TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CNTT&TT

BÁO CÁO
KHAI KHOÁNG DỮ LIỆU

Đề tài

XÂY DỰNG WEBSITE GOM NHÓM DỮ LIỆU

Giảng viên hướng dẫn:

GV: Lưu Tiến Đạo

Sinh viên thực hiện:

Đào Nguyễn Duy Khanh

MSSV: B1609773

Nguyễn Hoàng Châu

MSSV: B1609762

Học kì 1,2019-2020

LỜI CẢM ƠN

Lời đầu tiên chúng em xin được gửi tới Thầy Lưu Tiến Đạo đã hướng dẫn, giúp đỡ và cho những lời khuyên quý báo để em có thể hoàn thành đề tài niên luận cơ sở ngành Khoa học máy tính lần này. Trong quá trình thực hiện em đã gặp không ít khó khăn và những sai sót trong suốt quá trình làm đề tài, nếu không có sự hướng dẫn và hỗ trợ của cô, có lẽ em đã khó lòng hoàn thành được và có thể cho ra được thành quả như hôm nay. Xin chân thành gửi lời cảm ơn chân thành nhất và sâu sắc nhất đến cô đã hướng dẫn em trong suốt quá trình.

Xin cảm ơn các thầy cô trường Đại Học Cần Thơ và đặc biệt là các thầy cô Khoa Công Nghệ Thông Tin và Truyền Thông. Cám ơn các thầy cô đã luôn lo lắng, truyền đạt những kiến thức quý giá giúp cho chúng em có đủ năng lực chuyên môn và kinh nghiệm để có thể hoàn thành tốt đề tài.

Mặc dù đã cố gắng hoàn thành đề tài niên luận cơ sở ngành Khoa học máy tính trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Em rất mong nhận được sự cảm thông và góp ý tận tình từ quý thầy cô và bạn bè. Qua đó, em có thể tích luỹ những kinh nghiệm cho bài luận văn quan trọng sắp tới.

Cần Thơ, ngày 5 tháng 06 năm 2020

Người viết

NHẬN XÉT CỦA GIẢNG VIÊN

MỤC LỤC

DANH MỤC HÌNH ẢNH

PHẦN GIỚI THIỆU

1.Đặt vấn đề

Ngày nay, trong thời kì Công nghiệp hóa-Hiện đại hóa. Công nghệ thông tin đã trở nên phổ biến rộng và có tầm quan trọng trên hầu hết mọi lĩnh vực. Các công việc cũng được tin học hóa một cách rộng rãi. Dữ liệu cũng từ đó mà được sinh ra ngày một nhiều. Để có thể xử lý dữ liệu đó đòi hỏi người dùng phải có tư duy về lập trình cũng như hiểu biết về công nghệ thông tin. Nhưng không phải ai cũng là dân trong nghề. Không phải ai cũng có thể ngồi lập trình từng dòng code. Để khắc phục hạn chế này. Trước đó đã từng có xuất hiện một số website được viết bằng ngôn ngữ R khá là phức tạp, giao diện cũng khó cho người sử dụng nhưng chi phí để vận hành hệ thống khá là cao. Trước các hạn chế đó thì việc xây dựng một trang web dựa trên ngôn ngữ mới Python với giao diện phù hợp với mọi đối tượng, được sử dụng miễn phí là khá cần thiết.

2.Tóm tắt lịch sử giải quyết vấn đề

Đã có các trang web được xây dựng để xử lý dữ liệu như:

www.asaanai.com hệ thống xử lý dữ liệu bằng các mô hình mà không cần tới một dòng code.

3.Mục tiêu đề tài

Đề tài tập trung vào việc xây dựng và thiết kế trang web dành cho người dùng để xử lý dữ liệu. Áp dụng hai giải thuật K-means và agg. Tạo ra chức năng gom nhóm dữ liệu mà không cần đến một dòng code.

4.Đối tượng phạm vi nghiên cứu

Đối tượng và phạm vi nghiên cứu là những người dùng trong mọi lĩnh vực cần xử lý dữ liệu.

5.Nội dung nghiên cứu

- Xây dựng bố cục trang web
- Thiết kế trang web

- Cài đặt giải thuật K-means và Agg
- Tích hợp giải thuật vào trang web
- Tạo chức năng trang web
- Chạy trên trình duyệt

6.Bố cục

Bố cục bài báo cáo gồm 3 phần:

Phần giới thiệu: Đặt vấn đề và đưa ra hướng giải quyết dựa vào yêu cầu thực tế. Xác định cụ thể đối tượng và phạm vi nghiên cứu. Xây dựng nội dung cụ thể của việc nghiên cứu. Tóm tắt bố cục niên luận để người đọc dễ theo dõi.

Phần nội dung:

Chương 1: Tìm hiểu về Website. Các khái niệm cũng như phân loại các website. Mô tả bài toán dựa trên yêu cầu thực tế. Đưa ra các chức năng chính của trang web, môi trường vận hành và các tính năng của hệ thống.

Chương 2: Tìm hiểu về giải thuật K-means và Agg. Cài đặt trên ngôn ngữ python. Tích hợp vào website để tạo chức năng gom nhóm dữ liệu cho người dùng.

Chương 3: Kiểm thử và đánh giá bao gồm kế hoạch kiểm thử, các trường hợp kiểm thử, kết quả kiểm thử và cách giải quyết nếu có sai sót

Phần kết luận: Những kết quả đạt được và hướng phát triển của trang web.

