# BÁO CÁO TIẾN ĐỘ THỰC HIỆN ĐỒ ÁN 3

**Đề tài: K-Means Clustering**

**NHÓM THỰC HIỆN:**

**Họ và tên MSSV**

**Trần Bình Cơ 17110105**

**Nguyễn Thái Phương Thảo 17110227**

# I.Tổng quan về thuật toán gom cụm dữ liệu:

## ****1. Giới thiệu về kỹ thuật phân cụm trong Khai phá dữ liệu (Clustering Techniques in Data Mining)****

* Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp **Unsupervised Learning** trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau.
* Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu của của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data reduction, “natural clusters”, “useful” clusters, outlier detection

## 2. Phân loại các phương pháp gom cụm dữ liệu tiêu biểu:

* Phân hoạch (partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó thông qua các giải thuật như K-means, K-medoids,…
* Phân cấp (hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí nào đó.
* Dựa trên mật độ (density-based): dựa trên connectivity and density functions.
* Dựa trên lưới (grid-based): dựa trên a multiple-level granularity structure.
* Dựa trên mô hình (model-based): một mô hình giả thuyết được đưa ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất.

## 3. Kỹ thuật phân cụm có thể áp dụng trong rất nhiều lĩnh vực như:

* Marketing: Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng giá trị, phân loại và dự đoán hành vi khách hàng,…) sử dụng sản phẩm hay dịch vụ của công ty để giúp công ty có chiến lược kinh doanh hiệu quả hơn;
* Biology: Phận nhóm động vật và thực vật dựa vào các thuộc tính của chúng;
* Libraries:  Theo dõi độc giả, sách, dự đoán nhu cầu của độc giả…;
* Insurance, Finance: Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng, phát hiện gian lận tài chính (identifying frauds);
* WWW:  Phân loại tài liệu (dodwcument classification);  phân loại người dùng web (clustering weblog);…jj v

# II.Thuật toán K-means

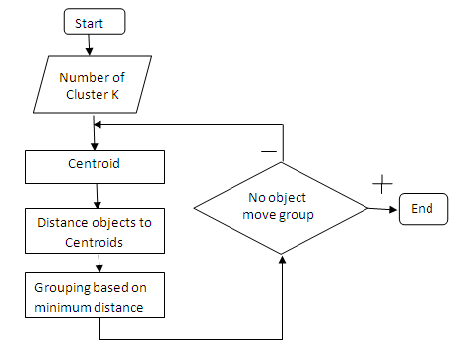
## 1.Giới thiệu về thuật toán K-Means

* K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát(tức là dữ liệu không có nhãn)( **Unsupervised)** và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.
* Thuật toán phân cụm k-means thường được sử dụng trong các ứng dụng cỗ máy tìm kiếm, phân đoạn khách hàng, thống kê dữ liệu,…

## 2.Thuật toán K-Means là gì?

Thuật toán phân cụm k-means là một phương pháp được sử dụng trong phân tích tính chất cụm của dữ liệu. Nó đặc biệt được sử dụng nhiều trong khai phá dữ liệu và thống kê. Nó phân vùng dữ liệu thành k cụm khác nhau. Giải thuật này giúp chúng ta xác định được dữ liệu của chúng ta nó thực sử thuộc về nhóm nào.

## 3.Ý tưởng của thuật toán K-Means



*Hình 1. Ý tưởng thuật toán K-Means*

1. Khởi tạo K điểm dữ liệu trong bộ dữ liệu và tạm thời coi nó là tâm của các cụm dữ liệu của chúng ta.
2. Với mỗi điểm dữ liệu trong bộ dữ liệu, tâm cụm của nó sẽ được xác định là 1 trong K tâm cụm gần nó nhất.
3. Sau khi tất cả các điểm dữ liệu đã có tâm, tính toán lại vị trí của tâm cụm để đảm bảo tâm của cụm nằm ở chính giữa cụm.
4. Bước 2 và bước 3 sẽ được lặp đi lặp lại cho tới khi vị trí của tâm cụm không thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.

## 4.Cách thức thực hiện

### a) Lựa chọn số cụm

* Chỉ việc lựa chọn số cụm k đã có thể tách thành 1 bài toán riêng. Không có 1 con số k nào là hợp lý cho tất cả các bài toán. Bạn có thể đọc hiểu tập dữ liệu của mình để xác định xem trong đó có thể có bao nhiêu cụm? Nhưng không phải lúc nào bạn cũng có thể làm thế. Cách làm duy nhất là bạn hãy thử với từng giá trị k=1,2,3,4,5,… để xem kết quả phân cụm thay đổi như thế nào. Một số nghiên cứu cho thấy việc thay đổi k sẽ có hiệu quả nhưng sẽ dừng lại ở 1 con số nào đó. Như vậy bạn hoàn toàn có thể thử xem dữ liệu của mình tốt với giá trị k nào đó.

### b) Khởi tạo k vị trí ban đầu

* khởi tạo k tâm cụm này phân bố đồng đều trên không gian của bộ dữ liệu. Điều đó có thể làm khi bạn có thể xác định được không gian và tính chất của dữ liệu. Nhưng ít nhất, các tâm cụm mà bạn khởi tạo cũng đừng quá gần nhau, cũng đừng trùng nhau.
* Còn 1 cách cuối cùng là bạn sẽ chạy thuật toán nhiều lần để lấy kết quả tốt nhất trong các lần chạy đó. Với điều kiện là bạn khởi tạo tâm của k cụm ngẫu nhiên.

### c) Về vấn đề tính dừng (hội tụ)

* Đối với những trường hợp dữ liệu phức tạp, thuật toán k-means sẽ rất lâu hoặc không bao giờ hội tụ. Tức là sẽ không bao giờ xác định được tâm cụm cố định để kết thúc bài toán. Hoặc là phải chạy qua rất nhiều bước lặp. Trong những trường hợp như vậy, thay vì phải tìm được k tâm cụm cố định thì ta sẽ dừng bài toán khi sự thay đổi ở một con số chấp nhận được. Tức là giữa hai lần cập nhật tâm cụm thì chênh lệch vị trí giữa tâm cũ và mới nhỏ hơn một số delta cho phép nào đó.

