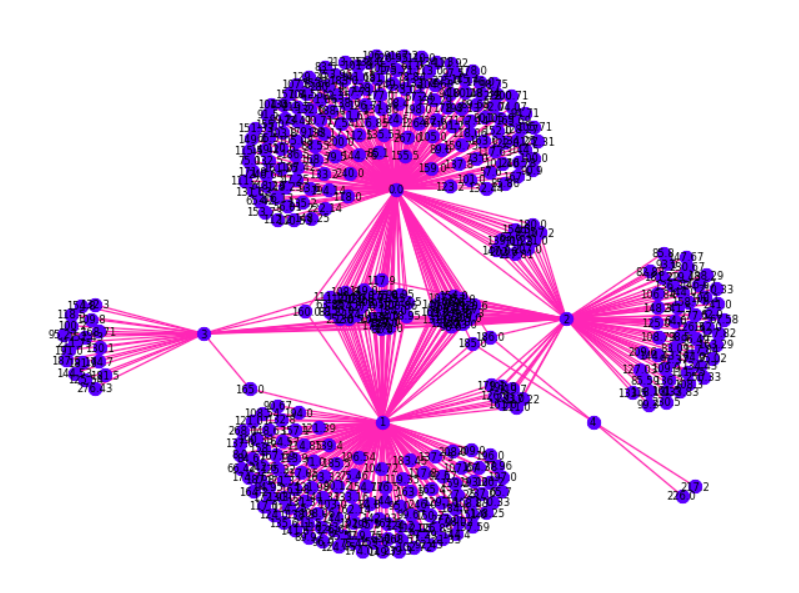
**\* Nhận xét :** Nếu tổng số yêu cầu đặc biệt mà khách hàng yêu cầu trong quá trình đặt phòng ít, có thể cho thấy rằng các khách hàng của khách sạn không quá đòi hỏi và yêu cầuhiều dịch vụ hơn so với những khách sạn khác. Điều này có thể làm tăng tính khả thi của khách sạn về việc đáp ứng các yê cầu đặc biệt của khách hàng, đồng thời giảm thiểu thành công việc đặt phòng khách hàng bị hủy hoặc không xuất hiện.

## 4. Trực quan hóa đồ thị 1 phía trên gephi

* Node: Số doanh thu khách sạn nhận được từ khách hàng có yêu cầu đặc biệt trong quá trình đặt phòng.
* Edge: Hai nút doanh thu khách sạn nhận được từ khách hàng có yêu cầu đặc biệt trong quá trình đặt phòng được nối với nhau tạo thành cạnh, ý nghĩa nói lên sự thu nhập doanh thu của dịch vụ phát triển khi có chung số lượng yêu cầu đặc biệt.
* Weight: Trọng số là số yêu cầu đặc biệt của khách hàng trùng nhau.

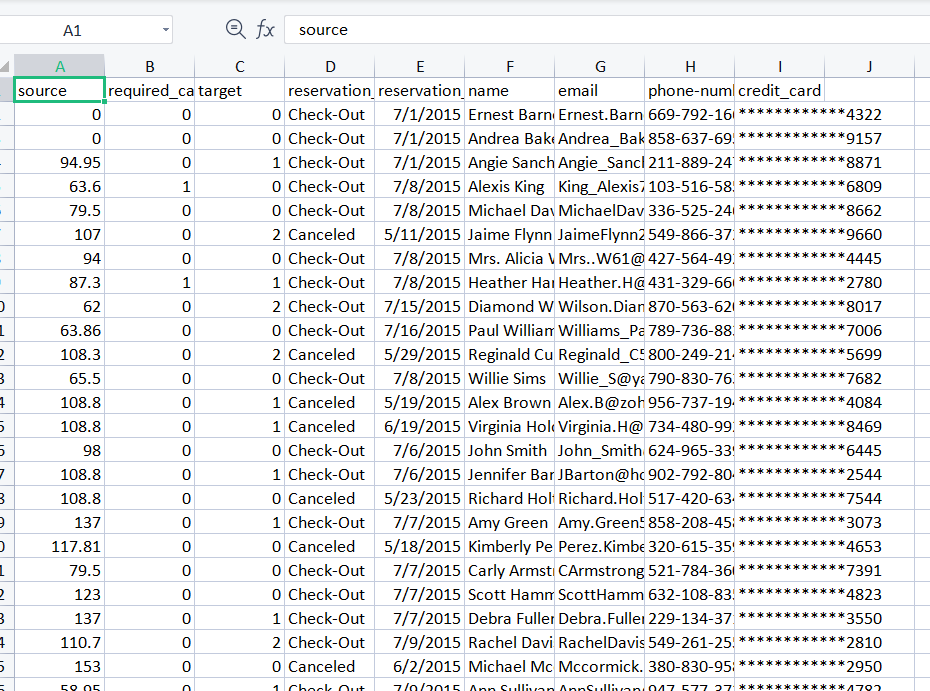


*Hình 8. Đồ thị 1 phía thực thi bằng code python*



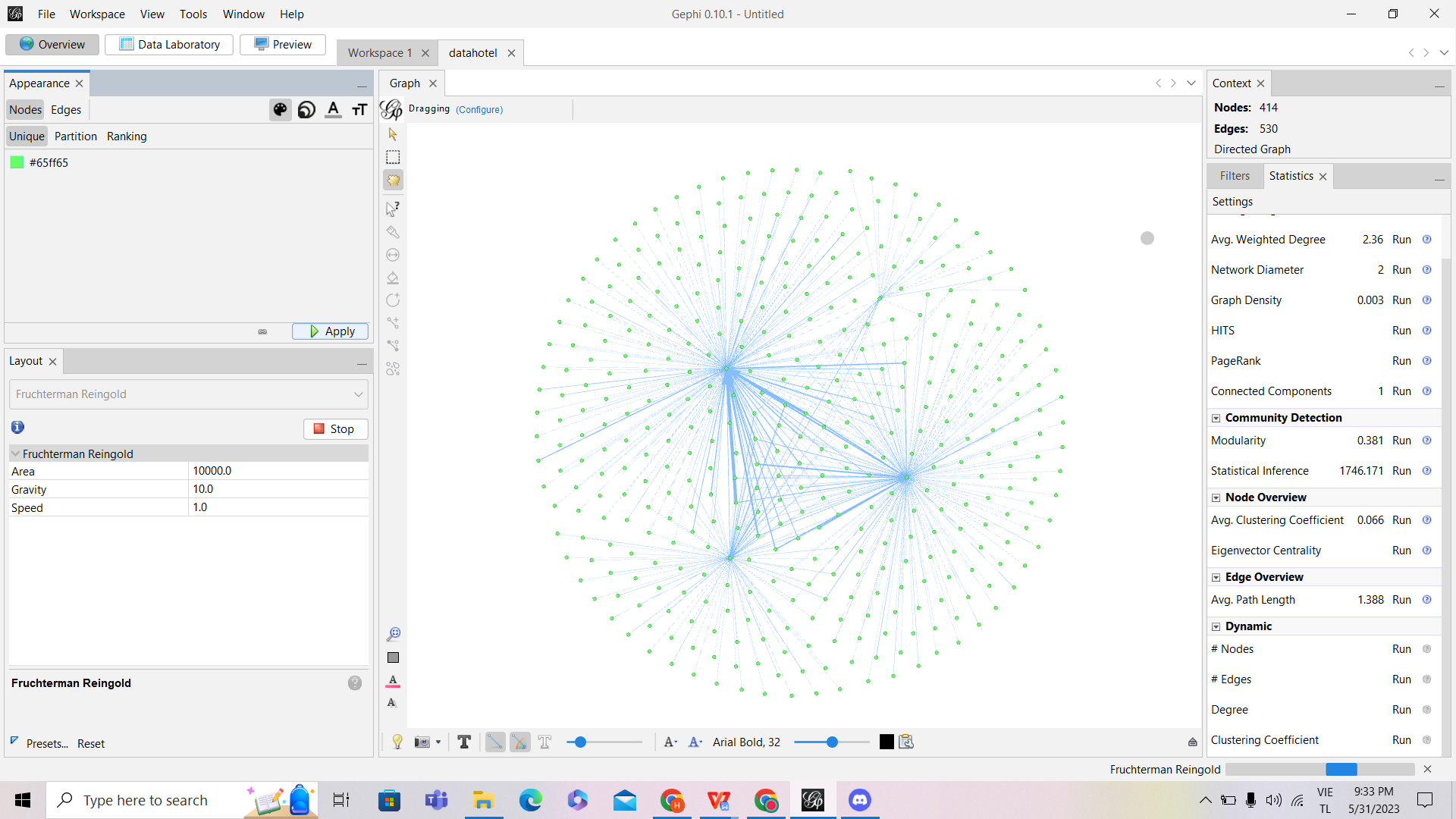
*Hình 9. Kết quả đồ thị 1 phía thực thi bằng code python*