CHƯƠNG 1: MÔ TẢ BÀI TOÁN

1. Khái niệm website:

Website còn gọi là trang web hoặc trang mạng, là một tập hợp các trang web, thường chỉ nằm trong một tên miền hoặc tên miền phụ trên World Wide Web của Internet. Một trang web là tập tin HTML hoặc XHTML có thể truy nhập dùng giao thức HTTP. Trang mạng có thể xây dựng từ các tệp tin HTML(trang mạng tĩnh) hoặc vận hành bằng các CMS chạy trên máy chủ(trang mạng động).

Trang mạng có thể được xây dựng bằng nhiều ngôn ngữ lập trình khác nhau(PHP, ASP.NET, Java, Puby on Rails, Pert,…).

2. Phân loại Website:

- **Phân loại theo dữ liệu:**

-Website động (Dynamic website) là website có cơ sở dữ liệu, được cung cấp công cụ quả lý website (Admin Tool) để có thể cập nhật thông tin thường xuyên, quản lý các thành phần trên website. Loại website này thường được viết bằng các ngôn ngữ lập trình như PHP, Asp.net, JSP, Perl,..., quản trị Cơ sở dữ liệu bằng SQL hoặc MySQL,...

Với website động khi xây dựng sẽ bao gồm 2 phần.

+ Một phần hiển thị trên trình duyệt mà khi truy cập internet ta thường thấy

+ Phần ngầm bên dưới dùng để điều khiển nội dung của trang web, phần nội dung ngầm này thường thì chỉ những người quản trị website đó mới có thể có quyền truy cập vào.

-Website tĩnh (Static Website): Web tĩnh ở đây được hiểu theo nghĩ là dữ liệu không thay đổi thường xuyên, do lập trình bằng ngôn ngữ HTML theo từng trang như brochure, không có cơ sở dữ liệu và không có công cụ quản lý thông tin trên website. Bạn phải biết kỹ thuật thiết kế trang web (thông thường bằng các

phần mềm như FrontPage, Dreamwaver,...) khi muốn thiết kế hoặc cập nhật thông tin của những trang web này.

Với dạng web này người để thay đổi nội dung trên trang web người sở hữu phải truy cập trực tiếp vào các mã lệnh để thay đổi thông tin. Không có cơ sở dữ liệu bên dưới hệ thống, không có công cụ để điều khiển nội dung gián tiếp. Dạng file của trang website tĩnh thường là html,htm,..

- **Phân loại theo đối tượng sở hữu**

Có thể là công việc của một cá nhân, một doanh nghiệp hoặc các tổ chức, và thường dành riêng cho một số chủ đề cụ thể hoặc mục đích. Bất kỳ trang web có thể chứa một siêu liên kết vào bất kỳ trang web khác, do đó phân biệt các trang web cá nhân, như cảm nhận của người sử dụng. Tạm thời phân loại như sau:

+ Thiết kế website cá nhân: Các đối tượng như diễn viên, ca sĩ, người nổi tiếng, người thiết kế đồ hoạ, hoặc bất kỳ cá nhân nào thích giớ thiệu bản thân mình đều có thể tạo ra một website cho cá nhân mình

+Thiết kế website tin tức**:** Đây là một dạng website cung cấp thông tin chính trị, xã hội, kinh tế, khoa học, giáo dục, sức khoẻ,... thể loại này được phát triển trên nền tảng từ các thể loại báo giấy truyền thống.

+ Thiết kế website Mạng xã hội (blog): Là dạng web dành cho người sử dụng được quyền tạo cho mình một không gian riêng gồm nhiều trang độc lập, ở đây người dùng có thể đăng tải thông tin cá nhân,sở thích,viết nhật ký tại trang của mình. Các mạng xã hội nổi tiếng như: Yahoo, Facebook,Workpress, opera,...

+ Thiết kế website cho doanh nghiệp: thiết kế web với mục đích quảng bá công ty, giới thiệu các chức năng hoạt động, cập nhật những tin tức, sản phẩm mới của công ty nhằm dễ dàng tiếp cận đến khách hàng thông qua một kênh quảng bá mới là internet.

+ Thiết kế website Diễn đàn: Là dạng web tương tác với người dùng mà bất kỳ xem nào cũng có thể đăng ký tham gia là thành viên và được quyền tăng tải bài viết của mình và dĩ nhiên diễn đàn luôn có người kiểm soát thông tin người dùng đăng tải và có quyền can thiệp vào việc hiển thị thông tin đó hay không

+ Thiết kế website bán hàng: là các dạng web cho phép bán hàng trực tuyến, việc thanh toán có nhiều hình thức như: tiền mặt, chuyển khoản, thanh toán bằng thẻ, hoặc thông qua cổng thanh toán của các dịch vụ hỗ trợ.

+ Thiết kế website dành cho các tổ chức, cơ quan nhà nước: Các bộ, sở, ban ngành, hiệp hội tổ chức,…là đối tượng sở hữu website dạng này.

+ Thiết kế website giải trí: đăng tải phim ảnh, nhạc, game,..

