TRƯỜNG ĐẠI HỌC SƯ PHẠM KĨ THUẬT THÀNH PHỐ HỒ CHÍ MINH

KHOA ĐÀO TẠO CHẤT LƯỢNG CAO

NGÀNH CÔNG NGHỆ THÔNG TIN

**⬩⬩🙠🕮🙢⬩⬩**



**BÁO CÁO ĐỒ ÁN 1**

**ĐỀ TÀI**

**TÌM HIỂU THUẬT TOÁN**

**AGGLOMERATIVE HIERARCHICAL CLUSTERING VÀ ỨNG DỤNG**

**GVHD: TS.Lê Văn Vinh**

**SVTH1 MSSV**

**Phạm Thị Như Hảo 16110068**

**Đặng Thanh Tú 16110253**

**Thành phố Hồ Chí Minh, tháng 12 năm 2018**

TRƯỜNG ĐẠI HỌC SƯ PHẠM KĨ THUẬT THÀNH PHỐ HỒ CHÍ MINH

KHOA ĐÀO TẠO CHẤT LƯỢNG CAO

NGÀNH CÔNG NGHỆ THÔNG TIN

**⬩⬩🙠🕮🙢⬩⬩**



**BÁO CÁO ĐỒ ÁN 1**

**ĐỀ TÀI**

**TÌM HIỂU THUẬT TOÁN**

**AGGLOMERATIVE HIERARCHICAL CLUSTERING VÀ ỨNG DỤNG**

**GVHD: TS.Lê Văn Vinh**

**SVTH1 MSSV**

**Phạm Thị Như Hảo 16110068**

**Đặng Thanh Tú 16110253**

**Thành phố Hồ Chí Minh, tháng 12 năm 2018**

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Giáo viên hướng dẫn

(ký và ghi họ tên)

NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Giáo viên phản biện

(ký và ghi họ tên)

LỜI CẢM ƠN

Lời đầu tiên, cho phép em được gửi lời cảm ơn chân thành đến thầy Lê Văn Vinh, người đã phụ trách hướng dẫn chúng em trong việc thực hiện và hoàn thành đề tài một cách tối ưu, hiệu quả nhất. Những sự chỉ dạy, nhật xét, đóng góp ý kiến từ thầy đã góp phần rất quan trọng trong việc giúp chúng em hoàn hành đề tài một cách hoàn thiện nhất.

Cảm ơn các thầy cô trong khoa công nghệ thông tin đã luôn nhiệt tình, cống hiến, tận tình giải đáp các thắc mắc của chúng em. Bên cạnh đó, chúng em xin được cảm ơn các bạn cùng khóa, các anh chị khóa trên đã không ngần ngại chia sẻ các thông tin, kinh nghiệm bổ ích giúp chúng em hoàn thiện hơn đề tài của mình.

Bài thu hoạch được thực hiện trong vòng 8 tuần, một khoảng thời gian có hạn, cùng với những hạn chế về mặt kiến thức nên việc sai sót là điều không thể tránh khỏi trong quá trình thực hiện. Chúng em kính mong nhận được những lời nhận xét, ý kiến đóng góp quý báo từ quý Thầy Cô để chúng em có thêm kinh nghiệm và hoàn thành tốt hơn những đề tài sau này.

Chúng em xin chân thành cảm ơn.

Thành phố Hồ Chí Minh, ngày 16 tháng 12 năm 2018

Sinh viên thực hiện 1 Sinh viên thực hiện 2

(kí và ghi rõ họ tên) (kí và ghi rõ họ tên)

TÓM TẮT

Trong những năm trở lại đây, công nghệ thông tin đã và đang không ngừng phát triển một các mạnh mẽ. Sự ra đời của công nghệ thông tin đã một phần nào giúp cho xã hội, đời sống của người dân vươn lên một tầm cao mới và một cách rõ rệt, ngoài ra công nghệ thông tin còn góp phần không nhỏ cho sự phát triển của nhân loại khắp nơi trên thế giới.