* Xuất dữ liệu thành file csv rồi tải về để thực hiện sẽ đồ thị trên gephi

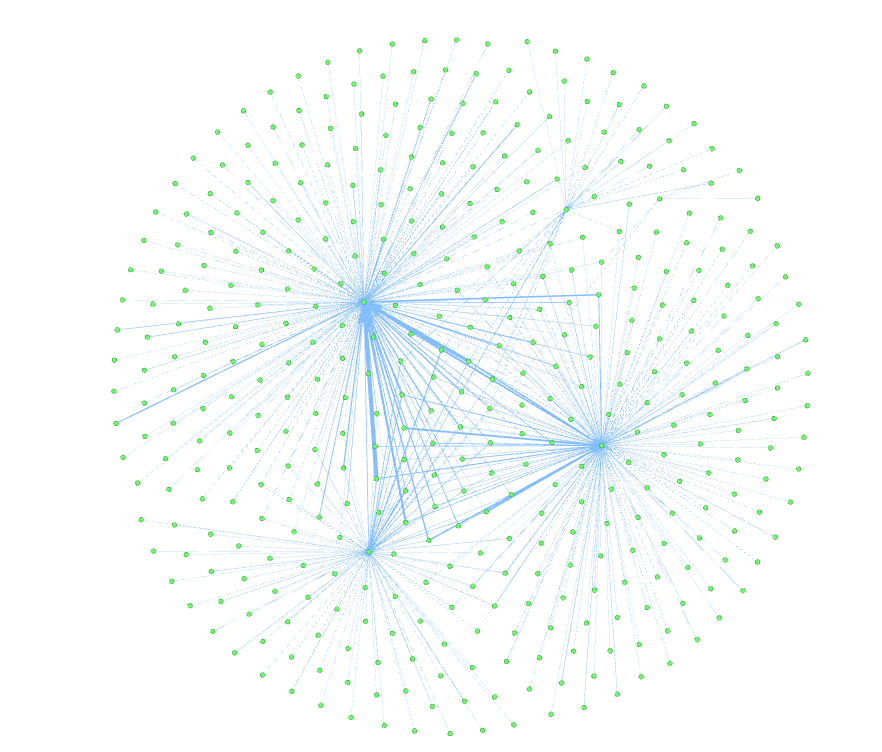


*Hình 10. File csv tải về*

* Đồ thị 1 phía trên gephi



*Hình 11. Đồ thị 1 phía trên gephi*

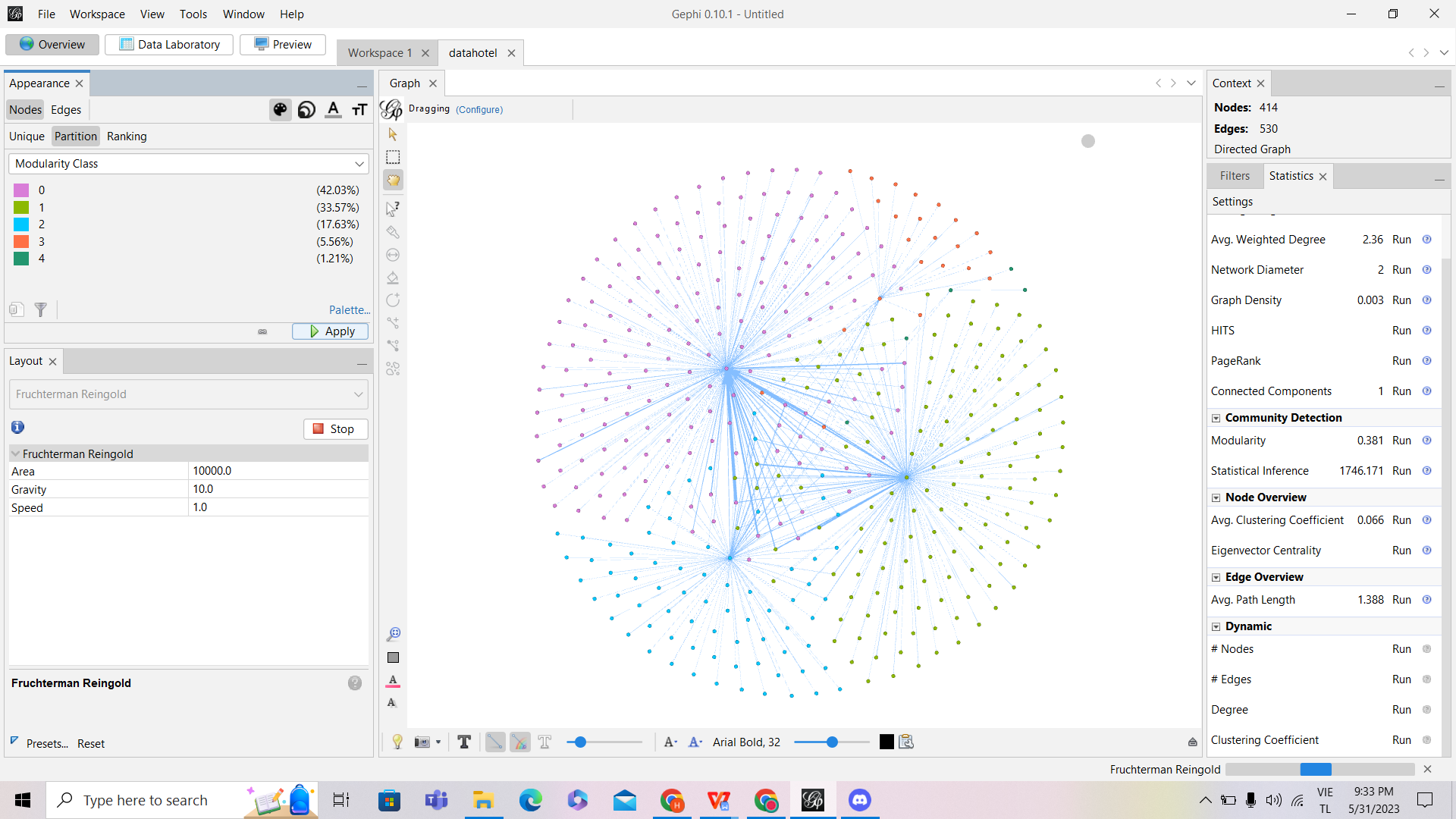


*Hình 12. Đồ thị 1 phía trên gephi full*

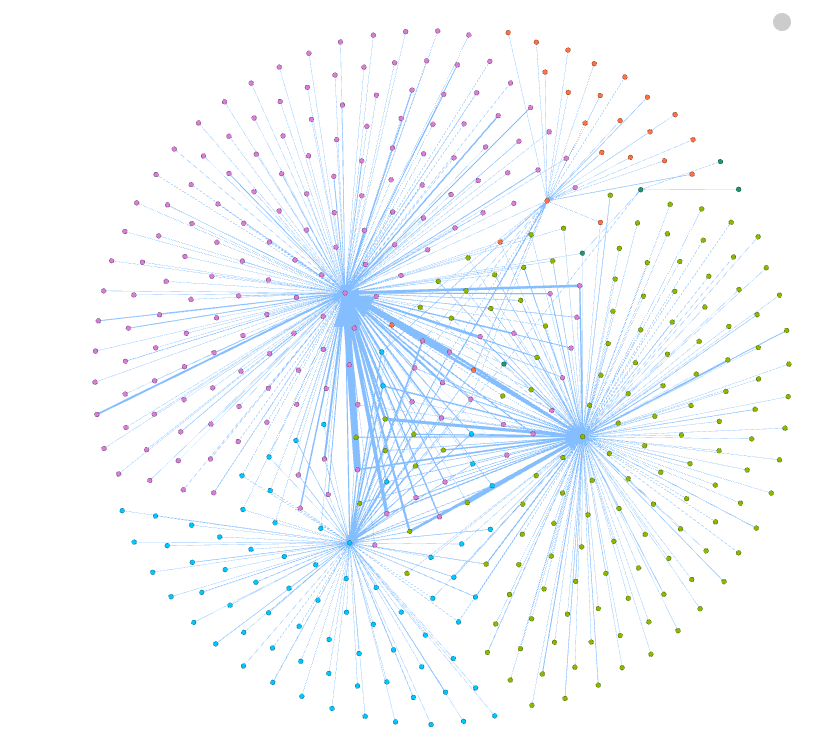
# CHƯƠNG 3: PHÂN TÍCH VÀ TRỰC QUAN HÓA BẰNG CÁC THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG

## 1. Thuật toán phát hiện cộng đồng Louvain

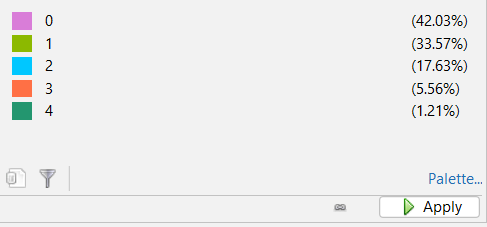