3.Mô tả tổng quan

3.1. Bối cảnh sản phẩm

Hiện nay, các dữ liệu cá nhân cũng như dữ liệu công ty doanh nghiệp đều được lưu trữ xử lý trên trên không gian mạng. Các dữ liệu này rất đa dạng và phong phú. Đòi hỏi người dùng phải có tư duy cũng như các kỹ năng chuyên ngành để có thể phân tích và xử lý chúng. Việc đó đem đến nhiều khó khăn hạn chế rất nhiều cho người dùng. Để xử lý vấn đề này việc xây dựng một trang web xử lý dữ liệu là một việc rất cần thiết. Nó đem lại lợi ích về mặt thời gian công sức cũng như giảm bớt được chi phí và tăng nhiều lợi nhuận cho công ty. Trang web được xây dựng với ngôn ngữ mới giúp cho người dùng có thể dễ dàng tiếp cận các chức năng mà không cần kiến thức về chuyên môn và hoàn toàn miễn phí.

3.2. Các chức năng của trang web

- Chức năng nhập dữ liệu: Cho phép người dùng nhập các file dữ liệu như: csv, excel,…

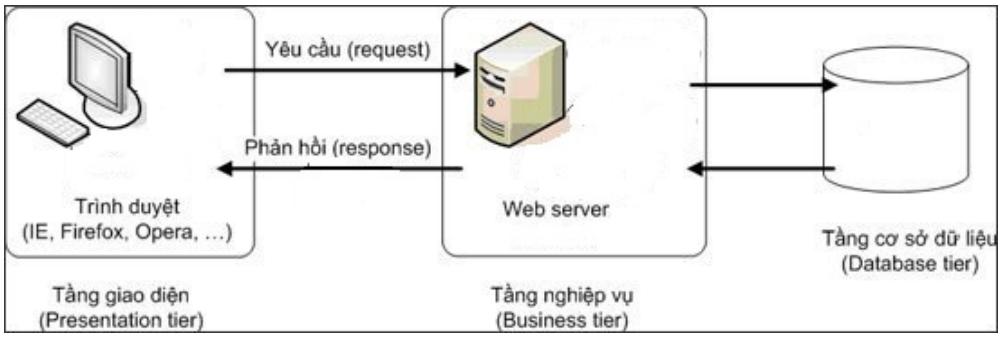
- Chức năng chọn cột: Hiển thị số cột trong tập dữ liệu của người dùng. Cho phép người dùng đánh dấu vào số cột của dữ liệu cần thiết để gom nhóm.
- Chức năng xuất đồ thị: Chức năng này xuất ra đồ thị dựa vào số cột mà người dùng chọn để gợi ý số nhóm phù hợp.
- Chức năng chọn nhóm: Chức năng này giúp người dùng có thể lựa chọn số nhóm phù hợp để gom nhóm cho dữ liệu của mình.
- Chức năng hiển thị hình ảnh: Chức năng này hiện thị hình ảnh của dữ liệu khi đã gom nhóm. Giúp người dùng có thể dễ dàng nhận biết dữ liệu của mình

3.3. Môi trường vận hành

Trang web được vận hành trên host. Với domain được mua trên các máy chủ đặt ở Việt Nam hoặc nước ngoài.

CHƯƠNG 2: THIẾT KẾ VÀ CÀI ĐẶT

1.Kiến trúc hệ thống



Hình 1: Sơ đồ kiến trúc hệ thống

Các thành phần của hệ thống:

-*Trình duyệt:* Đảm nhận vai trò của client. Nhận thông tin từ Webserver rồi hiển thị cho người dùng. Nhận dữ liệu từ người dùng lựa chọn, nhập vào rồi gửi về cho Webserver.

-*Web server:* Lưu trữ mã nguồn và vận hành hệ thống. Tiếp nhận các yêu cầu từ phía client, xử lý, thao tác với cơ sở dữ liệu rồi lưu trữ, trả kết quả về client.

-*Clustering:* Là giải thuật gom nhóm được tích hợp vào trang web. Sau khi tiếp nhận được dữ liệu từ phía webserver thì Clustering sẽ xử lý dữ liệu và trả về phía webserver để hiển thị cho người dùng.

2. Thiết kế trang web

2.1. HTML

HTML (tiếng anh, viết tắt cho từ HyperText Markup Language, hay là "ngôn ngữ đánh dấu siêu văn bản") là một ngôn ngữ đánh dấu được thiết kế ra để tạo nền các trang web với các mẫu thông tin được trình bày trên World Wide Web.

Các thẻ(tag) được sử dụng:

- **<link>:** Thẻ link sử dụng để chèn các đường link của các file Css, Javascrpipt.
- **<nav>:** Tạo thanh công cụ taskbar để người dùng có thể tùy chọn các menu nhanh chóng
- **
:** Các thẻ dùng để tạo các menu và menu con
- **<a>:** Dùng để chèn các link của các menu trên thanh taskbar
- ** :** Thẻ dùng để tạo khung bố cục trang web
- **<img>:** Thẻ dùng để chèn hình ảnh vào trang Web
- **<h1>:** Tạo tiêu đề
- **<p>:** Chèn các ký tự chuỗi ký tự
- **<hr>:** Vẽ các đường ngang để phân bố cục
- **<script>:** Chèn các đoạn javascipt

2.2. CSS

CSS là viết tắt của Cascading Style Sheets. Đây là một style sheet được sử dụng để mô tả giao diện và định dạng của một tài liệu viết bằng ngôn ngữ đánh dấu(markup).

