

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 2**

-----🙞🙜🕮🙞🙜-----

****

**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

**MÔN: PHÁT TRIỂN HỆ THỐNG THÔNG MINH**

**XÂY DỰNG WEB BÁN THUỐC TÂY**

**GIẢNG VIÊN HƯỚNG DẪN: NGUYỄN NGỌC DUY**

**NHÓM 21**

**N20DCCN040 – Phùng Đức Mạnh**

**N20DCCN079 - ĐẶNG ĐỨC TRỌNG**

**Thành phố Hồ Chí Minh, 4/11/2023**

Mục lục

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 1](#_Toc150016336)

[1.1. Lý do chọn đề tài 1](#_Toc150016337)

[1.2. Mục tiêu nghiên cứu 2](#_Toc150016338)

[1.3. Đối tượng và phạm vi nghiên cứu 2](#_Toc150016339)

[1.4. Phương pháp nghiên cứu 2](#_Toc150016340)

[1.5. Một số hạn chế dự kiến 2](#_Toc150016341)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 2](#_Toc150016342)

[2.1. Giới thiệu về học máy 2](#_Toc150016343)

[2.2. Giới thiệu về bài toán phân cụm 2](#_Toc150016344)

[2.3. Thuật toán Hierarchical Divisive Clustering 4](#_Toc150016345)

[2.3.1. Mã giả: 4](#_Toc150016346)

[2.3.2. Công thức toán học trong Hierarchical Divisive Clustering: 4](#_Toc150016347)

[2.4. Hình ảnh minh họa sau khi phân cụm 4](#_Toc150016348)

[2.5. Dữ liệu 5](#_Toc150016349)

[2.5.1. Nghiệp vụ 5](#_Toc150016350)

[2.5.2. Xác định thuộc tính và mô tả dữ liệu 6](#_Toc150016351)

[CHƯƠNG 3: XÂY DỰNG CẤU HÌNH 7](#_Toc150016352)

[3.1. Đọc dữ liệu từ bảng Thuốc trong database và xử lý dữ liệu 7](#_Toc150016353)

[3.2. Áp dụng thuật toán phân cụm cho dữ liệu 9](#_Toc150016354)

[3.3. Chèn dữ liệu là kết quả của thuật toán vào lại bảng Thuốc liên quan trong database 14](#_Toc150016355)

[CHƯƠNG 4: THIẾT KẾ HỆ THỐNG 17](#_Toc150016356)

[4.1. Thiết kế cơ sở dữ liệu 17](#_Toc150016357)

[4.2. Thiết kế giao diện 19](#_Toc150016358)

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

* 1. Lý do chọn đề tài
* Trong thời đại ngày nay, sử dụng tri thức đã trở thành động lực chủ chốt cho tăng

trưởng kinh tế quốc gia, cho tăng cường năng lực cạnh tranh của doanh nghiệp. Đồng

thời, dung lượng dữ liệu số tăng rất nhanh chóng, chúng đã trở thành nguồn tài nguyên

ẩn thông tin và tri thức có tiềm năng hữu ích cho phát triển kinh tế và tăng cường năng

lực cạnh tranh. Nghiên cứu và triển khai các phương pháp tự động phát hiện các mẫu

mới, có giá trị, hữu ích tiềm năng và hiểu được trong khối dữ liệu đồ sộ, khắc phục hiện

tượng giàu về dữ liệu mà nghèo về thông tin, hướng tới mục tiêu tăng cường tài nguyên

tri thức là hết sức cần thiết và có ý nghĩa. Theo đó, việc định nghĩa khai phá dữ liệu được

hiểu theo nghĩa đơn giản chính là việc rút trích thông tin hay tri thức mới và có ích từ

nguồn dữ liệu khổng lồ.

* Trong lĩnh vực thương mại, tính cạnh tranh kinh doanh là rất cao, cho nên việc

phân tích dữ liệu để đưa đến người dùng những dịch vụ tốt hơn, có nhiều tiện ích cho

khách hàng hơn là điều cần thiết. Cụ thể trong chuyện mua bán điện thoại, khi đưa vào

thông tin của một chiếc máy nào đó, người ta đồng thời phải đưa ra quyết định để biết

loại máy đó thuộc nhóm nhu cầu nào. Tuy nhiên, với mỗi thông tin được đưa vào, người

đánh giá, xếp loại chúng vào những nhóm nhu cầu phải có kiến thức, phải là chuyên gia

am hiểu về các dòng máy, am hiểu về các thông tin chi tiết của máy. Như vậy nếu ta đưa

vào rất nhiều thông tin, người này sẽ phải tốn rất nhiều thời gian để xem xét và đưa ra

quyết định, như vậy sẽ dẫn đến nguy cơ mất cơ hội trong kinh doanh. Do đó, việc khai

phá dữ liệu nhằm giúp đỡ con người trong công việc luôn được thúc đẩy mạnh mẽ, làm

sao với các dữ liệu đã thu thập, những hành động tương ứng từ dữ liệu đó có thể mang

lại lợi ích kịp thời, tối đa cho doanh nghiệp.