**\*** Thực thi thuật toán Louvain bằng Gephi



*Hình 13. Sử dụng thuật toán Louvain*



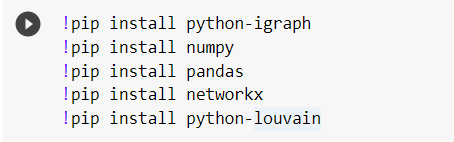
*Hình 14. Sử dụng thuật toán Louvain full*



*Hình 15. Thuật toán chia thành 5 cụm màu*

**Nhận xét**: Thuật toán chia ra thành 5 cụm

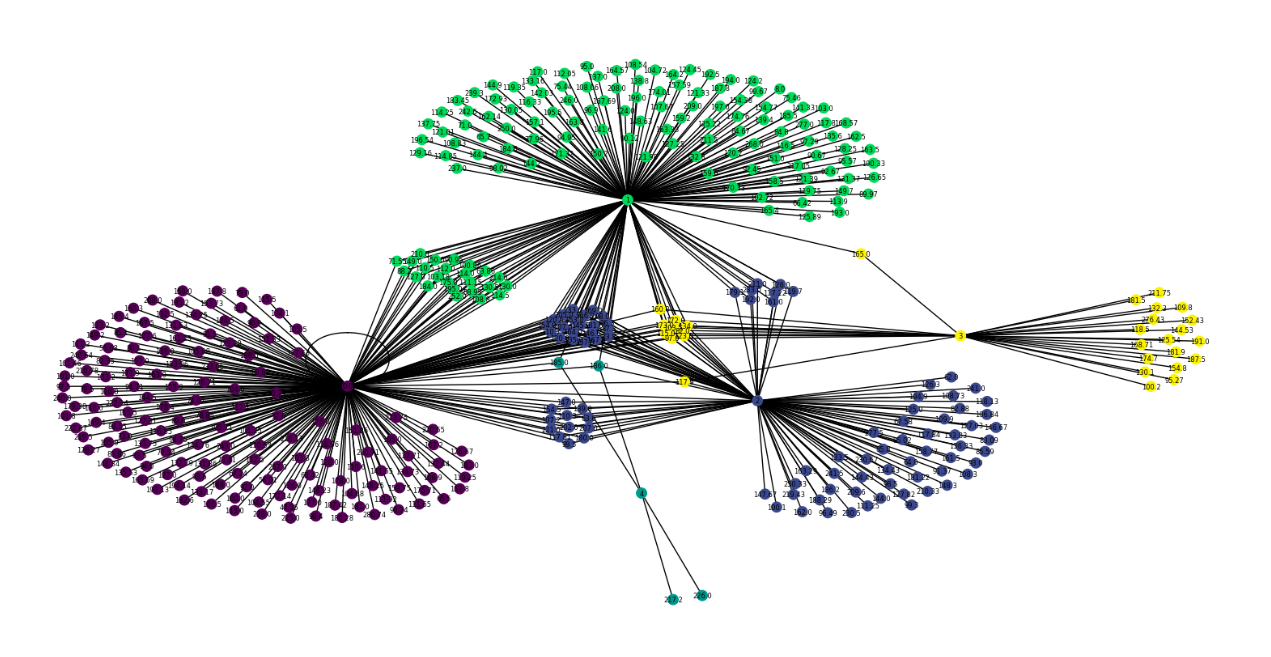
\* Thực thi bằng Python



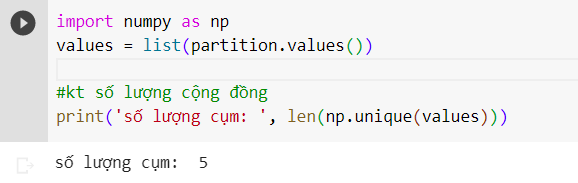
*Hình 16. Khai báo , cài đặt thư viện code python*



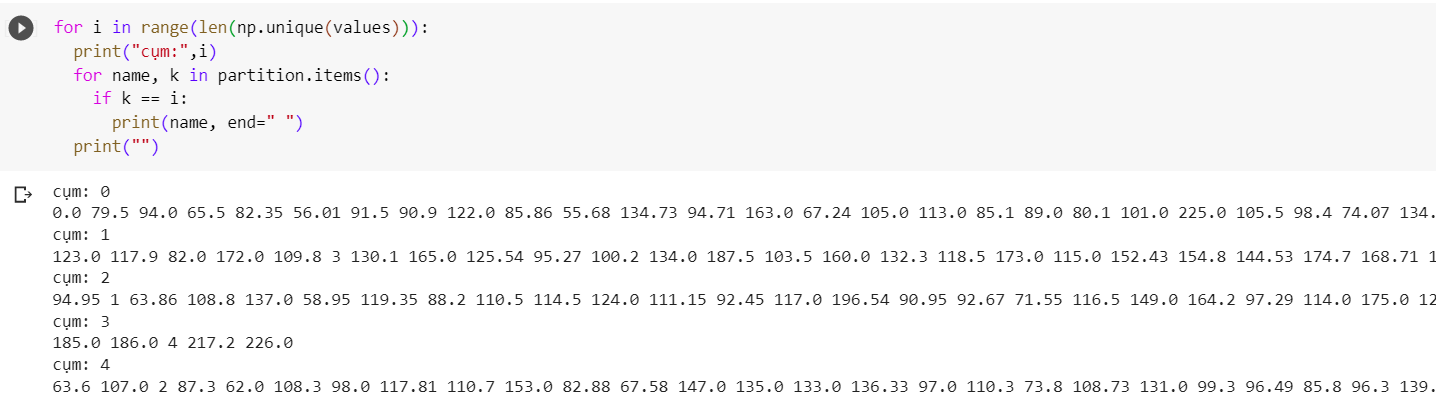
*Hình 17. Code thực thi thuật toán louvain bằng python*