Các class được sử dụng và các thuộc tính của class:

- Body: Class của body:
  - background-image: Chèn ảnh làm ảnh nền
  - background-repeat: không cho lặp hình ảnh
  - background-size: đặt size cho nền
  - background-position: Vị trị của nền
  - height: chiều dài của hình nền
- Các class: howdo, start, upload. Dùng để căn chỉnh và định dạng các ký tự và hình ảnh.

- padding (left, right, top, bottom): dùng để căn chỉnh lề trên dưới trái phải của ký tự và hình ảnh

- text-aligh: Dùng để căn chỉnh ký tự

-font-size: Kích cỡ ký tự

-font-family: Kiểu chữ

-font-weight: Phong cách hiển thị

2.3. Bootstrap

Bootstrap là nền tảng Framwork HTML, CSS hay JavaScript cho phép lập trình viên có thể thiết kế đượ website dựa trên reponsive web mobile.

- Jumbotron: Class của tiêu đề của trang web-> Sử dụng thư viện có sẵn của Bootstrap -> Tạo hiệu ứng cho chữ
- Container: Class dùng để phân chia bố cục khung hiển thị của trang web -> Sử dụng thư viện của Bootstrap ->Khung web được căn lề
- Row-Col: Class dùng để phân chia hàng và cột của một khung trong thẻ -> Sử dụng thư viện Bootsrap -> Các phần được phân chia rõ ràng theo từng phần được cài đặt sẵn
- Card: Class dùng để tạo một khung hiển thị chi tiết nội dung-> Sử dụng thư viện Bootsrap -> Tạo một khung riêng biệt để làm nổi bật nội dung.

3.Cài đặt giải thuật

3.1. Giải thuật gom cụm Clustering

- **Tổng quát:**

- Là mô hình gom cụm dữ liệu (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất tượng tự nhau và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau.

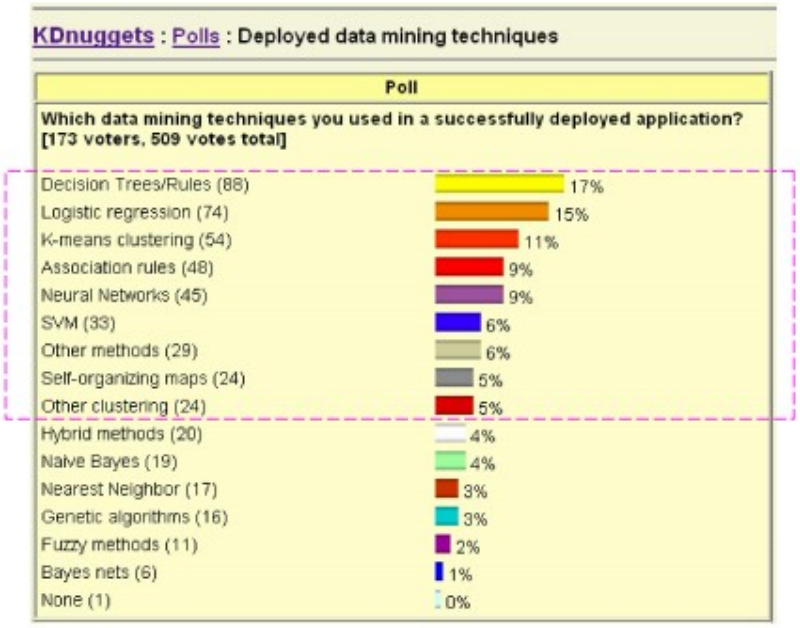
- Có nhiều nhóm giải thuật khác nhau: hierarchical clustering, partitioning, density-based, model-based, ect.

- Được sử dụng nhiều nhất: K-Means, Đẻnogram, SOM, EM

- Thường được tính dựa trên cơ sở khoảng cách nên phải chuẩn hóa dữ liệu

- Khoảng cách được tính theo từng kiểu của dữ liệu: số, nhị phân, symbol (interval, histogram, taxonomy)

- Được ứng dụng thành công trong hầu hết các lĩnh vực tìm kiếm thông tin, phân tích dữ liệu.



Biểu đồ tỉ lệ thành công của một số kỹ thuật khai thác dữ liệu dựa trên các bình chọn (173 người bình chọn và 509 tổng phiếu bình chọn)

- Cây quyết đinh/ Luật: 88 phiếu chiếm 17%

- Hồi quy tuyến tính: 74 phiếu chiếm 15%

- K-Means clustering: 54 phiếu chiếm 11%

- Association Rules: 48 phiếu chiếm 9%

- Neural Network: 45 phiếu chiếm 9%

- SVM: 38 phiếu chiếm 6%

- Các phương pháp khác: 29 phiếu chiếm 6%

- Self-organizing maps: 24 phiếu chiếm 5%

- Các phương pháp gom cụm khác: 24 phiếu chiếm 5%

- **Kiểu số:**

  - Khoảng cách Minkowski

$$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+...+|x_{i_p}-x_{j_p}|^q)}$$

  Trong đó i = (x_{i1}, x_{i2}, …, x_{ip}) và j = ($x_{j1,}$ x_{j2}, …, x_{jp}) là 2 phần tử dữ liệu trong *p*-dimensional, *q* là số nguyên dương.