* Với những lý do trên cùng với sự cho phép của thầy Nguyễn Ngọc Duy, nhóm

chúng em quyết định chọn đề tài “**Xây dựng web bán thuốc tây**” với chức năng thông minh: **gợi ý thuốc liên quan với thuốc mà người dùng đang xem**

* 1. Mục tiêu nghiên cứu
* Nghiên cứu mô hình thuật toán phân cụm: **Hierarchical Divisive Clustering**
* Ứng dụng của thuật toán phân cụm Hierarchical Divisive Clustering vào việc gợi ý những thuốc liên quan (chung cụm) với thuốc đang xem
  1. Đối tượng và phạm vi nghiên cứu
* Đối tượng nghiên cứu: dữ liệu về điện thoại
* Phạm vi nghiên cứu: Thị trường điện thoại di động trên toàn thế giới
  1. Phương pháp nghiên cứu
* Phương pháp thực nghiệm khoa học: Dựa vào tập dữ liệu đã nhận được, tiến hành

tiền xử lý và sử dụng các thuật toán phân cụm Hierarchical Divisive Clustering, đánh giá tính hiệu quả của thuật toán dựa trên tập dữ liệu kiểm thử

* 1. Một số hạn chế dự kiến
* Số lượng mẫu dữ liệu càng nhiều thì thời gian tính toán của thuật toán càng lâu

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

* 1. Giới thiệu về học máy
* **Học máy** (machine learning ) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc

nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải

quyết những vấn đề cụ thể.

* **Học máy** được chia thành 2 loại chính bao gồm: học có giám sát và học không giám

sát.

* + Học có giám sát là phương pháp sử dụng những dữ liệu được gán nhãn sẵn

để suy luận ra quan hệ giữa đầu vào và đầu ra. Sau khi tìm hiểu cách tốt nhất

để mô hình hóa các mối quan hệ cho dữ liệu được gán nhãn, thuật toán huấn

luyện sẽ được sử dụng cho các bộ dữ liệu mới. Học tập có giám sát có thể

được nhóm lại thành các vấn đề về phân loại và hồi quy.

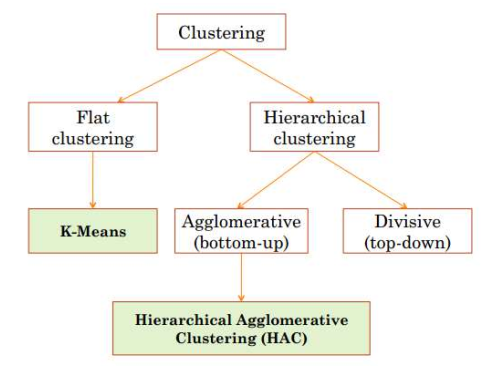
* + Học không giám sát sử dụng những dữ liệu chưa được gán nhãn sẵn để suy

luận và tìm cách để mô tả dữ liệu cùng cấu trúc của chúng. Ứng dụng của

học không giám sát đó là hỗ trợ phân loại thành các nhóm có đặc điểm tương

đồng

* 1. Giới thiệu về bài toán phân cụm
* Khái niệm: **Phân cụm (Clustering)** thuộc loại học không giám sát (Unsupervised learning) là một dữ liệu là bài toán gom nhóm các đối tượng dữ liệu vào thánh từng cụm (cluster) sao cho các đối tượng trong cùng một cụm có sự tương đồng theo một tiêu chí nào đó
  + Bài toán theo đề tài: phân nhóm các thuốc liên quan với nhau dựa theo 4 đặc trưng: tên thuốc, công dụng, số lượng dược chất, tổng trọng số dược chất
* Phân cụm chia thành các loại:



Hình 2.2.1. Các loại phân cụm

* Ở đề tài này, ta sẽ tìm hiểu về Phân cụm theo cấp bậc
* Phân cụm theo cấp bậc là mô hình phân cụm dựa trên kết nối, nhóm các điểm dữ liệu gần nhau dựa trên thước đo độ tương tự hoặc khoảng cách.
  + Giả định là các điểm dữ liệu gần nhau thì giống nhau hoặc có liên quan hơn so với các điểm dữ liệu ở xa nhau hơn.
  + **Agglomerative Clustering** (phân cụm kết tụ): Nó còn được gọi là cách tiếp cận từ dưới lên hoặc phân cụm kết tụ phân cấp (HAC).Một cấu trúc có nhiều thông tin hơn tập hợp các cụm không có cấu trúc được trả về bởi phân cụm phẳng.Thuật toán phân cụm này không yêu cầu chúng ta xác định trước số lượng cụm. Các thuật toán từ dưới lên xử lý mỗi dữ liệu dưới dạng một cụm đơn ngay từ đầu và sau đó liên tục kết hợp các cặp cụm cho đến khi tất cả các cụm được hợp nhất thành một cụm duy nhất chứa tất cả dữ liệu.
  + **Divisive Clustering** (Phân cụm phân chia) ngược lại với phân cụm HAC với cách tiếp cận từ trên xuống. Nó bắt đầu với tất cả các điểm dữ liệu trong một cụm duy nhất, sau đó phân chia đệ quy các cụm thành các cụm con nhỏ hơn dựa trên sự khác biệt của chúng
* Lý do chọn thuật toán Divisive Clustering:
  + Phân cụm tổng hợp đưa ra quyết định bằng cách xem xét các mô hình cục bộ hoặc các điểm lân cận mà không tính đến việc phân phối dữ liệu toàn cầu ngay từ đầu.Những quyết định ban đầu này không thể được hoàn tác. Trong khi đó, phân cụm phân chia sẽ xem xét phân phối dữ liệu toàn cầu khi đưa ra quyết định phân vùng cấp cao nhất.
  1. Thuật toán Hierarchical Divisive Clustering
     1. Mã giả:

Mã giả thuật toán hierarchical divisive cluster áp dụng cho đề tài

B1: tiền xử lý dữ liệu: chuyển dữ liệu text sang số, tính khoảng cách giữa các điểm dữ liệu với nhau và lưu vào matrix

B2: chọn một cụm có đường kính lớn nhất giữa 2 điểm trong cụm đó ra để phân cụm (nếu phân lần đầu thì ta lấy cụm gốc)