*Hình 18. Đồ thị thuật toán louvain bằng python*



*Hình 19. Số lượng cụm của thuật toán Louvain*

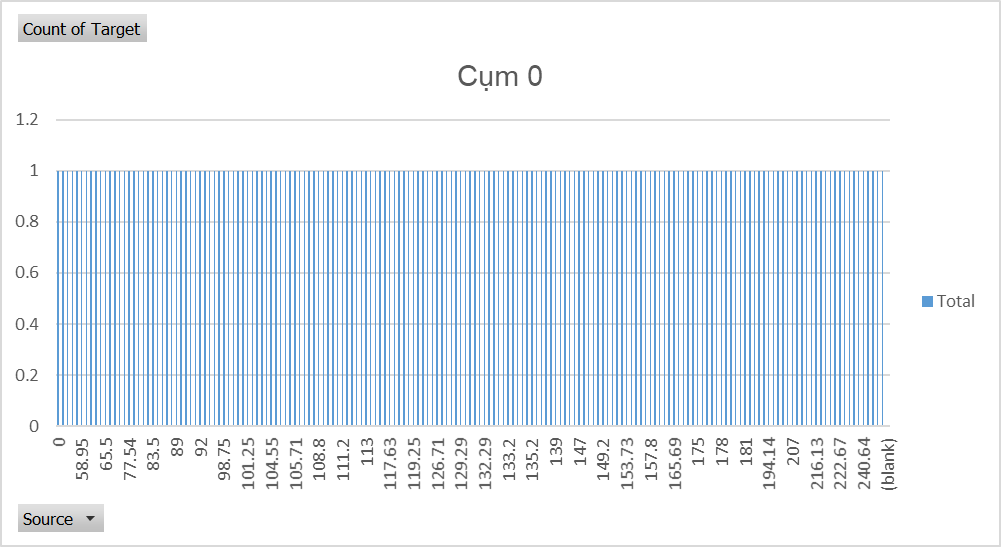


*Hình 20. Code đếm số lượng cụm*

***Số cụm : 5***

\*Cụm 0: [0.0, 79.5, 94.0, 65.5, 82.35, 56.01, 91.5, 90.9, 122.0, 85.86, 55.68, 134.73, 94.71, 163.0, 67.24, 105.0, 113.0, 85.1, 89.0, 80.1, 101.0, 225.0, 105.5, 98.4, 74.07, 134.1, 117.63, 47.25, 73.0, 93.6, 83.5, 61.0, 89.68, 155.0, 105.08, 65.42, 119.25, 149.4, 152.0, 159.75, 112.2, 110.6, 105.9, 153.96, 126.67, 100.0, 104.68, 213.75, 178.0, 129.29, 132.29, 105.71, 78.84, 90.71, 128.0, 103.8, 175.71, 110.0, 134.25, 101.8, 73.41, 132.44, 106.9, 131.86, 216.13, 98.75, 148.23, 123.2, 198.0, 151.33, 144.5, 155.5, 107.6, 133.75, 92.0, 128.27, 136.0, 168.3, 132.6, 132.5, 249.0, 186.5, 177.0, 153.73, 124.5, 110.53, 117.71, 116.85, 135.2, 111.65, 104.0, 133.17, 240.64, 167.5, 148.34, 111.2, 200.0, 151.86, 178.4, 101.46, 112.5, 222.67, 161.25, 118.0, 240.0, 176.0, 131.63, 233.05, 88.55, 133.2, 167.2, 115.5, 188.0, 141.65, 99.24, 280.74, 137.8, 126.71, 219.5, 111.92, 57.6, 157.8, 118.06, 177.14, 222.14, 194.14, 193.13, 210.78, 135.53, 180.28, 159.0, 104.55, 144.76, 220.55, 77.54, 171.9, 148.25, 183.0, 149.2, 267.0, 181.0, 200.71, 228.0, 138.0, 165.69, 127.31, 145.4, 101.25, 179.38]

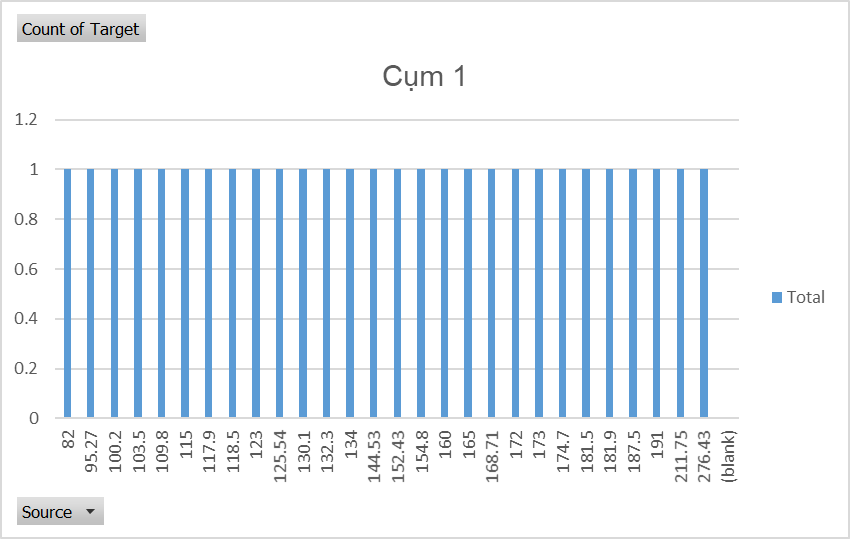
**Biểu đồ cụm 0 có ý nghĩa :** Cụm 0 chứa các giá trị giá cả thấp đến trung bình. Đa dạng về giá trị và không có sự tương đồng. Trong cụm này có giá trị cao nhất là 280.74 và thấp nhất là 47.25.



*Hình 21. Cụm 0*

\*Cụm 1: [123.0, 117.9, 82.0, 172.0, 109.8, 3, 130.1, 165.0, 125.54, 95.27, 100.2, 134.0, 187.5, 103.5, 160.0, 132.3, 118.5, 173.0, 115.0, 152.43, 154.8, 144.53, 174.7, 168.71, 181.5, 181.9, 191.0, 211.75, 276.43]