# III.Demo tìm hiểu và ứng dụng thuật toán K-Means vào dữ liệu thực

## 1)Ý tưởng thực hiện:

Sử dụng thuật toán K-Means để phân tích dữ liệu thực tế về sở thích xem phim của mọi người dựa vào đánh giá của họ đối với các tựa phim được phân chia thành nhiều thể loại khác nhau. Sau đó đưa ra gợi ý cho họ về việc chọn lựa phim cho phù hợp

## 2)Cài đặt môi trường

-Thực hiện trên Visual Studio 2019 với ngôn ngữ Python

-Cài đặt các thư viện và import vào chương trình Python:

+Pandas: hỗ trợ xử lý dữ liệu theo cấu trúc Dataframe

+Matplotlib: hỗ trợ vẽ đồ thị

+Numpy: hỗ trợ khả năng tính toán

+Scipy.sparse: hỗ trợ xử lý ma trận

+mpl\_toolkits: hỗ trợ tính toán khoảng cách

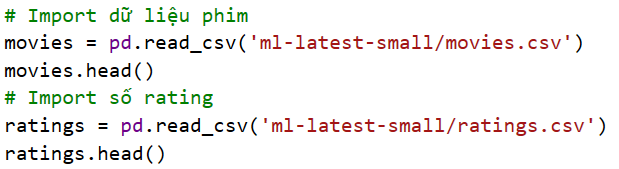
+sklearn: hỗ trợ tính sai số toàn phương trung bình, hệ số hình chiếu của mỗi mẫu và tất cả các mẫu.

+itertools hỗ trợ xử lý các vòng lặp

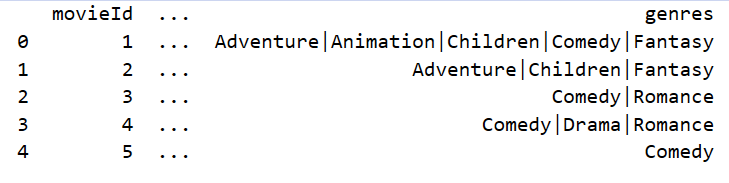
-Tiền xử lý dữ liệu:

+Import dữ liệu phim, rating và xem trước một phần của dữ liệu ban đầu:

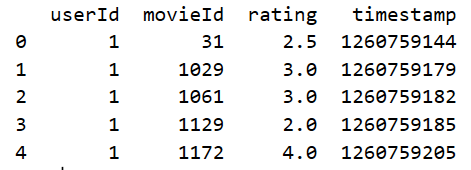
Input:



Output: Phim bao gồm các thể loại



Lượng rating của người xem đối với từng phim



+Ta sẽ tìm hiểu cấu trúc của dữ liệu và xem có bao nhiêu dữ liệu trong file bằng cách:

Input:

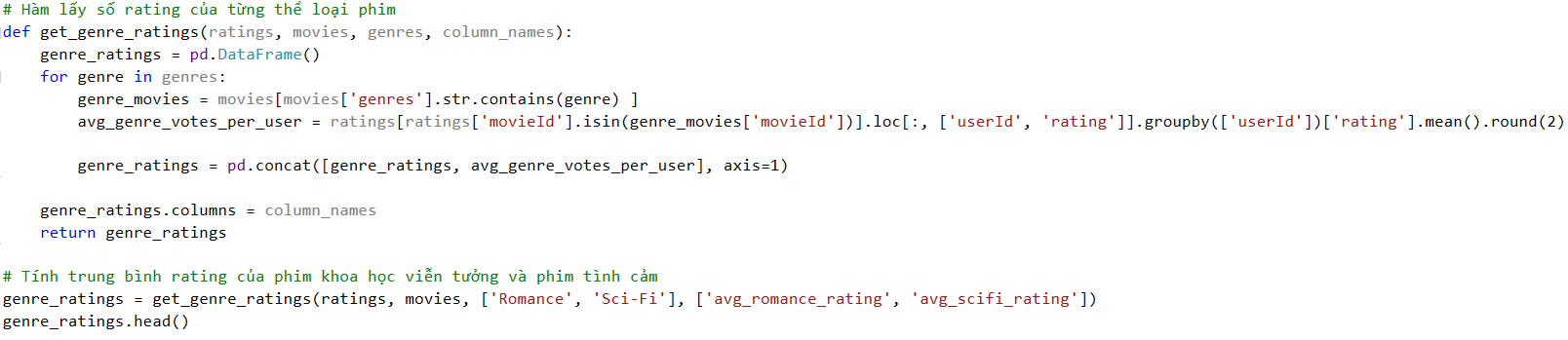


Output:

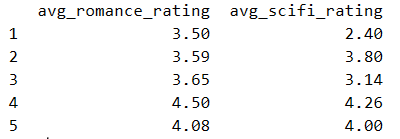


+Xét phim khoa học viễn tưởng và phim tình cảm bắt đầu bằng việc xem xét một nhóm nhỏ người xem về thể loại yêu thích của họ

Input:

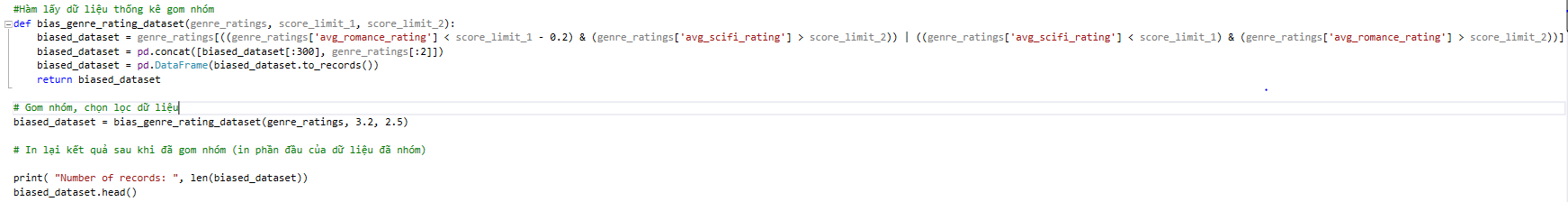


Output:



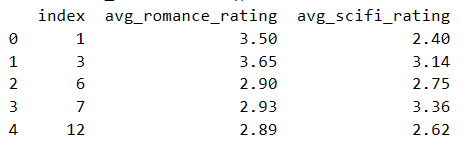
+Để chính xác hơn trong việc tính toán, ta sẽ thống kê và gom nhóm những rating từ những người chỉ thích phim khoa học viễn tưởng và phim tình cảm:

Input:



Output:

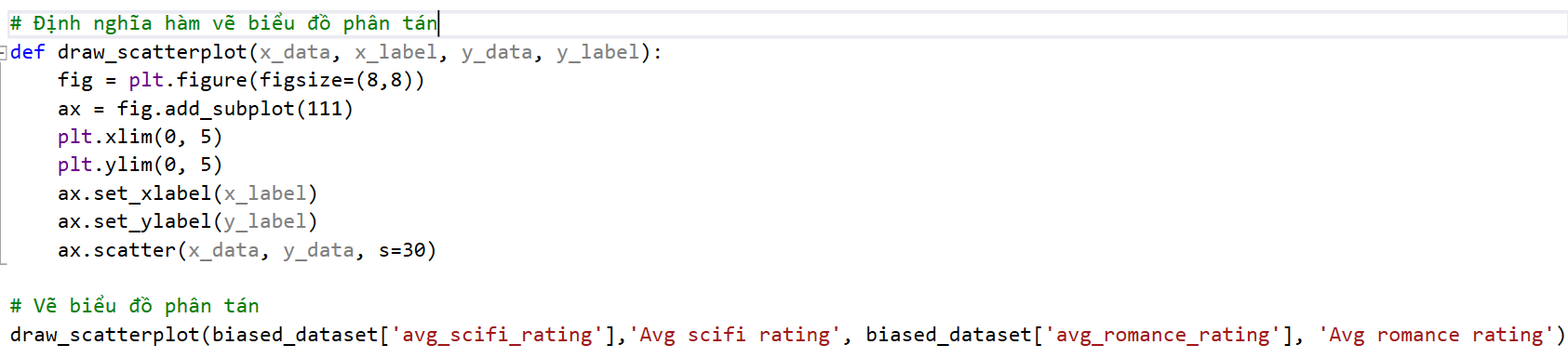




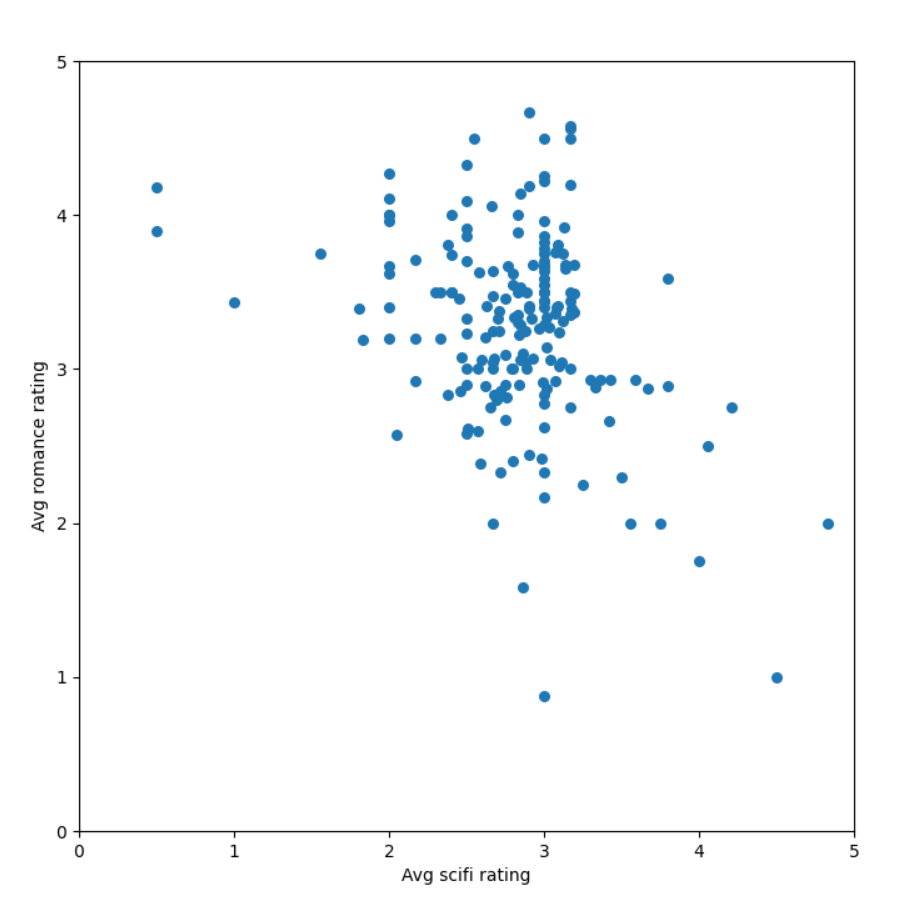
Ta thấy có 183 records và ứng với mỗi records sẽ có một rating cho một bộ phim khoa học viễn tưởng và phim tình cảm

+Tiếp theo ta sẽ thực hiện phân tích hình ảnh hóa để có một cái nhìn tổng quát hơn về tập dữ liệu đã chọn lọc và các đặc điểm của nó:

Input:



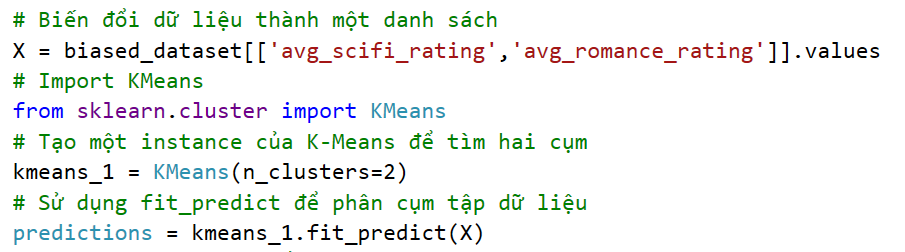
Output:

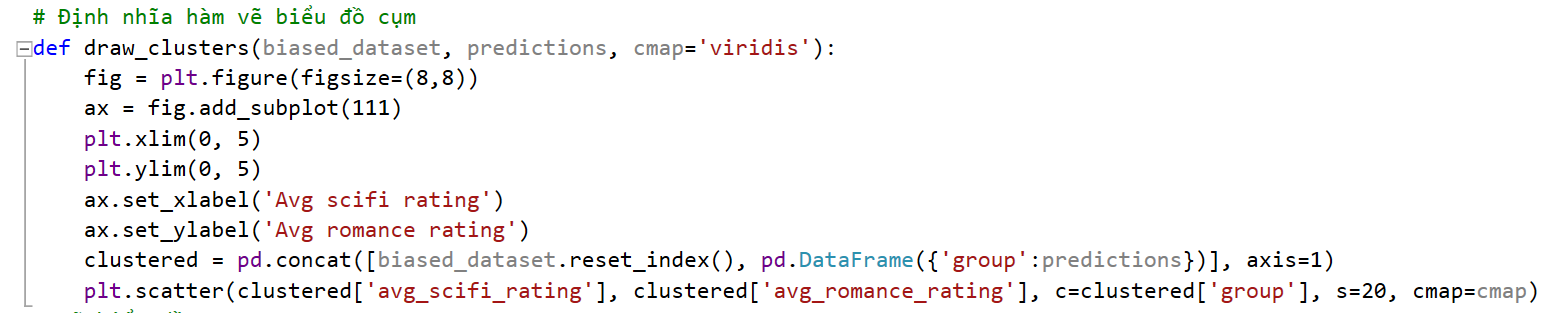


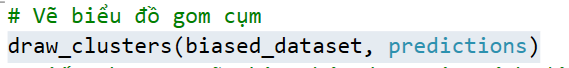
-Áp dụng thuật toán K-Means:

Dữ liệu đã lọc ở phần trước đã hoàn toàn rõ ràng, ta sẽ áp dụng thuật toán K-Means để chia nhỏ mẫu thành hai nhóm riêng biệt:

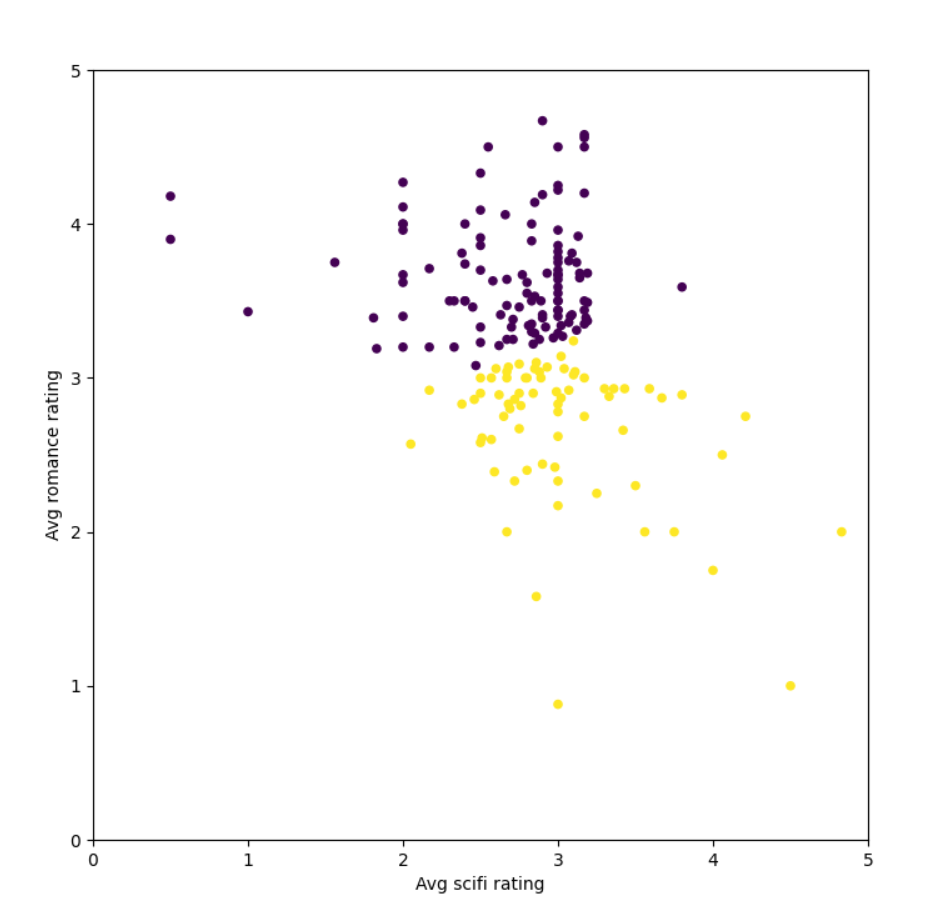
Input:







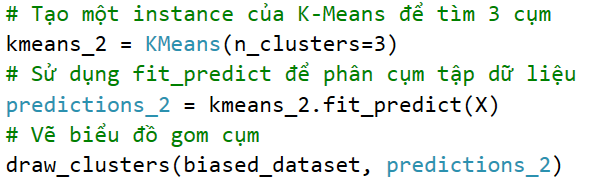
Output:



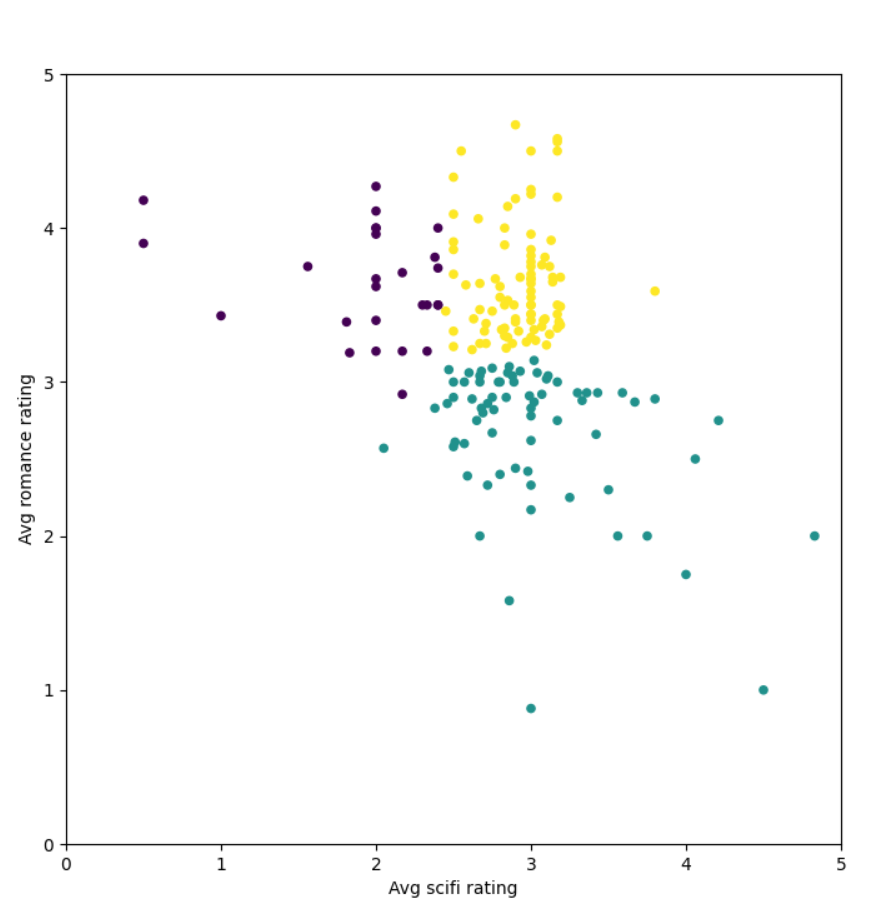
Chúng ta có thể thấy rằng các nhóm chủ yếu dựa trên cách đánh giá phim tình cảm của mỗi người. Nếu xếp hạng phim tình cảm trung bình của họ trên 3 sao, thì họ thuộc một nhóm. Nếu không, họ thuộc nhóm khác.