B3: lấy mỗi điểm trong cụm cần phần và tính toán khoảng cách trung bình tới các điểm còn lại trong cùng cụm

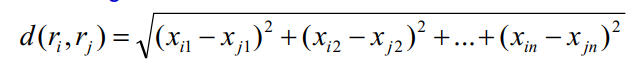
B4: lấy mỗi điểm trong cụm cần phần và tính toán khoảng cách trung bình tới các điểm còn lại khác cụm

B5: lấy hiệu 2 khoảng cách trung bình tính ở B3 và B4, nếu hiệu nào lớn nhất thì ta lấy điểm đó ra khỏi cụm và thêm tạm thời vào cụm mới

B6: quay lại B3, đến khi nào hiệu <= 0 thì quay lại bước 2, khi nào đường kính lớn nhất <=0 thì chuyển sang bước 7

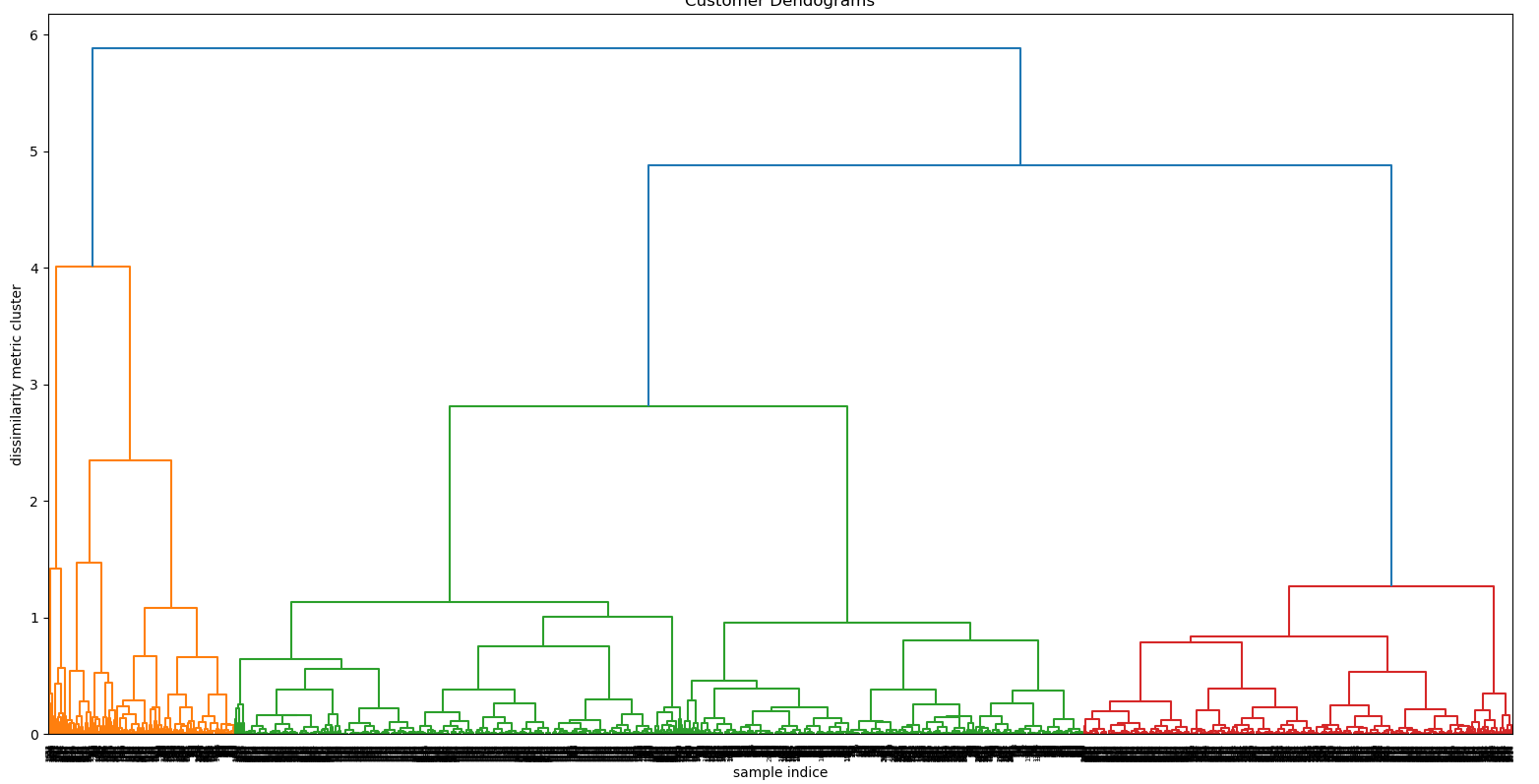
B7: trả về danh sách cụm được phân theo sơ đồ cây. Kết thúc thuật toán

* + 1. Công thức toán học trong Hierarchical Divisive Clustering:
* Ở bước tiền xử lý dữ liệu, sử dụng công thức khoảng cách **Euclidean** để tính khoảng cách giữa các mẫu thuốc với nhau. Sau khi tính toán xong sẽ lưu lại dưới dạng ma trận kích thước NxN (với N: số lượng mẫu thuốc).