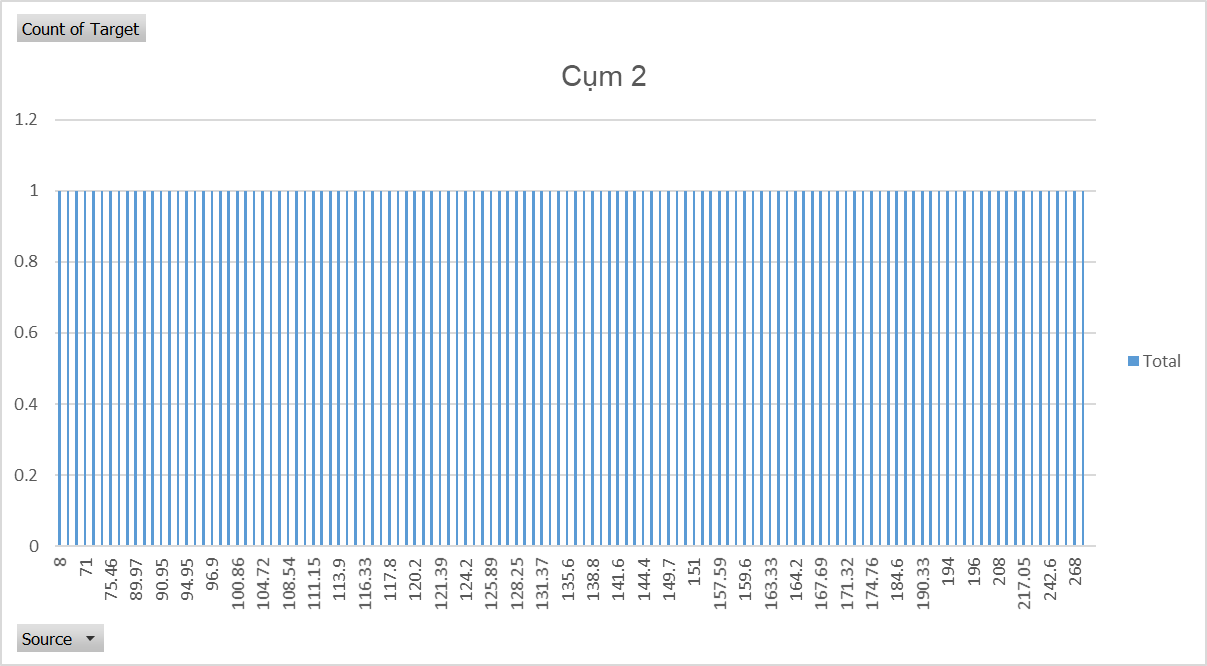
**Biểu đồ cụm 1 có ý nghĩa :** Cụm 1 chứa các giá trị giá cả trung bình đến cao. Có các giá trị tập trung trong khoảng từ khoảng 100 đến 200. Các giá trị cao nhất trong cụm là 276.43 và thấp nhất là 82.



*Hình 22. Cụm 1*

\*Cụm 2: [94.95, 1, 63.86, 108.8, 137.0, 58.95, 119.35, 88.2, 110.5, 114.5, 124.0, 111.15, 92.45, 117.0, 196.54, 90.95, 92.67, 71.55, 116.5, 149.0, 164.2, 97.29, 114.0, 175.0, 124.45, 171.32, 210.0, 117.8, 71.0, 151.0, 157.1, 185.5, 195.0, 193.0, 150.0, 126.65, 104.72, 66.42, 77.96, 112.0, 75.46, 197.0, 108.83, 90.67, 144.4, 139.4, 89.97, 144.9, 113.9, 164.57, 125.22, 130.0, 133.16, 120.2, 130.05, 196.0, 100.86, 132.8, 144.2, 163.8, 130.5, 184.6, 8.0, 127.0, 65.7, 168.57, 107.67, 96.9, 190.33, 135.6, 184.0, 121.33, 150.2, 75.44, 127.25, 84.8, 116.33, 137.75, 98.02, 95.0, 128.25, 158.5, 131.37, 129.16, 214.0, 141.33, 121.01, 172.93, 192.5, 141.6, 163.33, 95.57, 112.05, 103.18, 119.75, 167.69, 194.0, 217.05, 149.7, 159.2, 121.98, 142.03, 90.12, 114.25, 209.0, 174.01, 124.2, 183.45, 159.6, 102.72, 114.85, 138.8, 108.54, 165.4, 108.06, 148.63, 125.89, 170.33, 252.0, 237.0, 162.5, 195.5, 163.5, 174.76, 162.14, 187.8, 242.6, 268.0, 239.3, 154.38, 121.39, 250.0, 208.0, 157.59, 246.0, 211.5, 277.0]

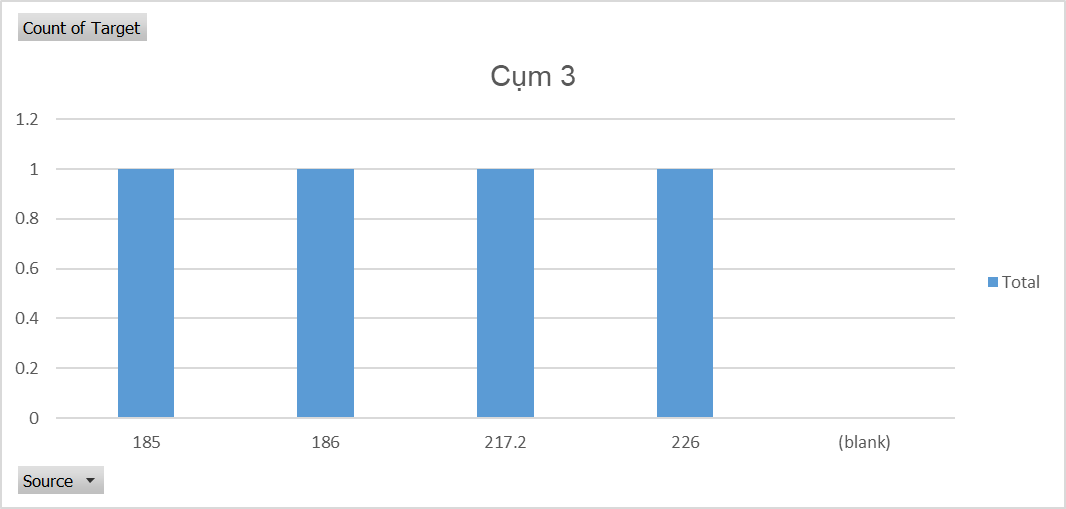
**Biểu đồ cụm 2 có ý nghĩa :** Cụm 2 chứa các giá trị giá cả thấp đến trung bình, nhưng cũng có một số giá trị cao. Bao gồm rất nhiều giá trị, bao phủ một khoảng giá trị rộng từ khoảng 60 đến 280. Các giá trị cao nhất trong cụm là 277 và thấp nhất là 1. Tuy nhiên, số lượng giá trị ở phần lớn khoảng giá trị nhỏ.



*Hình 23. Cụm 2*

\*Cụm 3: [185.0, 186.0, 4, 217.2, 226.0]

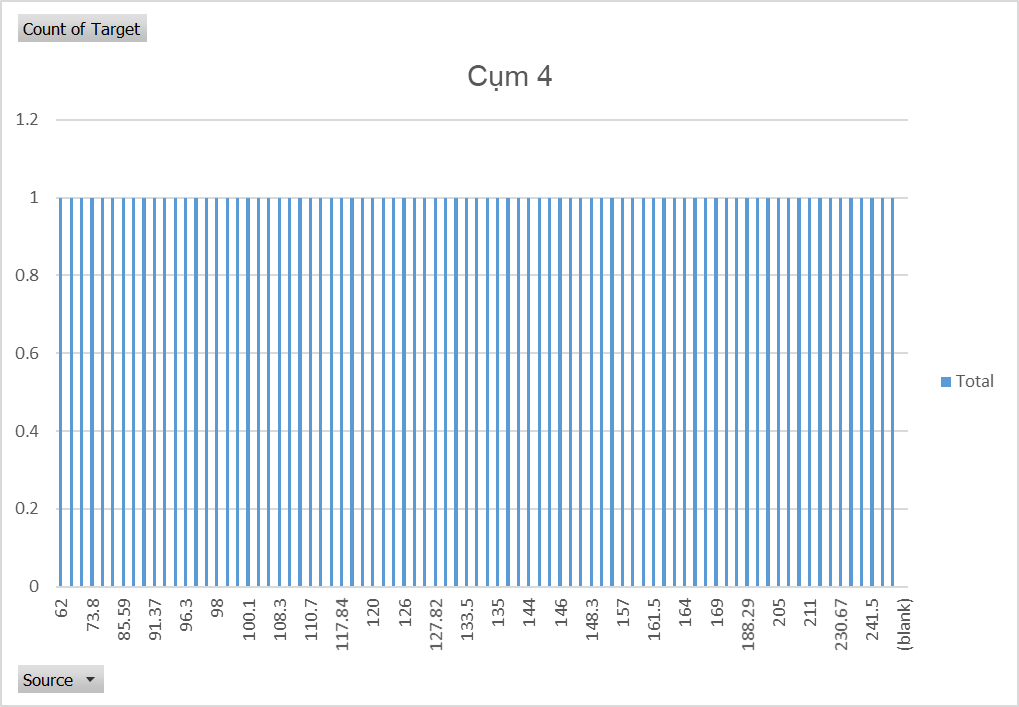
**Biểu đồ cụm 3 có ý nghĩa :** Cụm 3 chỉ có một số giá trị và giá cả của chúng cao hơn rất nhiều so với các cụm khác. Bao gồm các giá trị cao trong khoảng từ 185 đến 226, có chỉ một giá trị nằm trong khoảng còn lại. Điểm chung giữa các giá trị trong cụm này là chúng đều nằm trong khoảng 185-226.