+Tiếp theo ta sẽ phân thành ba nhóm:

Input:



Output:



Từ đồ thị ta thấy những người thích phim khoa học viễn tưởng cũng được phân định rõ ràng

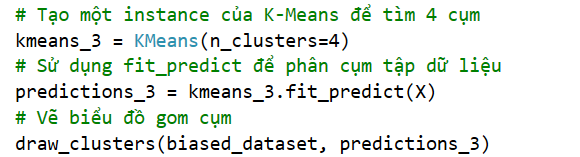
Những người thích khoa học viễn tưởng và lãng mạn thuộc nhóm màu vàng.

Những người thích khoa học viễn tưởng nhưng không thích lãng mạn thuộc nhóm màu xanh .

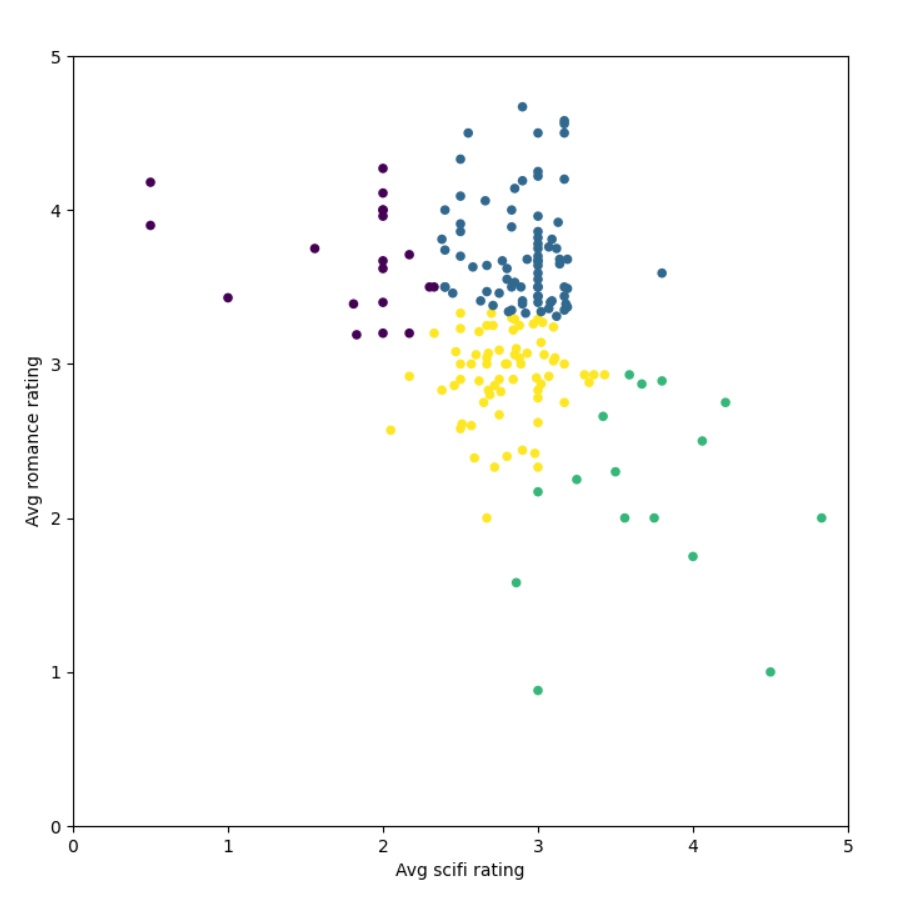
Những người thích lãng mạn nhưng không thích khoa học viễn tưởng thuộc nhóm màu tím.

+Tiếp tục thêm một nhóm khác:

Input:



Output:



Từ phân tích này,ta có thể nhận ra khi chia tập dữ liệu thành càng nhiều nhóm thì sở thích của những người thuộc mỗi nhóm càng giống nhau.

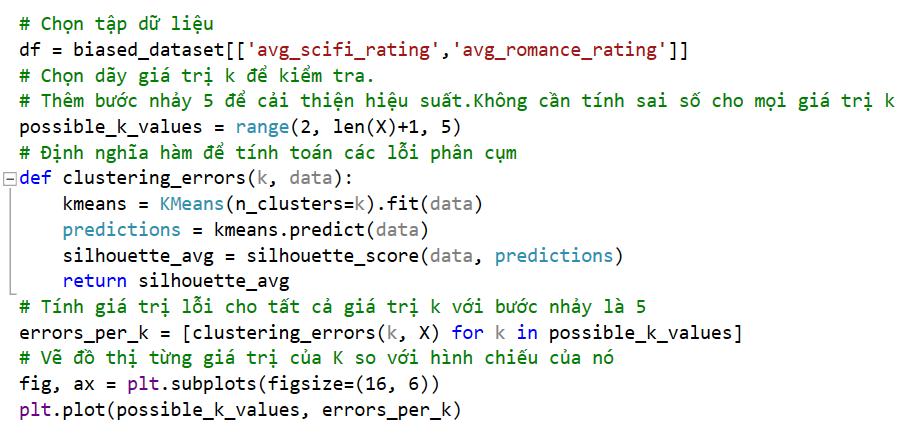
+ Lựa chọn đúng số cụm bằng phương pháp khuỷu tay (Elbow Method):

Elbow Method : được sử dụng để xác định số lượng chính xác các cụm trong tập dữ liệu. Nó hoạt động bằng cách vẽ biểu đồ các giá trị tăng dần của K so với tổng sai số thu được khi sử dụng K đó. Mục tiêu là tìm k mà đối với mỗi cụm sẽ không làm tăng đáng kể phương sai, để hiểu rõ hơn về phương pháp này, khi chúng ta nói về phương sai, chúng ta đang đề cập đến lỗi. Một trong những cách để tính toán sai số này là: Đầu tiên, trừ đi khoảng cách Euclid từ mỗi điểm của mỗi cụm đến tâm của nhóm tương ứng. Sau đó, bình phương giá trị này (để loại bỏ số âm). Và cuối cùng, cộng tất cả các giá trị đó để có được tổng sai số, bây giờ ta muốn tìm ra số lượng cụm phù hợp cho tập dữ liệu, ta sẽ thực hiện phương pháp elbow cho tất cả các giá trị có thể có của K sẽ nằm trong khoảng từ 1 đến tất cả các phần tử của tập dữ liệu.