Hình 2.3.2.1 Công thức khoảng cách **Euclidean**

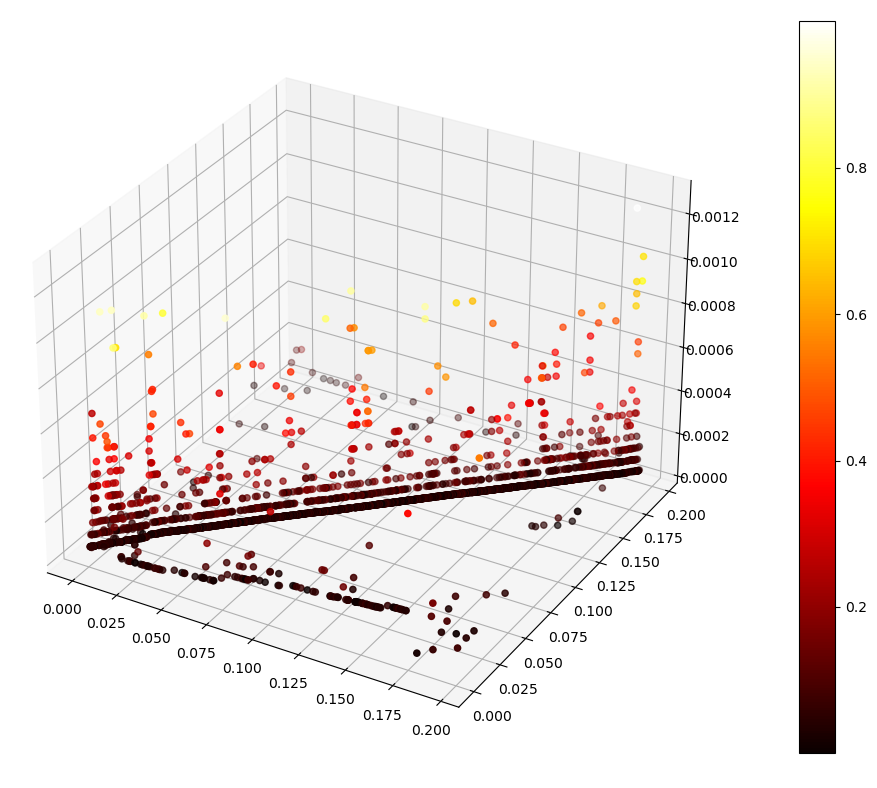
* Trong thuật toán sẽ có những công thức:
  + Tính khoảng cách trung bình từ một mẫu thuốc với các thuốc khác nằm chung cụm.
  + Tính khoảng cách trung bình từ một mẫu thuốc với các thuốc khác nằm khác cụm.
  1. Hình ảnh minh họa sau khi phân cụm



Hình 2.4.1. Hình ảnh các cụm được phân

* 1. Dữ liệu
     1. Nghiệp vụ

Hình ảnh mô phỏng dữ liệu không gian 4 chiều do mỗi thuốc có 4 thuộc tính:



Hình 2.5.1.1 Hình ảnh mô phỏng các điểm dữ liệu thuốc

* + 1. Xác định thuộc tính và mô tả dữ liệu

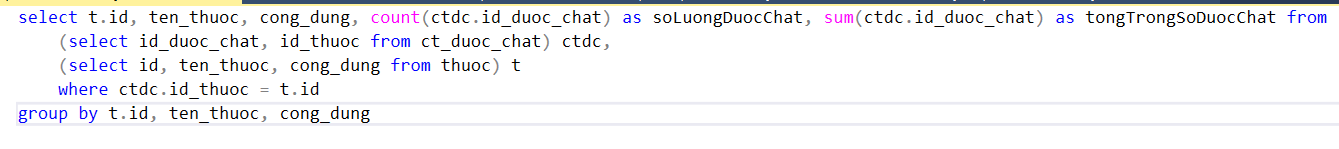
Dữ liệu lấy từ database có dạng sau:



Hình 2.5.2.1. Dữ liệu thuốc lấy từ database

* soLuongDuocChat: một thuốc sẽ chứa nhiều dược chất, thì thuộc tính này sẽ lưu tính toán tất cả dược chất có trong thuốc đó.
* tongTrongSoDuocChat: mỗi dược chất sẽ có một id riêng biệt, với thuộc tính này sẽ lưu kết quả tính tổng các id của các dược chất có trong thuốc đó. Vẫn có hạn chế chưa khắc phục được: có trường hợp số lượng dược chất giống nhau, trọng số bằng nhau nhưng dược chất trong thuốc đó khác nhau.

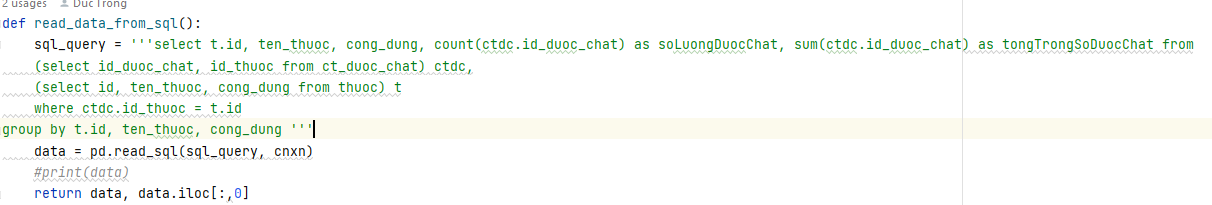
Với lệnh truy vấn sql:



Hình 2.5.2.2. Lệnh truy vấn để lấy dữ liệu từ database

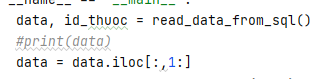
# CHƯƠNG 3: XÂY DỰNG CẤU HÌNH

* 1. Đọc dữ liệu từ bảng Thuốc trong database và xử lý dữ liệu
* Lấy dữ liệu từ database



Hình 3.1.1. Lấy dữ liệu từ database và lưu vào biến (có kiểu pandas) trong python