*Hình 24. Cụm 3*

\*Cụm 4: [63.6, 107.0, 2, 87.3, 62.0, 108.3, 98.0, 117.81, 110.7, 153.0, 82.88, 67.58, 147.0, 135.0, 133.0, 136.33, 97.0, 110.3, 73.8, 108.73, 131.0, 99.3, 96.49, 85.8, 96.3, 139.0, 167.0, 85.59, 120.6, 91.37, 146.0, 100.1, 118.13, 120.0, 106.84, 161.0, 134.43, 144.43, 133.83, 99.5, 166.0, 111.25, 169.0, 93.0, 117.22, 126.0, 148.3, 119.7, 64.0, 107.2, 98.5, 181.22, 83.09, 136.2, 161.5, 230.67, 180.0, 109.9, 164.0, 125.0, 145.0, 194.9, 192.0, 126.3, 154.0, 147.67, 157.0, 241.5, 127.03, 154.5, 117.84, 179.1, 207.0, 233.0, 163.29, 158.77, 95.02, 162.0, 219.43, 250.33, 146.67, 202.0, 211.0, 188.29, 127.82, 144.0, 209.6, 205.0, 221.0, 230.5, 241.0, 277.5, 133.5, 210.33]

**Biểu đồ cụm 4 có ý nghĩa :** Cụm 4 chứa các giá trị giá cả thấp đến trung bình. Tuy nhiên, có một số giá trị khá thấp. bao gồm các giá trị thấp phân bố khá đều. Các giá trị cao nhất trong cụm là 277.5 và thấp nhất là 2. Cụm này cũng có thể đại diện cho các kết quả hoặc tình huống trung bình. Tuy nhiên, sự phân bố đều hơn so với cụm 1.