  - Nếu *q* = 1, d là khoảng cách Manhattan

  - Nếu *q* = 2, d là khoảng cách Euclid

  - Khoảng cách cosine: $d_{cos}(i, j) = i^T j/(||i||\ ||j||)$.

- **Kiểu nhị phân**

		Object *j*		
		1	0	*sum*
Object *i*	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	sum	*a+c*	*b+d*	*p*

  - Khoảng cách đối xứng:

$$d(i,j)=\frac{b+c}{a+b+c+d}$$

  - Khoảng cách bất đối xứng:

$$d(i,j)=\frac{b+c}{a+b+c}$$

- Hệ số Jaccard bất đối xứng:

$$sim_{Jaccard}(i,j) = \frac{a}{a+b+c}$$

- **Kiểu loại (nominal type)**
  - Ví dụ các thuộc tính màu sắc
  - Phương pháp matching đơn giản, m là số lượng matches và p là tổng số biến (thuộc tính), khaongr cách được định nghĩa:

$$d(i,j) = \frac{p-m}{p}$$

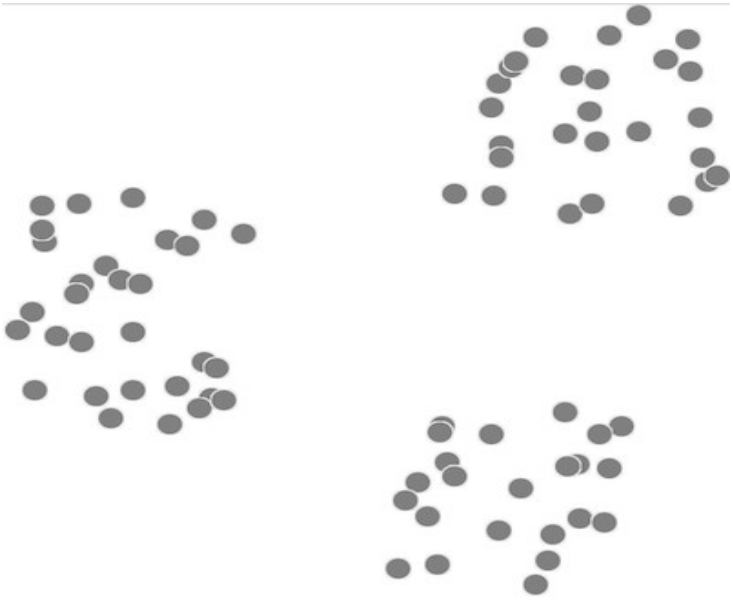
3.2. K-Means

- **Phương pháp K-Means là gì**

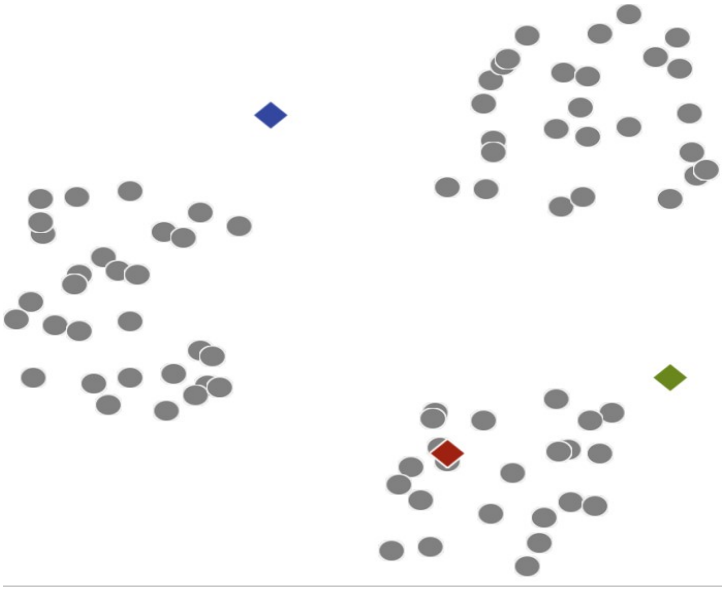
Phương pháp K-Means là một trong số những thuật toán gom cụm(clustering). Đầu vào tập dữ liệu cần phân cụm và số cụm(cluster), đầu ra chúng ta sẽ được kết quả dữ liệu đã được phân về các cluster.

- **Phân tích chi tiết**

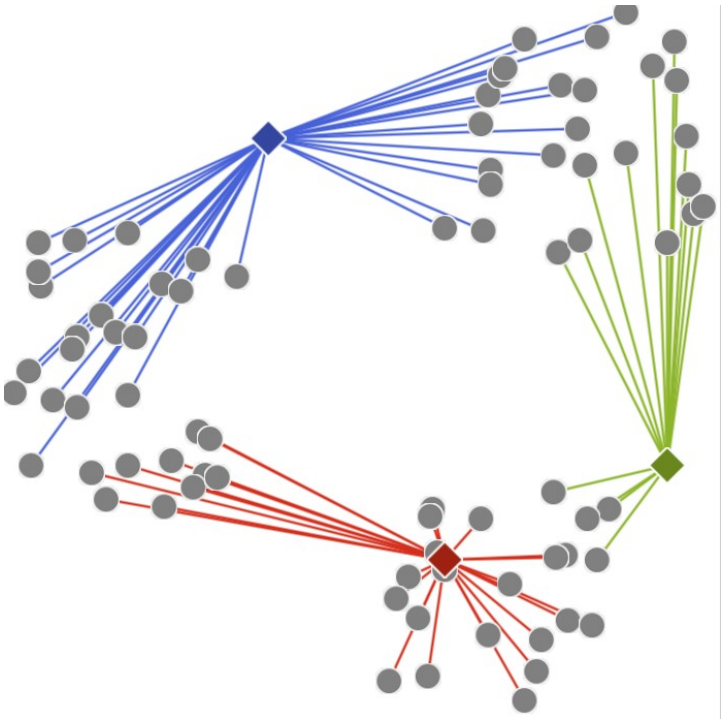
1. Đầu tiên chuẩn bị dữ liệu cần phân cụm. Tiếp theo chọn số lượng cụm(cluster) cần phân chia. Chọn số cluster là 3. Dữ liệu được biểu diễn dưới dạng các điểm. Cự ly của các dữ liệu được hiểu là độ dài đoạn thẳng nói giữa 2 điểm với nhau