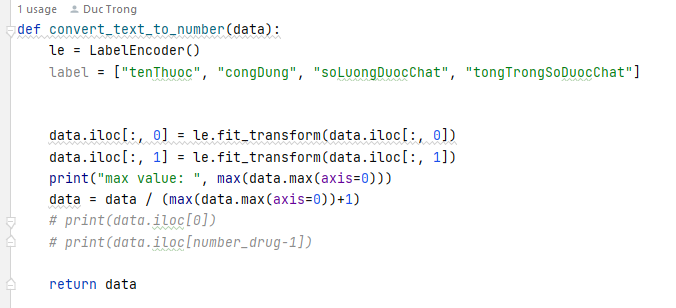
* Chia cắt dữ liệu: tách id thuốc vào mảng 1 chiều khác, 4 thuộc tính còn lại cất vào mảng hai chiều



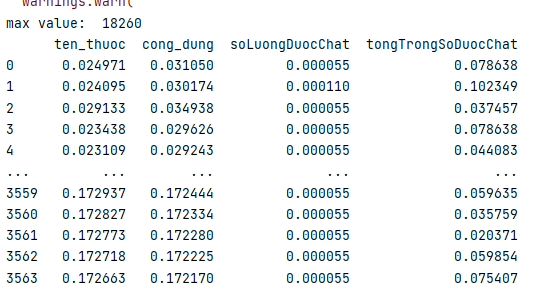
Hình 3.1.2. Tách dữ liệu đọc được từ database

* Chuyển đổi dữ liệu trong cột tenThuoc (tên thuốc) và congDung (công dụng) từ văn bản sang số, sau đó chia toàn bộ dữ liệu cho giá trị lớn nhất trong tất cả dữ liệu để chuẩn hóa dữ liệu sang số thập phân có phạm vi < 1





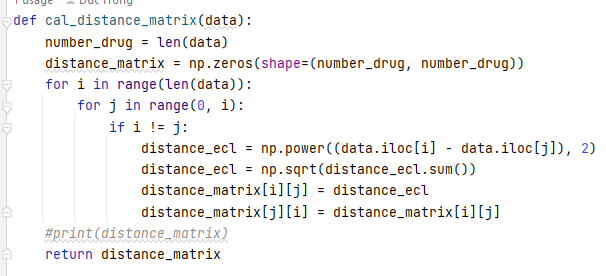
Hình 3.1.3. Chuyển dữ liệu từ văn bản sang số



Hình 3.1.4. Chia toàn bộ cho giá trị max là 18260 để có dữ liệu mới là các dữ liệu thập phân

* Tính toán khoảng cách **Euclidean** giữa các điểm dữ liệu (thuốc)





Hình 3.1.4. Hàm tính toán khoảng cách **Euclidean**



Hình 3.1.5. Lưu kết quả ma trận khoảng cách vào file nhị phân distance\_matrix2.npy



Hình 3.1.6. Lưu kết quả mảng một chiều chứa id thuốc vào file nhị phân id\_thuoc.npy

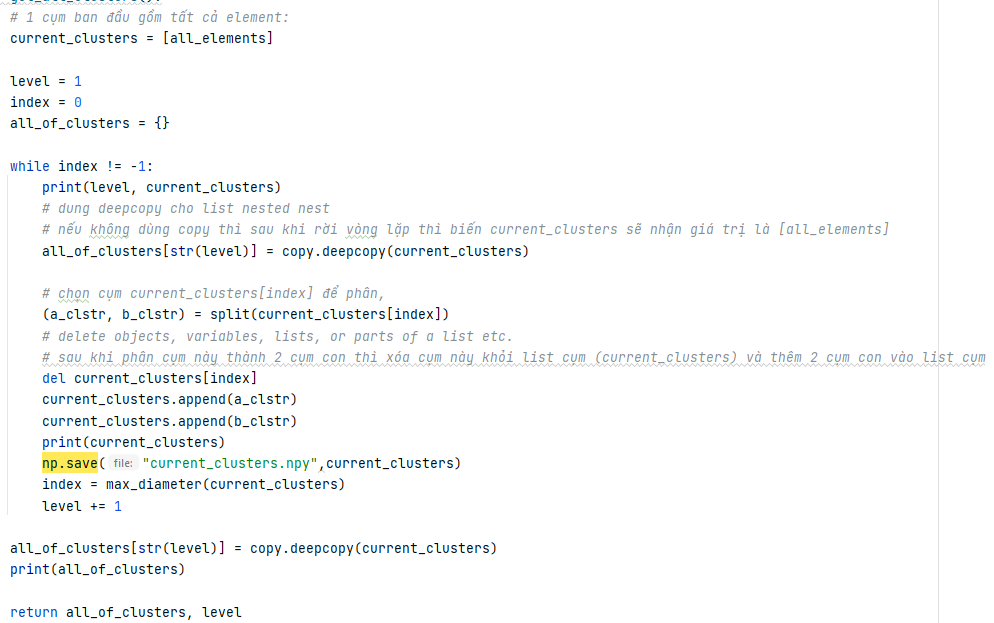
* 1. Áp dụng thuật toán phân cụm cho dữ liệu
* Đọc dữ liệu từ 2 file nhị phân distance\_matrix2.npy, d\_thuoc.npy



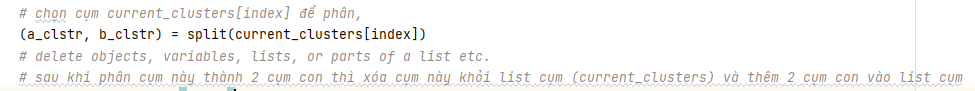
* Tạo một biến kiểu pandas index cột và hàng là id\_thuoc, giá trị của các hàng và cột là dữ liệu từ ma trận khoảng cách