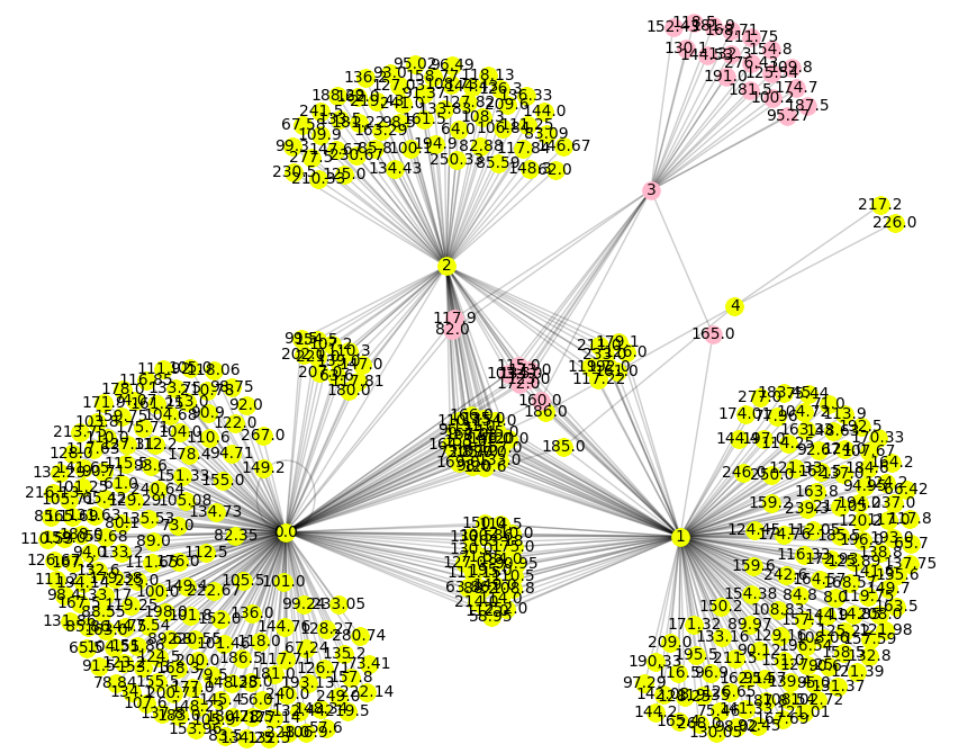
*Hình 25. Cụm 4*

## 2. Thuật toán Girvan Newman

\* Code python

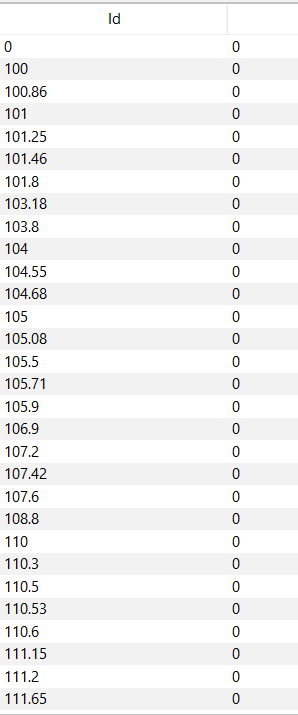


*Hình 26. Code python về thuật toán Girvan Newman*

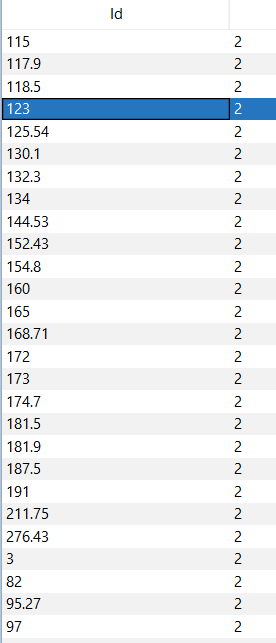


*Hình 27. Đồ thị thể hiện thuật toán Girvan Newman*

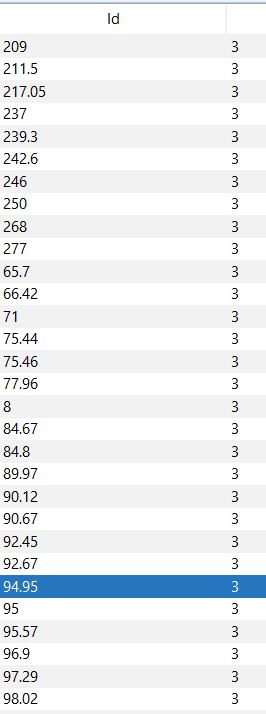
\* Thực thi trên Gephi



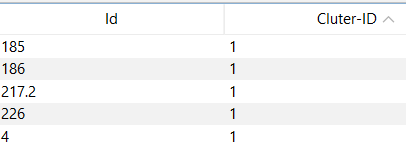
*Hình 28. Cụm 0 trên gephi*



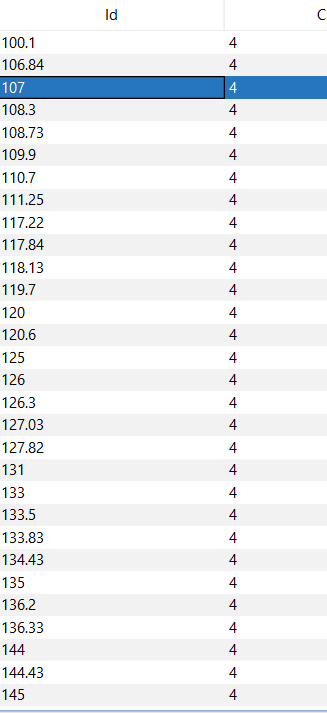
*Hình 29. Cụm 1 trên gephi*



*Hình 30. Cụm 2 trên gephi*



*Hình 31. Cụm 3 trên gephi*

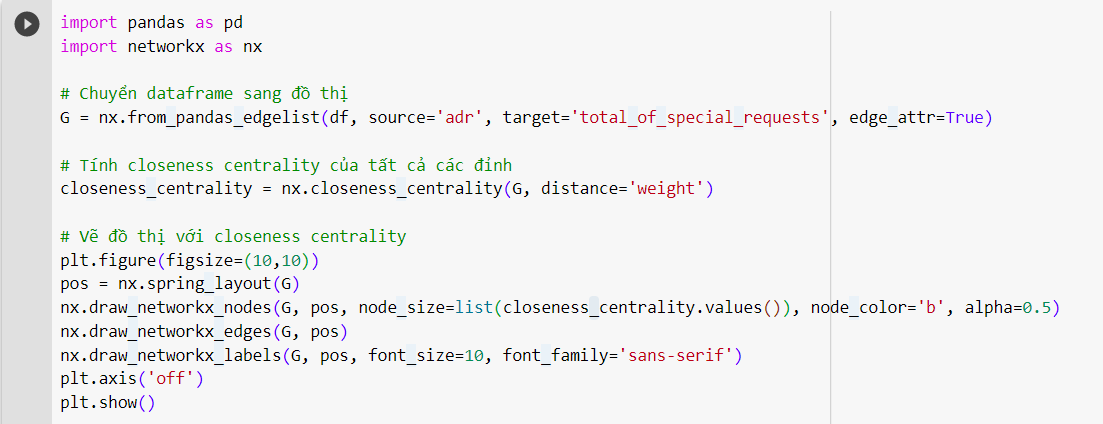


*Hình 32. Cụm 4 trên gephi*

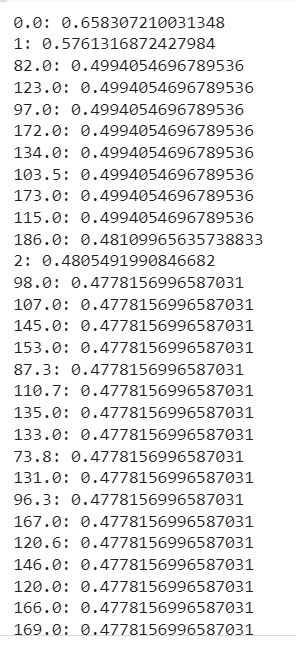
# CHƯƠNG 4: PHÂN TÍCH VÀ TRỰC QUAN HÓA BẰNG CÁC THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG

## 1. Đo độ Closeness centrality.

\* Thực thi bằng code python

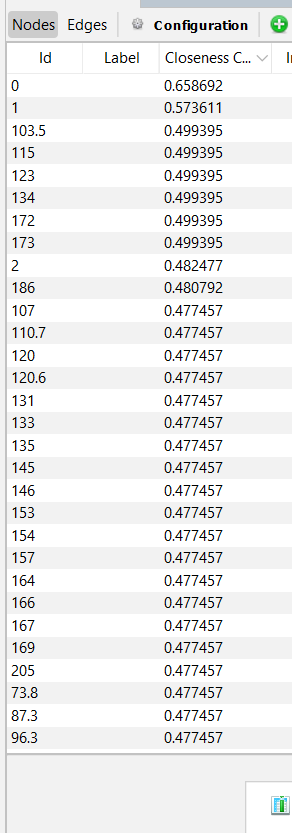


*Hình 33. code python closeness centrality*



*Hình 34. Kết quả code python closeness centrality*

\* Thực thi bằng gephi

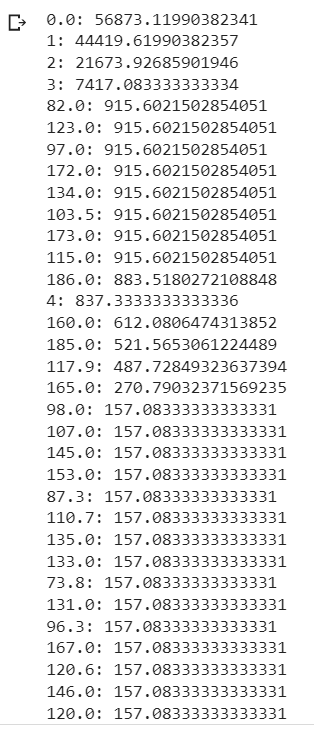


*Hình 35. Closeness centrality trên gephi*

## Đo độ Betweeness centrality.

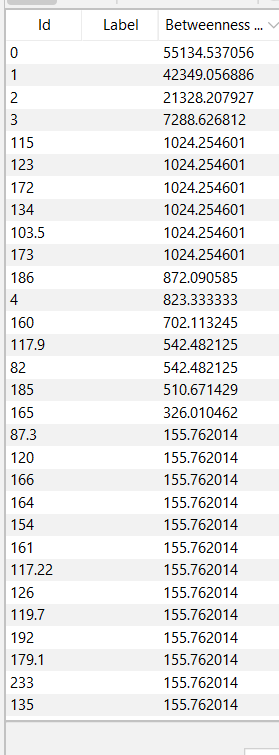


*Hình 36. Code python betweeness centrality*



*Hình 37. Kết quả code python betweeness centrality*

\* Thực thi betweeness centrality trên gephi



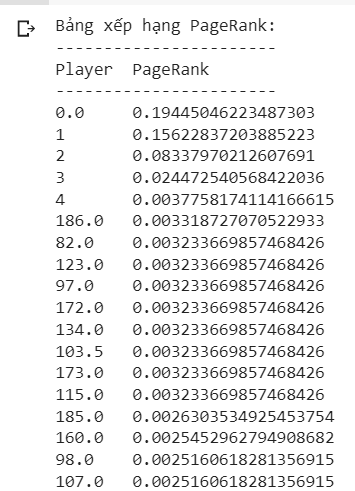
*Hình 38. Betweeness centrality trên gephi*