Trong thế kỷ 21 như hiện nay, trí tuệ nhân tạo là một lĩnh vực rất nóng và rất có tiểm năng phát triển do việc tính toàn của máy lúc nào cũng nhanh và chính xác hơn con người. Tại sự kiện Google I/O 2018 v diễn ra, “gã khổng lồ tìm kiếm” đã có màn trình diễn AI đầy ấn tượng trước hàng triệu người có mặt tại sân khấu cũng như qua các phương tiền truyền thông. Những người có mặt trức tiếp đã phải “Ồ” lên đầy ngạc nhiên, thích thú khi trợ lí ảo Google Assistant thực hiện cuộc gọi đặt lịch cắt tóc mà phái người nghe không hề biết đang nói chuyện với một AI. Sự thông minh và linh hoạt của AI được kỳ vọng sẽ giúp ích cho con người rất nhiều trong tương lai, giúp tăng năng suất và hiệu quả lao động. Do đó chúng em đã chọn đề tài **“Tìm hiểu về thuật toàn Agglomerative Hierarchical Clustering và ứng dụng”** để một phần biết được một phần kiến thức về trí tuệ nhân tạo, cách thức hoạt động và những ứng dụng của chúng trong cuộc sống.

MỤC LỤC

[1.Mô tả project 1](#_Toc532750009)

[2.Quá trình thực hiện project 1](#_Toc532750010)

[2.1. Tìm hiểu sơ lược về thuật toán Agglomerative Hierarchical Clustering (HAC) 1](#_Toc532750011)

[2.1.1. Đôi nét về Hierarchical Clustering 1](#_Toc532750012)

[2.1.2. Thuật toán Agglomerative Hierarchical Clustering 2](#_Toc532750013)

[2.2. Mẫu thiết kế code 3](#_Toc532750014)

[2.2.1. Liệt kê các lớp sử dụng trong chương trình (.java) 3](#_Toc532750015)

[2.2.2. Liệt kê các phương thức sử dụng trong chương trình (.java) 3](#_Toc532750016)

[2.2.3. Kết quả sau khi Run file Main.java 4](#_Toc532750017)

[2.3. Demo bằng đồ họa bằng ngôn ngữ python 6](#_Toc532750018)

[2.4. Ứng dụng của thuật toán AHC 8](#_Toc532750019)

[3.Mô tả phân công công việc 9](#_Toc532750020)

[4.Kết luận 9](#_Toc532750021)

[5.Tài liệu tham khảo 9](#_Toc532750022)

DANH MỤC CÁC HÌNH

[Hình 2. 1 Sơ đồ khối thuật toán AHC 2](#_Toc532752153)

[Hình 2. 2 Khởi tạo các đối tượng là các Point 4](#_Toc532752154)

[*Hình 2. 3 Khởi tạo và tính khoảng cách giữa các Cluster* 5](#_Toc532752155)

[*Hình 2. 4 Tìm ra khoảng cách giữa hai Cluster và tiến hành gộp 2 Cluster* 5](#_Toc532752156)

[*Hình 2. 5 Lặp lại các bước cho đến khi gộp hai Cluster cuối cùng lại thành một (Cluster CJAEHI và Cluster DFBG)* 6](#_Toc532752157)

[*Hình 2. 6 Import thư viện* 6](#_Toc532752158)

[*Hình 2. 7 Code vẽ tọa độ* 7](#_Toc532752159)

[*Hình 2. 8 Tọa độ các điểm* 7](#_Toc532752160)

[*Hình 2. 9 Code demo gom cụm* 7](#_Toc532752161)

[*Hình 2. 10 Quá trình gom cụm* 8](#_Toc532752162)

DANH MỤC CÁC BẢNG

[*Bảng 1 Liệt kê các Class trong project (Main.java)* 3](#_Toc532752489)

[*Bảng 2 Các phương thức sử dụng trong chương trình (Main.java)* 3](#_Toc532752490)

[*Bảng 3 Demo thuật toác HAC bằng ngôn ngữ python trên jupyter notebook* 6](#_Toc532752491)

[*Bảng 4 Bảng phân công công việc* 9](#_Toc532752492)

# 1.Mô tả project

Sản phẩm tạo ra nhằm minh họa một cách trực quan thuật toán Agglomerative Hierarchical Clustering (gọi tắt là HAC: thuật toán dùng để gom cụm dữ liệu). Project gồm 2 chương trình: một chương trình Java Application (lưu dưới dạng file .java) dùng để tạo nên các class mô tả thuật toán, một chương trình python (lưu dưới dạng file .py) sử dụng các thư viện có sẵn để mô tả trực quan bằng việc vẽ nên các quá trình thực thi thuật toán HAC. Chương trình Java Application sẽ mô tả việc khởi tạo các class, các phương thức, các hàm xử lí và gom cụm dữ liệu. Chương trình python sẽ mô tả vị trí tọa độ các điểm trên hệ trục tọa độ và vẽ nên sơ đồ cách thức gom cụm các điểm tạo thành các Cluster.