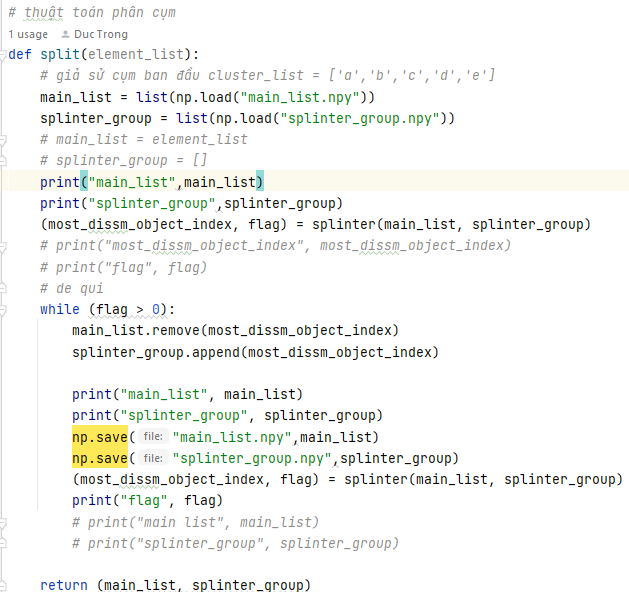


* Tiến hành phân cụm:
  + Level: cấp độ cụm, level = 0 là cụm ban đầu chứa tất cả thuốc, giá trị level tăng dần đến khi level = <giá trị lớn nhất> là cụm cuối cùng mỗi thuốc là một cụm
  + Current\_clusters: danh sách cụm sau mỗi level (mỗi lần phân cụm)



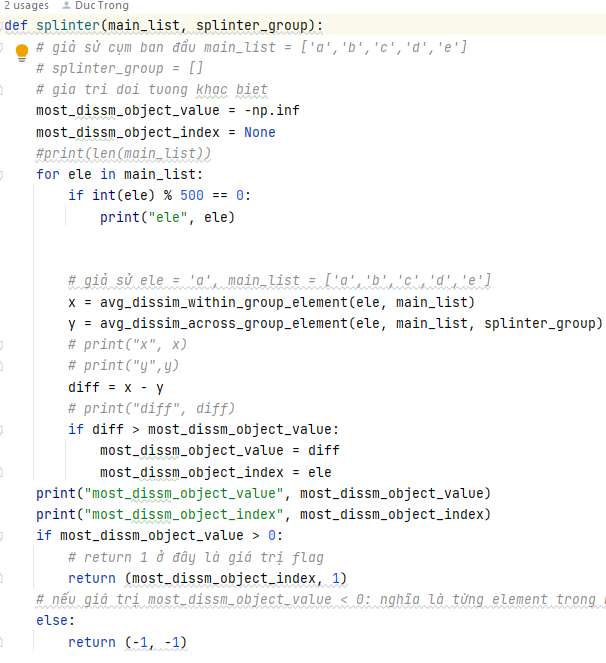
* Hàm split(cụm cần phân): thực hiện phân một cụm thành 2 cụm con
  + Main\_list: các id thuốc của cụm được phân, sẽ thay đổi khi id thuốc này được phân sang cụm splinter\_group
  + splinter\_group: cụm mới chứa các id thuốc được phân từ cụm main\_list sang
  + flag (cờ hiệu): đánh dấu cụm main\_list có thể phân được nữa không. Flag = 0: ngừng phân cụm và trả về kết quả 2 cụm con main\_list, splinter\_group, Flag = 1: tiếp tục phân cụm



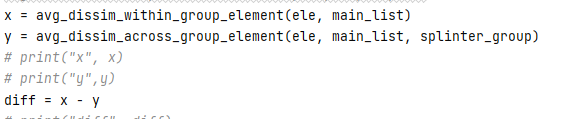


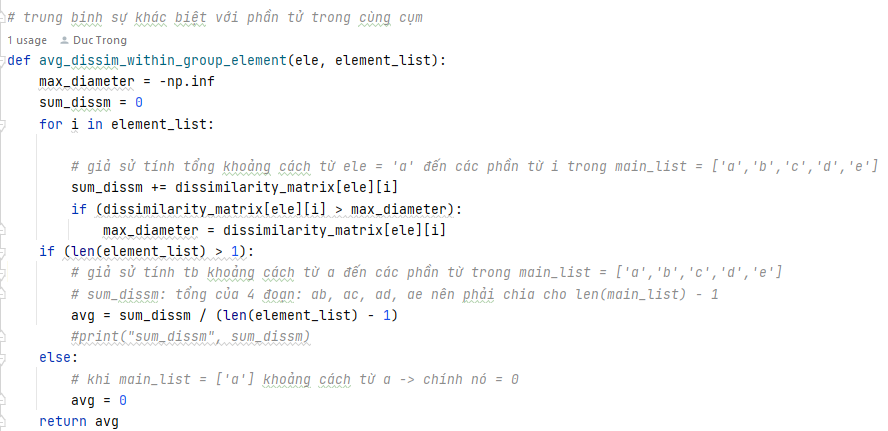
* Hàm splinter:
  + most\_dissm\_object\_value: giá trị hiệu của 2 khoảng cách trung bình:
    - từ một thuốc này tới các thuốc cùng cụm
    - từ một thuốc này tới các thuốc khác cụm
* sao cho giá trị hiệu này lớn nhất
* mục đích: giá trị hiệu lớn nhất tức khoảng cách thuốc này tới các thuốc trong cụm xa hơn các thuốc khác cụm nên chia thuốc này sang cụm khác
* nếu hiệu này nhỏ hơn không thì ngược lại: nên không cần chia thuốc này sang cụm khác
  + most\_dissm\_object\_index: id thuốc có giá trị hiệu lớn nhất ở trên