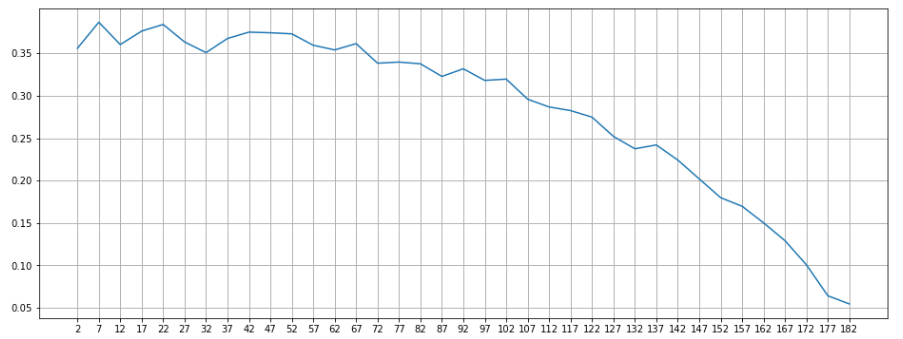
Nếu K = 1, chỉ tồn tại một nhóm cho tất cả các dữ liệu.

Nếu K = tất cả các điểm dữ liệu, mỗi điểm dữ liệu là một nhóm riêng biệt.

Input:



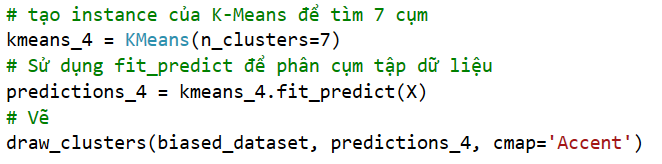
Output:



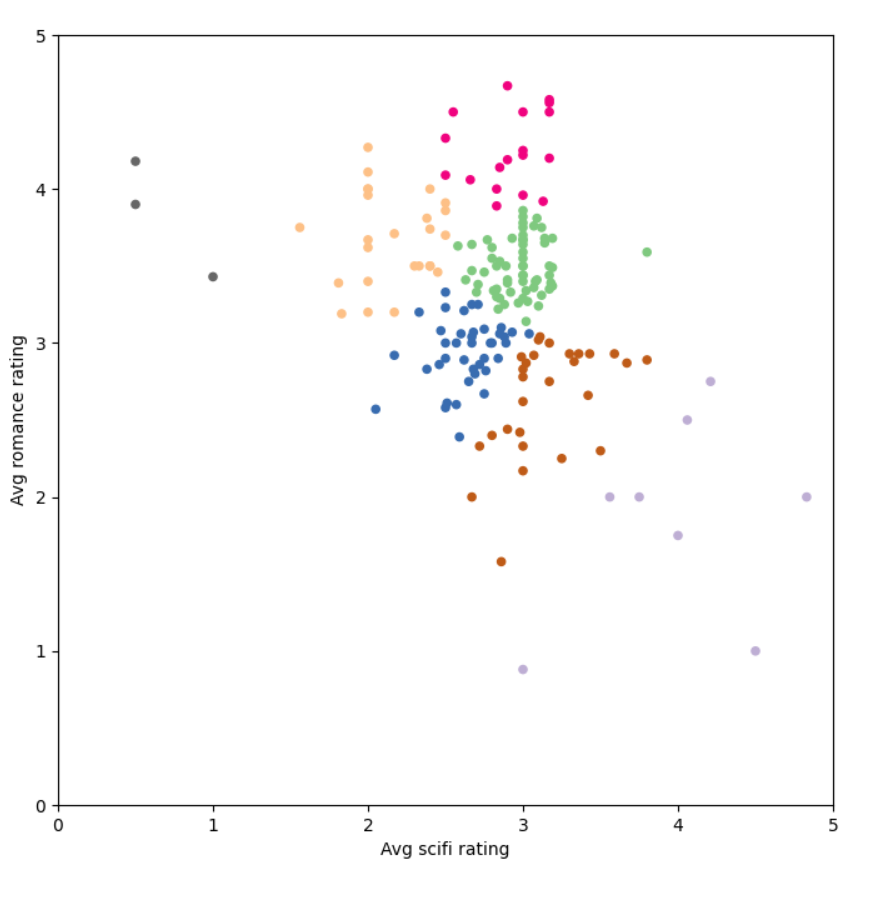
Dựa vào biểu đồ ta có thể thấy 7,22,32 là những lựa chọn K tốt nhất. Việc tăng số lượng các cụm vượt ra ngoài phạm vi đó dẫn đến các cụm xấu nhất theo hình chiếu của nó. Ta sẽ chọn K = 7 vì nó là điểm mang lại điểm tốt nhất và sẽ dễ hình dung hơn.

+Chọn k=7 đã tìm được để vẽ biểu đồ phân cụm thử:

Input:



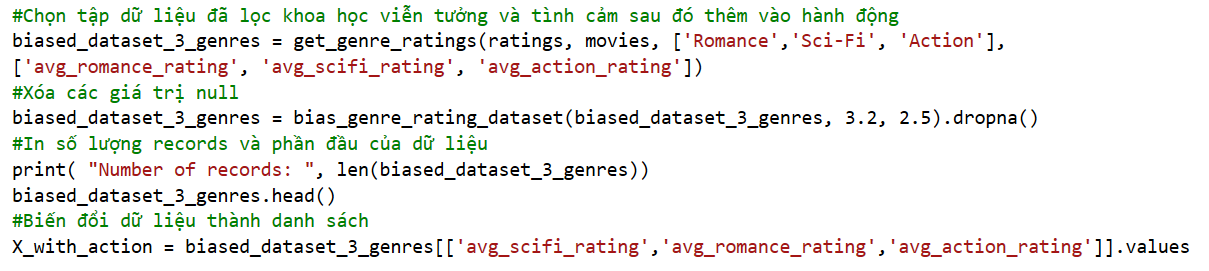
Output:



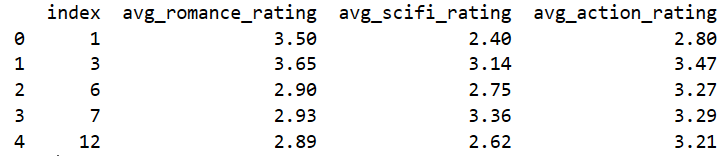
K đã được lựa chọn đúng nên biểu đồ phân bố cụm dễ dàng phân biệt. vì đang sử dụng màu của matplotlib “accent” nên không đủ màu để thử nghiệm các trường hợp k lớn hơn 10, các cụm khác nhau cùng màu sẽ rất khó phân biệt.