# 2.Quá trình thực hiện project

## 2.1. Tìm hiểu sơ lược về thuật toán Agglomerative Hierarchical Clustering (HAC)

### 2.1.1. Đôi nét về Hierarchical Clustering

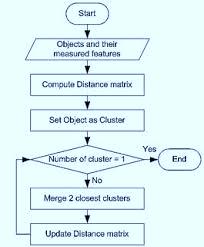
Hierarchical Clustering (phân cụm phân cấp): là một phương pháp phân tích cụm nhằm để xây dựng một hệ thống phân cấp của các cụm. Thuật toán phân cụm phân cấp được chia làm hai loại:

-Agglomerative : thực thi theo phương pháp bottom-up (từ dưới lên) mỗi điểm ban đầu được xem như một Cluster (cụm), sau đó sẽ tiến hành gom cụm 2 Cluster lại với nhau cho đến khi chỉ còn một Cluster bao gồm tất cả các điểm khởi tạo ban đầu.

-Divisive: thực thi theo phương pháp top-down (từ trên xuống) tất cả các điểm khởi tạo ban đầu được lưu trong một Cluster, sau đó sẽ tiến hành phân chia Cluster thành các điểm nhỏ (trái ngược với Agglomerative).

### 2.1.2. Thuật toán Agglomerative Hierarchical Clustering

#### 2.1.2a. Sơ đồ khối



Hình 2. 1 Sơ đồ khối thuật toán AHC

#### 2.1.2b. Thuật toán

Thuật toán HAC bao gồm bước:

-Bước 1: Khởi tạo các đối tượng gồm có 2 thuộc tính là X và Y

-Bước 2: Tính khoảng cách giữa các đối tượng

-Bước 3: Tạo ma trận khoảng cách

-Bước 4: Tìm hai điểm có khoảng cách nhỏ nhất gộp thành một Cluster

-Bước 5: Cập nhật lại số Cluster và lặp lại từ bước 2 tới bước 4 cho đến khi chỉ còn một cluster bao gồm tất cả các đối tượng đã khởi tạo.

## 2.2. Mẫu thiết kế code

### 2.2.1. Liệt kê các lớp sử dụng trong chương trình (.java)

*Bảng 1 Liệt kê các Class trong project (Main.java)*

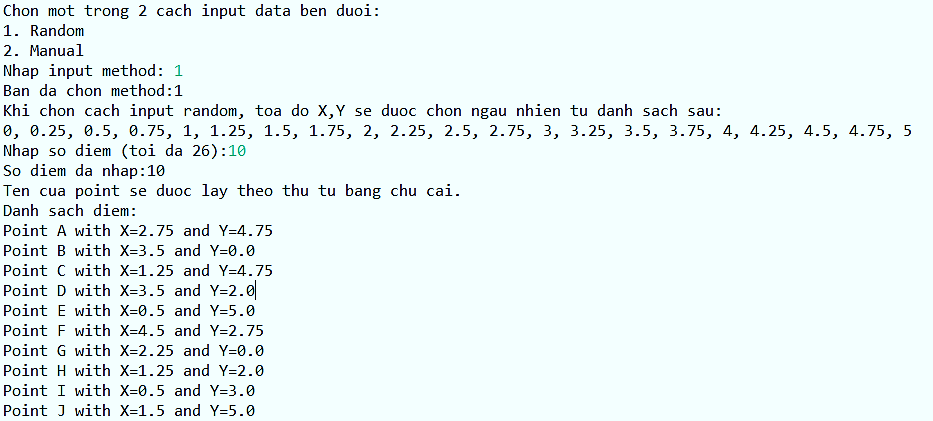
|  |  |  |  |
| --- | --- | --- | --- |
| TT | Class Name | SV phụ trách | Mục đích của Class |
| 1 | Point | Đặng Thanh Tú | Khởi tạo các đối tượng gọi là điểm với hai thuộc tính là x và y |
| 2 | Cluster | Phạm Thị Như Hảo | Khởi tạo các Cluster với các thuộc tính là name và danh sách các điểm (Point[]) |
| 3 | KhoangCach | Phạm Thị Như Hảo | Khởi tạo biến kc để lưu giá trị khoảng cách và lưu giá trị ấy vào mảng KhoangCach[] |