## Thuật toán PageRank

\* Thực thi trên python

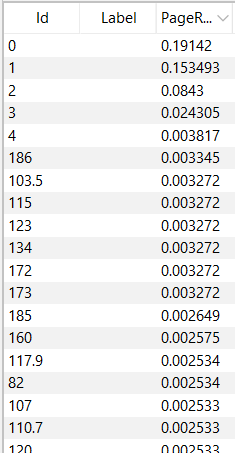


*Hình 39. Code python pagerank*



*Hình 40. Kết quả code python pagerank*

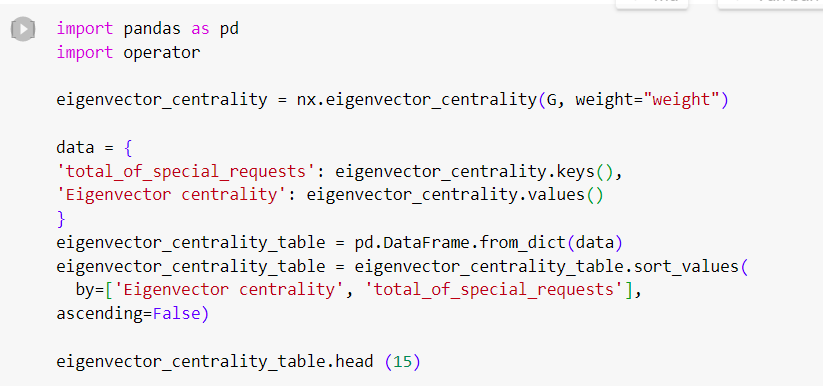
\* Thực thi trên Gephi



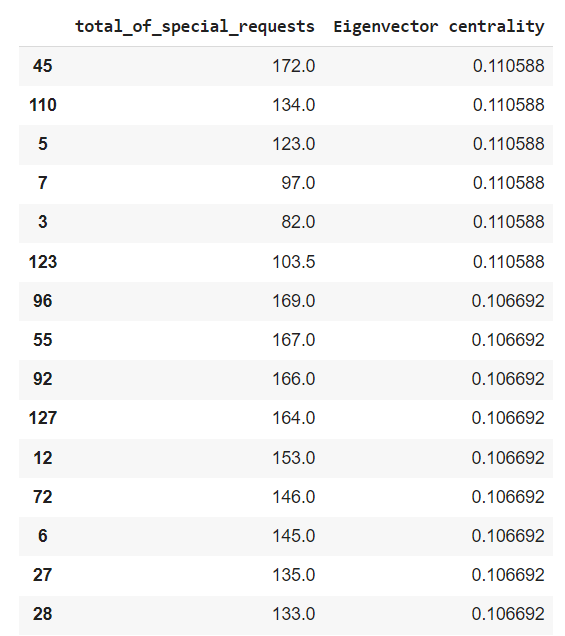
*Hình 41. Pagerank trên gephi*

## 4. Eigenvector và Eigenvalue

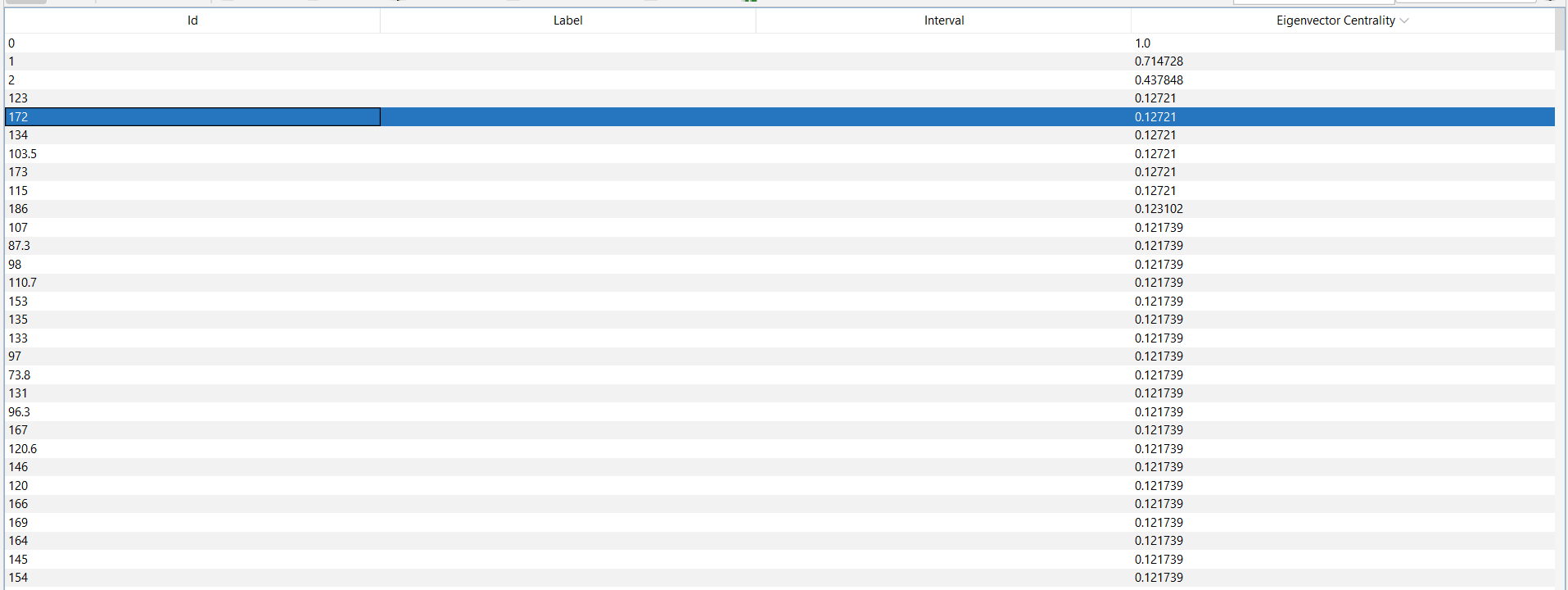
### 4.1. Eigenvector centrality



*Hình 42. Code python thuật toán Eigenvector*

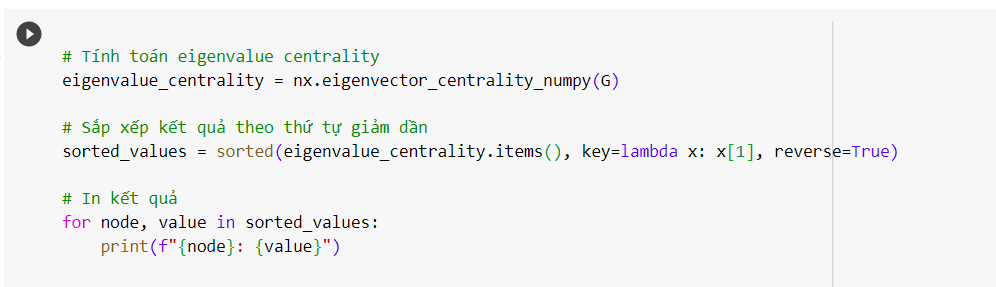


*Hình 43. Kết quả code python thuật toán Eigenvector*

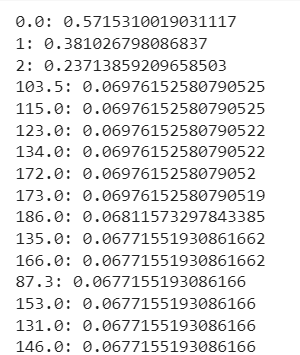


*Hình 44. Thuật toán Eigenvector trên gephi*

### 4.2. Eigenvalue



*Hình 45. Code python thuật toán Eigenvalue*



*Hình 46. Kết quả code python thuật toán Eigenvalue*

**TÀI LIỆU THAM KHẢO**

[1] [Link Code Python (Google Colab)](https://colab.research.google.com/drive/1Cko9nlZBEn6frq6fRlDEpuLm0EHgY8Yo?usp=sharing)

[2] [CodePython\_Sampleupdate.doc](https://uithcm-my.sharepoint.com/:w:/g/personal/20521645_ms_uit_edu_vn/EVm0s3WHnIhCgehw1H6iJAEBxxup3Xc_l8pHmqHGyg1POQ?e=PByBWD)

[3] [Link Dataset](https://uithcm-my.sharepoint.com/:x:/g/personal/20521645_ms_uit_edu_vn/EbCgbl3BPB5Eukt86Qph6LgBIa_Um_vs-rsMlQhF3R95ag?e=pKiuYk)

**Thông tin liên hệ**

Lê Hoàng Huy 0769699206

Huỳnh Thị Thanh Ngân 0865882023