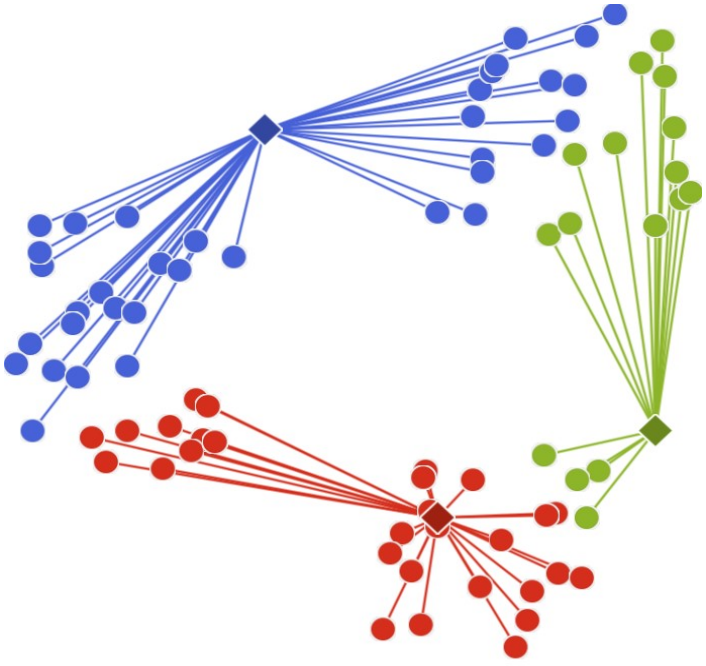
2. Chọn ngẫu nhiên 3 điểm làm trung tâm(center) của cluster



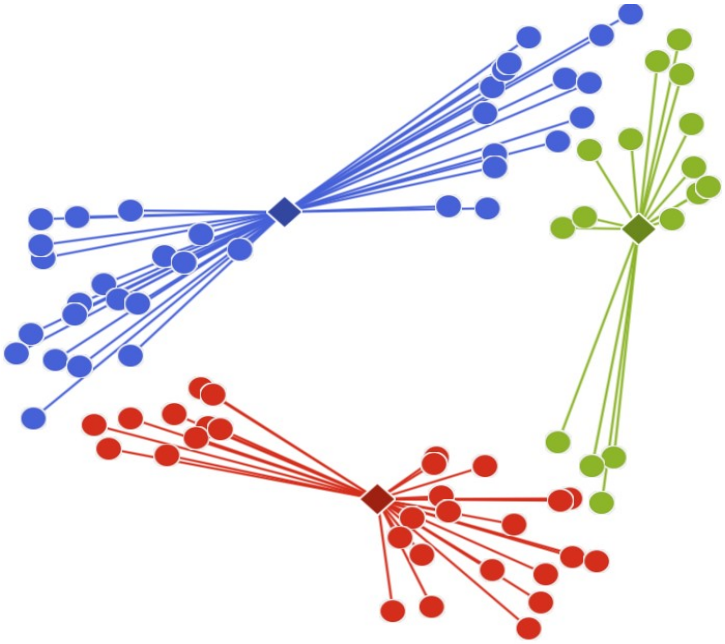
3. Với các điểm dữ liệu không được chọn là điểm trung tâm thì tính toán khoảng cách từ chính điểm đó đến các cluster và quyết định cluster nào gần với mình nhất.



4. Từ bước tính toán trên, tiến hành phân loại các điểm về các cluster đã quyết định (cluster gần nó nhất). Vậy là đã phân ra được 3 cụm.

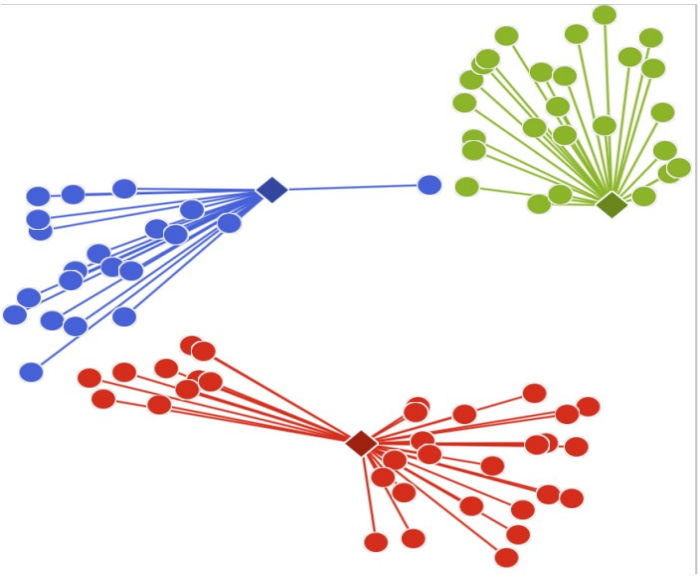


5. Bước trên chúng ta đã thu được 3 cụm, bây giờ tiến hành tính trọng tâm của các điểm dữ liệu của từng cụm. Sau đó di chuyển điểm trung tâm của cụm sang vị trí vừa tính được.