### 2.2.2. Liệt kê các phương thức sử dụng trong chương trình (.java)

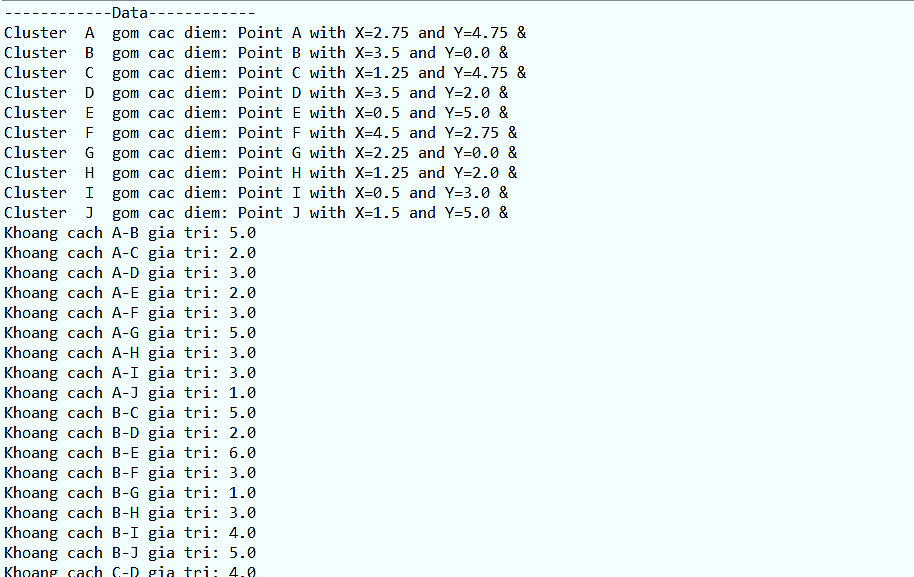
*Bảng 2 Các phương thức sử dụng trong chương trình (Main.java)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TT | Phương thức | Mục đích | Tên file, số thứ tự dòng chứa khai báo | Tên SV phụ trách viết |
| 1 | tinhkhoangcach(Point a, Point b) | Tính khoảng cách giữa hai đối tượng Point | Main.java  (368) | Đặng Thanh Tú |
| 2 | tinhkhoangcachcluser  (KhoangCach[] matrankc, Cluster a, Cluster b) | Tính khoảng cách giữa hai Cluster | Main.java  (396) | Đặng Thanh Tú |
| 3 | gopcluster(Cluster c1, Cluster c2) | Gộp 2 Cluster lại thành 1 | Main.java  (177) | Phạm Thị Như Hảo |
| 4 | methodSelection(Scanner scan) | Lựa chọn phương thức khởi tạo các dữ liệu | Main.java  (231) | Phạm Thị Như Hảo |
| 5 | removekhoangcachbyclustername(KhoangCach[] khoangCachCluster, String clusterName) | Xóa các giá trị khoảng cách trong ma trận khoảng cách dựa vào tên cluster | Main.java  (137) | Phạm Thị Như Hảo |
| 6 | removeclusterbyname(Cluster[] clusterList, String name) | Xóa cluster dựa vào tên cluster | Main.java  (192) | Phạm Thị Như Hảo |
| 7 | layclusternyname(Cluster[] clusterList, String name) | Trả về một Cluster dựa vào tên của Cluser | Main.java  (214) | Phạm Thị Như Hảo |
| 8 | inputRandom(Scanner scan, double[] toado) | Tạo ngẫu nhiên danh sách các điểm | Main.java  (249) | Phạm Thị Như Hảo |