+Thêm thể loại phim hành động vào để phân tích:

Input:

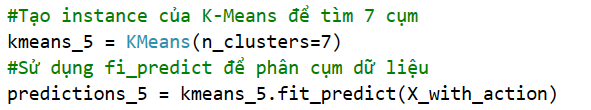


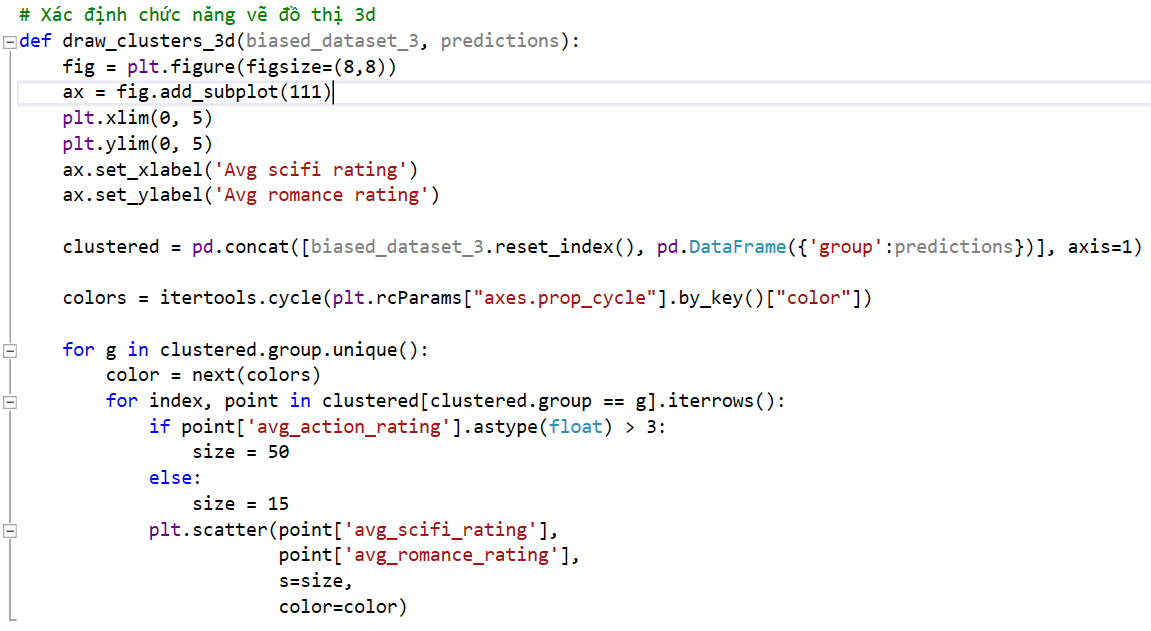
Output:



+Vẽ lại biểu đồ với k=7 sau khi đã thêm chiều dữ liệu vào bảng:

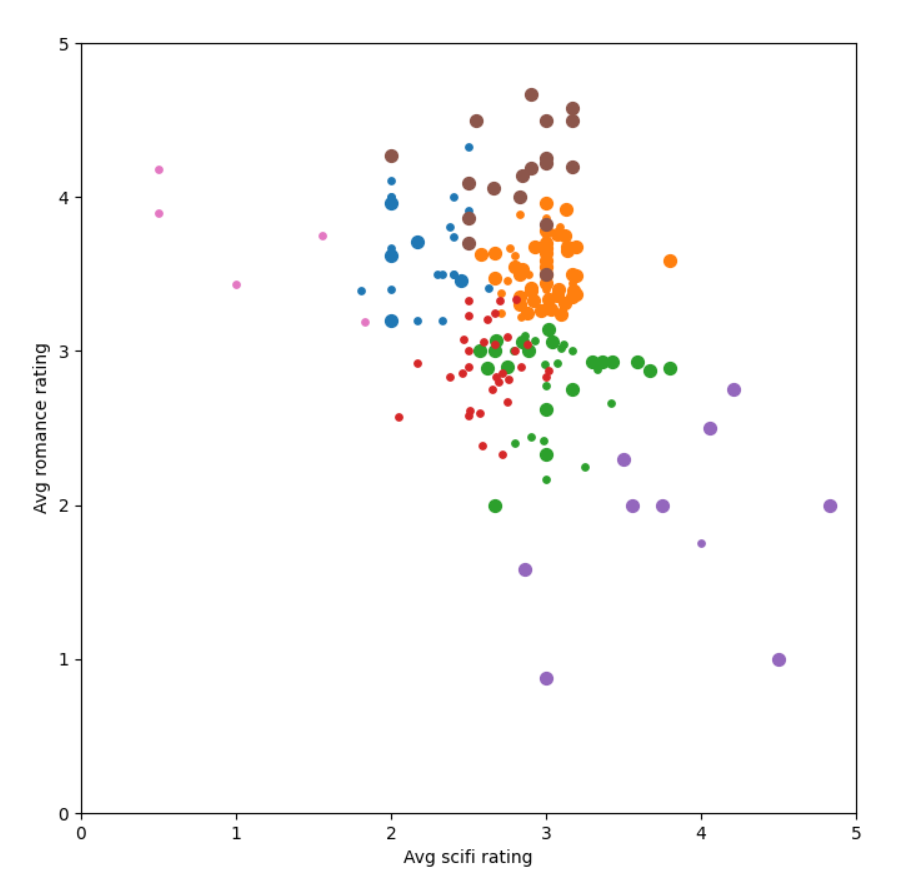
Input:







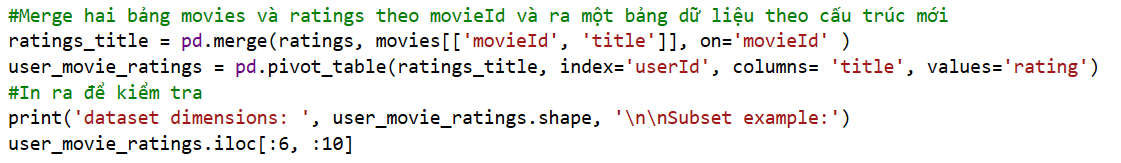
Output:



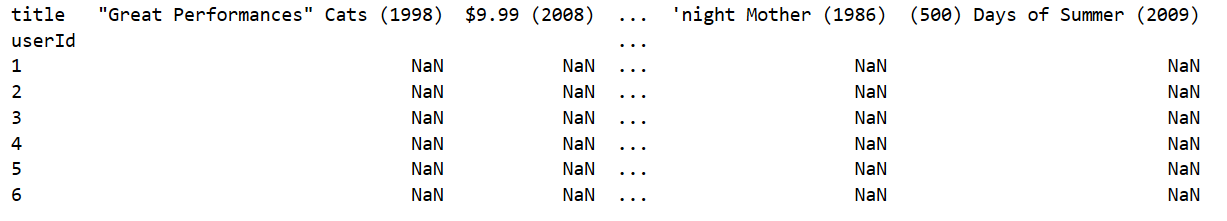
Như đã thấy sau khi vẽ bằng trục x,y để thể hiện ratings của phim tình cảm và khoa học giả tưởng,biến đổi kích thước để thể hiện ratings của phim hành động (kích thước càng lớn, độ ratings càng cao) khiến chúng ta khó hình dung một cách chính xác ở chiều dữ liệu lớn hơn 3.