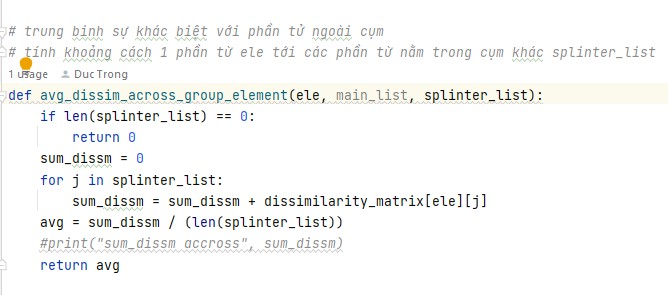


* Hàm avg\_dissim\_within\_group\_element: tính khoảng cách trung bình sự khác biệt với phần tử trong cùng cụm

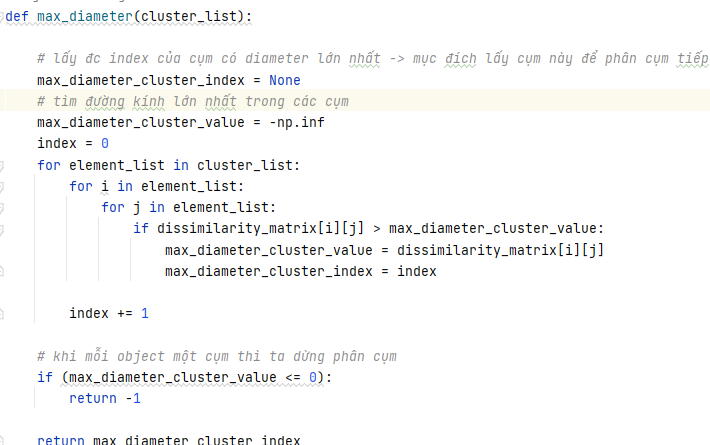




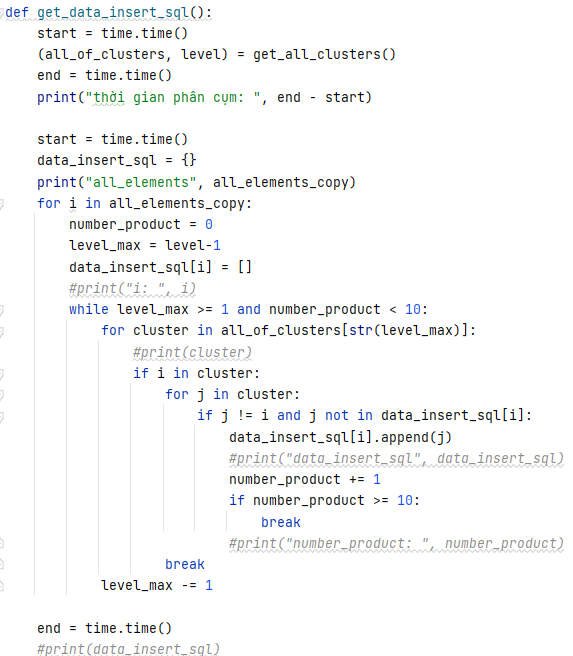
* avg\_dissim\_across\_group\_element: tính khoảng cách trung bình sự khác biệt với phần tử trong khác cụm

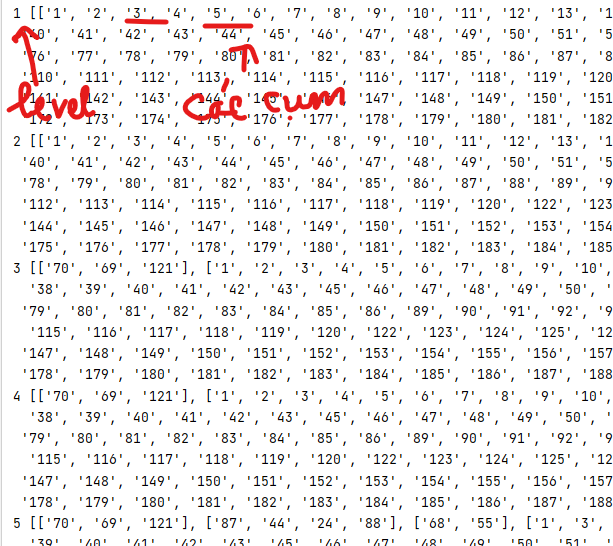


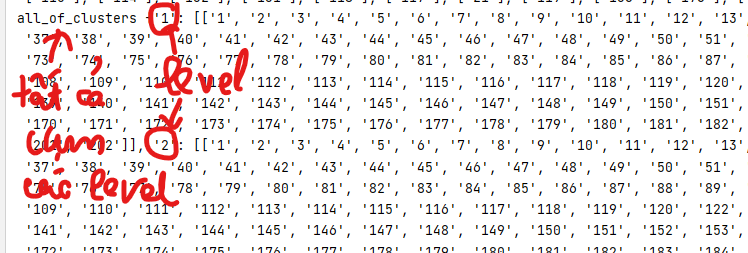
* Hàm max\_diameter: Thuật toán sẽ dừng nếu như việc gộp cụm tạo thành những cụm có độ gắn kết (cohension) thấp hơn. Độ gắn kết là một tiêu chuẩn để đo chất lượng cụm được tạo thành.Thông thường chúng ta có thể đo lường độ gắn kết dựa trên đường kính (diameter) của cụm sau gộp, đường kính được tính bằng khoảng cách lớn nhất giữa hai điểm trong cụm