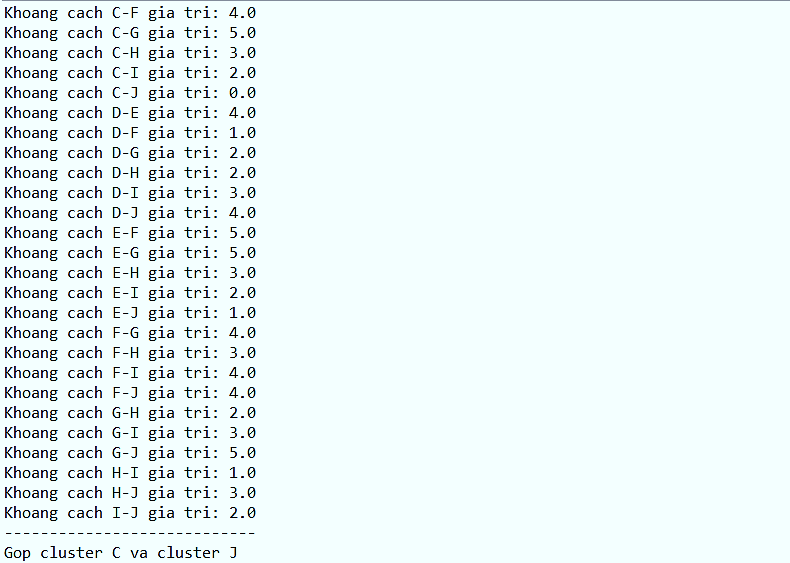
### 2.2.3. Kết quả sau khi Run file Main.java



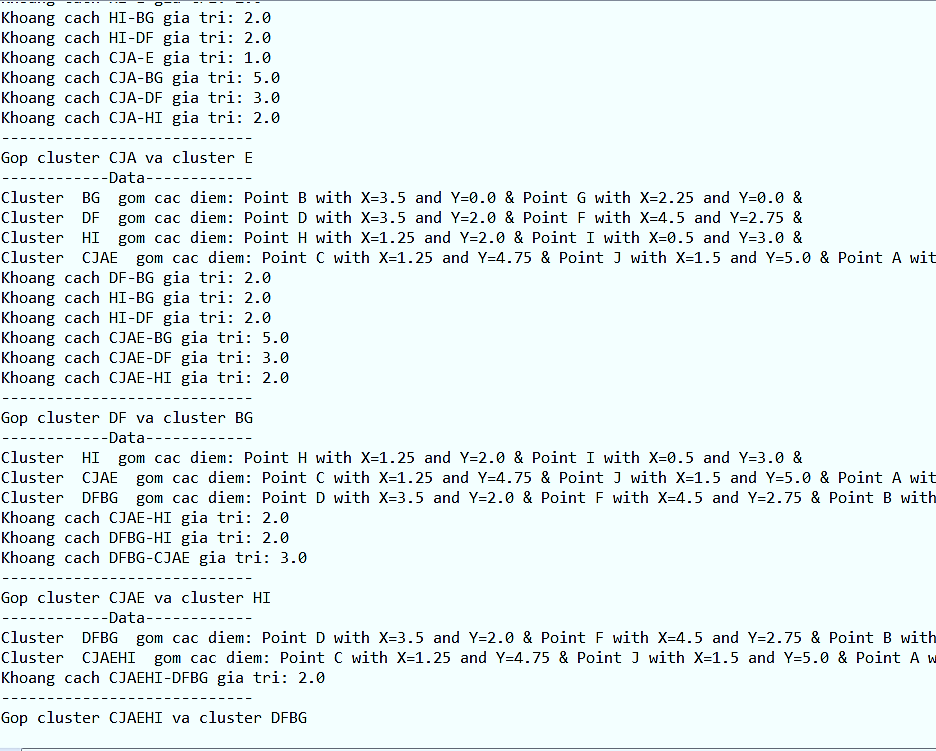
Hình 2. 2 Khởi tạo các đối tượng là các Point



*Hình 2. 3 Khởi tạo và tính khoảng cách giữa các Cluster*



*Hình 2. 4 Tìm ra khoảng cách giữa hai Cluster và tiến hành gộp 2 Cluster*



*Hình 2. 5 Lặp lại các bước cho đến khi gộp hai Cluster cuối cùng lại thành một (Cluster CJAEHI và Cluster DFBG)*