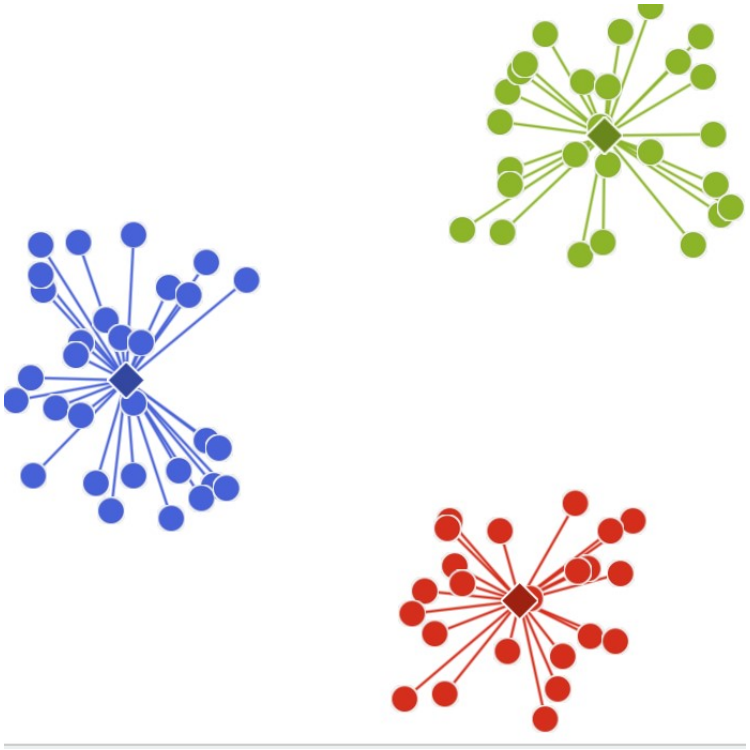
Vị trí mà 3 điểm trung tâm của cluster vừa di chuyển đến được hiểu ngắn gọn chính là điểm trung tâm đang di chuyển đến vị trí chính xác hơn.

6. Một lần nữa tiến hành bước 3, tính toán lại khoảng các các điểm đến các điểm trung tâm. Sau đó phân loại lại các điểm dữ liệu về các cụm.

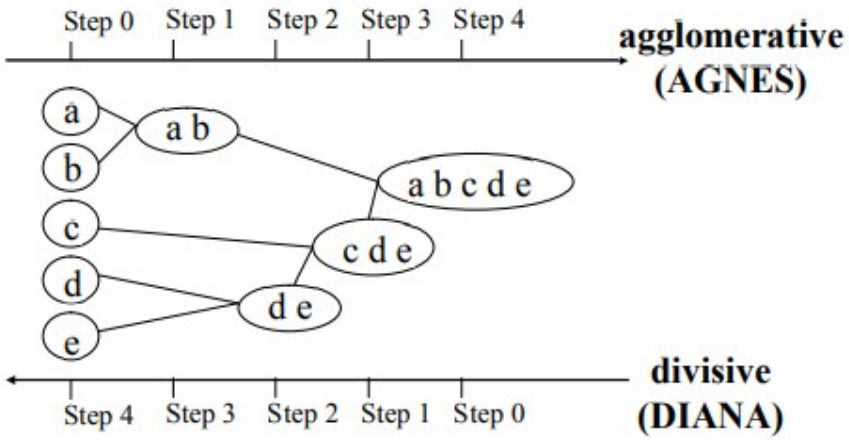


7. Sau đó lặp lại quá trình di chuyển cluster trung tâm và phân loại lại các điểm về các cụm gần nhất.

   Quá trình này sẽ dừng khi sau khi dữ liệu sau khi phân cụm lại không thay đổi gì so với lần trước.



3.3. Giải thuật AGG



- Bottom up

  - Bắt đầu với những clusters chỉ là 1 phần tử

  - Ở mỗi bước, merge 2 clusters gần nhau thành 1

  - Khoảng cách giữa 2 clusters: 2 điểm gần nhất từ 2 clusters, hoặc khoảng cách trung bình, etc.