* 1. Chèn dữ liệu là kết quả của thuật toán vào lại bảng Thuốc liên quan trong database
* Hàm get\_data\_insert\_sql():
  + Sau mỗi lần phân cụm: ta lưu vào biến all\_of\_clusters có kiểu dictionary (từ điển) với keys: level, value: danh sách cụm của level đó
  + Ta khởi tạo biến data\_insert\_sql kiểu dictionary (từ điển) với keys: id thuốc đang xem, value: list n thuốc liên quan (n ở đây chọn cụ thể là 10).
  + Từ biến all\_of\_clusters, ta duyệt ngược từ dưới lên trên (level max tới level = 0): để lấy được thuốc liên quan của từng thuốc và sau mỗi lần tìm thấy thuốc cùng cụm hoặc là cụm cha (level thấp hơn) thì chèn vào value trong biến data\_insert\_sql





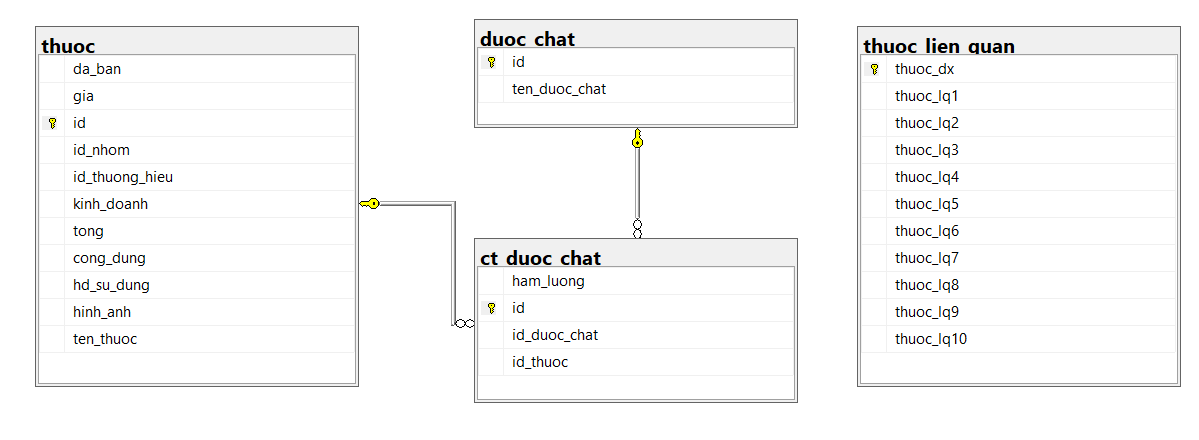


* Hàm write\_data\_to\_sql(): lấy dữ liệu trong biến data\_insert\_sql và chèn vào table thuoc\_lien\_quan trong database

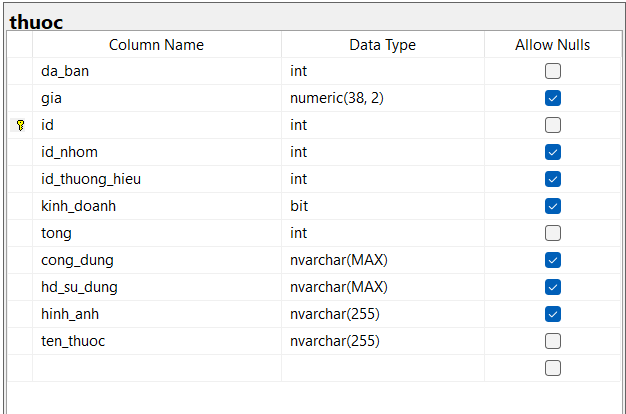


# CHƯƠNG 4: THIẾT KẾ HỆ THỐNG

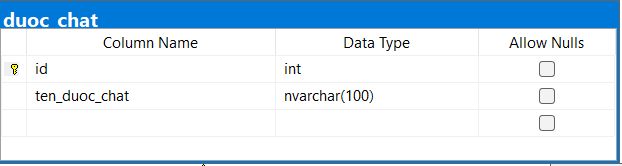
* 1. Thiết kế cơ sở dữ liệu



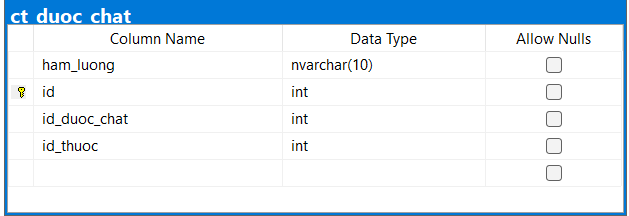
* Kiểu dữ liệu trong bảng Thuốc:



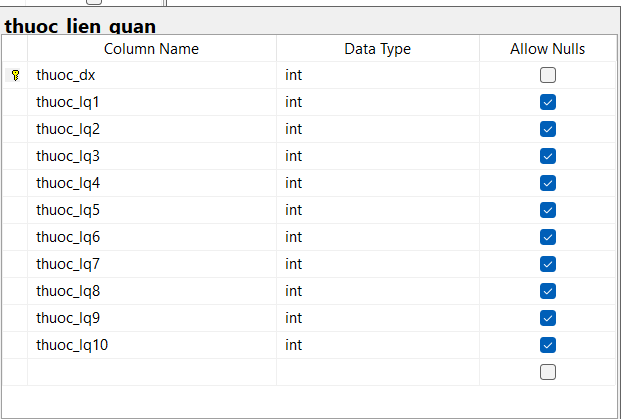
* Kiểu dữ liệu trong bảng dược chất



* Kiểu dữ liệu trong bảng chi tiết dược chất:



* Kiểu dữ liệu trong bảng thuốc liên quan:



* 1. Thiết kế giao diện