## 2.3. Demo bằng đồ họa bằng ngôn ngữ python

*Bảng 3 Demo thuật toác HAC bằng ngôn ngữ python trên jupyter notebook*

|  |  |  |  |
| --- | --- | --- | --- |
| TT | Màn hình | Giải thích | Mục đích chính |
| 1 | *Hình 2. 6 Import thư viện* | Imort các thư viện cần thiết và khởi tạo một mảng kiểu numpy | Nhằm sử dụng dễ dàng, thuận tiện cho việc demo |
| 2 | *Hình 2. 7 Code vẽ tọa độ* | Sử dụng các hàm chức năng có sẵn trong thư viện matplotlib.pylot | Nhằm biểu trị bị trí của các điểm trong hệ trục tọa độ |
| 3 | *Hình 2. 8 Tọa độ các điểm* | Kết quả sau khi run đoạn code số 2 | Vị trí các điểm được biểu diễn trên hệ trục tọa độ |
| 4 | *Hình 2. 9 Code demo gom cụm* | Sử dụng các module như linkage, dendrogram trong thư viện scipy.cluster.hierarchy | Dùng để biểu thị quá trình gom cụm các Cluster |
| 5 | *Hình 2. 10 Quá trình gom cụm* | Kết quả sau khi thực thi đoạn code số 4 | Kết quả và chức năng của hàm dendrogram |

## 2.4. Ứng dụng của thuật toán AHC

Thuật toán AHC có thể được dùng trong việc phân tích phát sinh gen trên vi khuẩn nước bọt. Phân tích quần thể vi khuẩn Streptococcus rất quan trọng vì những loài này có thể gây sâu răng, nha chu, hôi miệng (hôi miệng) và nhiều vấn răng miệng khác. Người ta nghiên cứu về mối quan hệ phát sinh gen giữa vi khuẩn Streptococcus trong nước bọt bằng cách sử dụng một cây phát sinh phương pháp phân cụm. Bắt đầu với trình tự DNA Streptococcus của vi khuẩn thu được từ GenBank, sau đó thực hiện trích xuất trình tự DNA đặc trưng. Kết quả trích xuất đặc trưng là dạng ma trận, sau đó thực hiện chuẩn hóa bằng cách chuẩn hóa tối thiểu và tính khoảng cách di truyền bằng khoảng cách Manhattan.

Kỹ thuật phân cụm kết tụ bao gồm liên kết đơn (single linkage), liên kết hoàn chỉnh (complet linkage) và liên kết trung bình (average linkage). Trong thuật toán kết tụ này, số lượng nhóm được bắt đầu bằng số lượng các loài riêng lẻ. Các loài tương tự nhất được nhóm lại cho đến khi độ tương tự giảm và sau đó tạo thành một nhóm duy nhất.

Kết quả của việc phân nhóm là một cây phát sinh và các nhánh tham gia vào một khoảng cách xác định, khoảng cách càng nhỏ thì sự giống nhau của việc thực hiện các loài càng lớn.

# 3.Mô tả phân công công việc

*Bảng 4 Bảng phân công công việc*

|  |  |  |
| --- | --- | --- |
| Tên SV | Đánh giá chung phần trăm đóng góp | Mô tả khái quát mảng công việc SV thực hiện trong đồ án. |
| Phạm Thị Như Hảo | 60% | Xây dựng ý tưởng, tìm kiếm tài liệu, viết code, xây dựng báo cáo |
| Đặng Thanh Tú | 40% | Tìm kiếm tài liệu, viết báo cáo, code một số hàm trong file Main.java |

# 4.Kết luận

Đồ án đã được hoàn thành 85% mục tiêu đề ra.

Khó khăn: Do đề tài còn mới mẻ nên việc tìm hiểu gặp nhiều khó khăn, đòi hỏi kĩ năng tìm đọc tài liệu tiếng anh cao

Ưu điểm: Hiểu được cơ chế làm việc của thuật toán AHC, demo được thuật toán bằng hai phương pháp là Java Application và sử dụng đồ họa python

Khuyết điểm: Chưa làm được các bài toán lớn trong thực tế vì quá phức tạp

# 5.Tài liệu tham khảo

Từ nguồn tài liệu giáo viên hướng dẫn cung cấp:

<https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>

Từ Internet:

<http://bis.net.vn/forums/p/571/1485.aspx?fbclid=IwAR1pzRoIZvkg01tGljPiqTs9kjLEkt1SxSvNJjR4OxTSg3rS2sGIKCftX_g>

<https://en.wikipedia.org/wiki/Hierarchical_clustering>

<https://aip.scitation.org/doi/abs/10.1063/1.4991230>

https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/