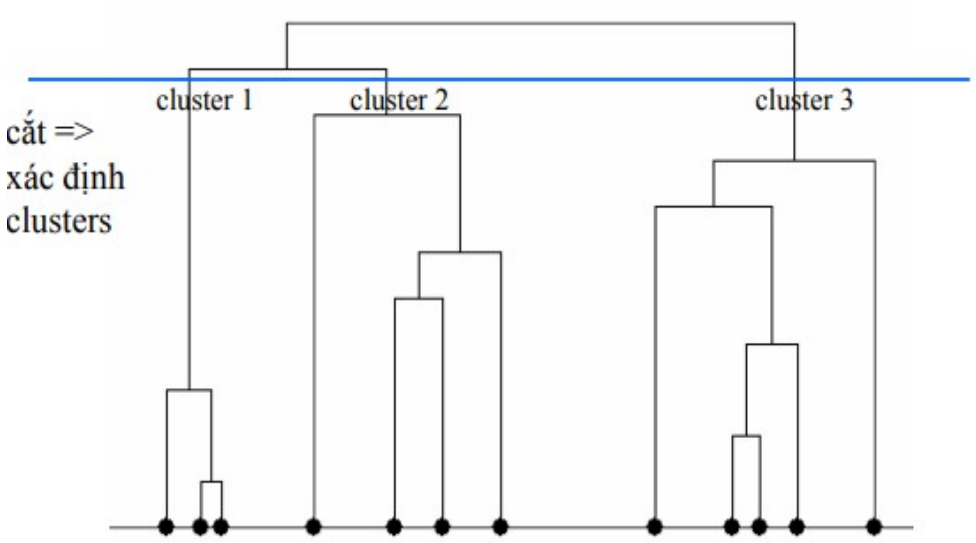
- Top down

  - Bắt đầu với 1 cluster là tất cả dữ liệu

  - Tìm 2 cluster con

  - Tiếp tục đệ quy trên 2 clusters con

- Kết quả sinh ra dendrogram



3.4. Tích hợp giải thuật vào trang web

- Để khởi tạo được chức năng chính của trang web là gom cụm dữ liệu thì việc kế tiếp sau khi cài đặt giải thuật là tích hợp giải thuật vào trang web.
- Giải thuật được cài đặt dựa trên ngôn ngữ python ngôn ngữ lập trình phổ biến nhất hiện nay.

- Python hỗ trợ các thư viện dùng để tích hợp giải thuật vào web. Thư viện Flask là thư viện được sử dụng.

- Thư viện Flask
  - Flask là một web framworks, nó thuộc loại micro-framwork được xây dựng bằng ngôn ngữ lập trình Python. Flask cho phép xây dựng các ứng dụng web từ đơn giản tới phức tạp. Nó có thể xây dựng các API nhỏ, ứng dụng web chẳng hạn như các trang web, blog, trang wiki hoặc một website dựa theo thời gian hay thậm chí một trang web thương mại. Flask cung cấp một công cụ, các thư viện và các công nghệ hỗ trợ làm những công việc trên
  - Flask là một micro-framework. Điều này có nghĩa là Flask là một môi trường độc lập, ít sử dụng các thư viện khác bên ngoài. Do vậy, Flask có ưu điểm là nhẹ, có rất ít lỗi do ít bị phụ thuộc cũng như dễ dàng phát hiện và xử lý các lỗi bảo mật.

CHƯƠNG 3: KIỂM THỬ VÀ ĐÁNH GIÁ

1. Mục tiêu

- Kiểm tra các chức năng của trang web có hoạt động đúng như mong muốn hay không.

- Phát hiển và khắc phục các lỗi, sự cố không mong muốn trong quá trình xây dựng và thiết kế trang web.

- Tìm ra những lỗi tiềm ẩn, đảm bảo trang web hoạt động đúng yêu cầu.

- Cung cấp cơ sở, tài liệu cho công đoạn bảo trì.

2. Nghi thức kiểm tra

- Kiểm thử giao diện: Có hiển thị đúng như mong muốn hay không.

- Kiểm thử cài đặt: Tìm và sửa các lỗi xảy ra trong quá trình cài đặt và thiết kế trang web.

- Kiểm thử các chức năng: Nhập dữ liệu, chọn cột, xuất đồ thị, chọn nhóm, xuất kết quả hình ảnh.

3.Kết quả

- Về giao diện: Bố cục hài hòa và đồ họa giao diện thân thiện dễ sử dụng

- Về chức năng:

1. Nhập dữ liệu -> Thành công
2. Chọn cột -> Thành công
3. Xuất đồ thị -> Thành công
4. Chọn nhóm -> Thành công
5. Xuất kết quả hình ảnh -> Thành công

PHẦN KẾT LUẬN

1.Kết quả đạt được

Xây dựng được trang web gom nhóm dữ liệu trên nền tảng ngôn ngữ Python tích hợp 2 giải thuật gom nhóm là K-Means là Agg. Đáp ứng nhu cầu cần xử lý dữ liệu của con người mà không cần tới code và hoàn toàn miễn phí.

2.Hạn chế

- Giao diện còn khá đơn giản chưa được đặc sắc
- Chức năng nhập dữ liệu chưa được đầy đủ vì chưa thể nhập các file khác như exe, pdf, …
- Chức năng chọn cột chỉ có thể chọn 2 cột.

3.Hướng phát triển

- Thêm màu sắc cho giao diện
- Thêm các giải thuật gom cụm mới để cho trang web phong phú
- Điều chỉnh các chức năng cho phù hợp với thực tiễn

TÀI LIỆU THAM KHẢO

1. Giáo trình, slide bài giảng Khai khoán dữ liệu của thầy Lưu Tiến Đạo.
2. Slide bài giảng "Giải thuật gom cụm Clustering algorithms" của thầy Đỗ Thanh Nghị