+Vẽ biểu đồ dạng khác để hình dung chính xác cụm dữ liệu tối đa 5 chiều

Input:

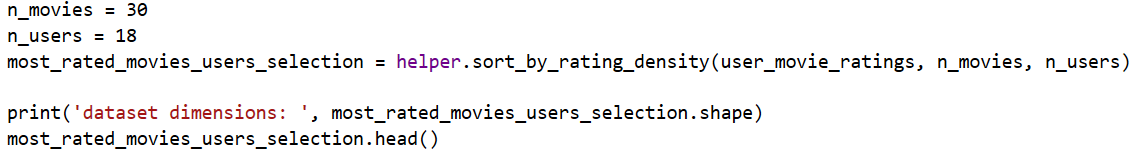


Output:

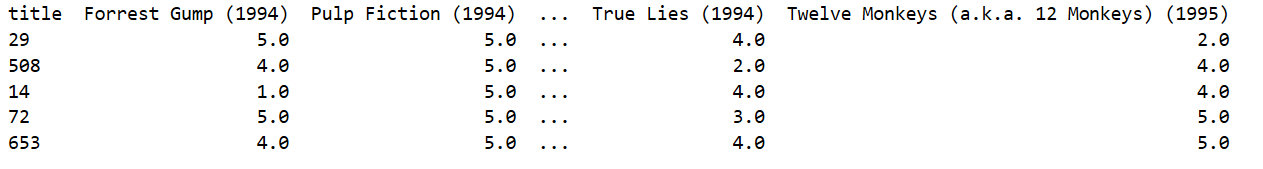


Tập dữ liệu chứa nhiều NaN do người xem không đánh giá phần lớn các bộ phim. Để khắc phục ta sẽ phân loại dataset theo phim được đánh giá cao nhất và người xem đã đánh giá số lượng phim nhiều nhất:

Input:



Output:

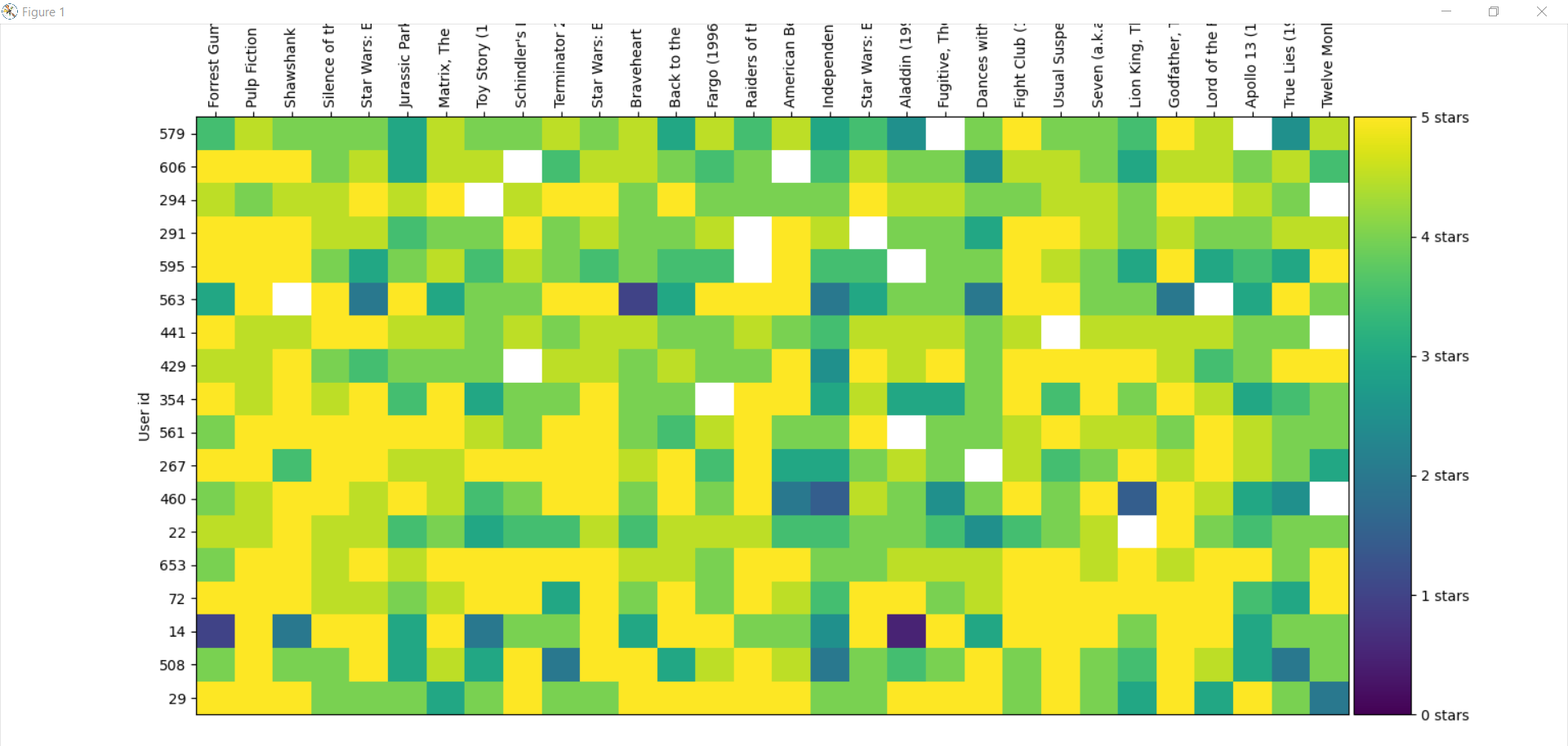


Dữ liệu đã được lọc theo ý muốn,để trực quan hơn cho ta có thể thấy rõ các xếp hạng (và sau đó là các cụm) khi ta xét các tập hợp con lớn hơn, ta sẽ thể hiện dữ liệu thông qua màu sắc thay vì số:

Input:



Output:



Qua biểu đồ ta đã nhận thấy rõ hơn: mỗi cột tượng trưng cho mỗi phim, mỗi hàng là một người xem, mỗi ô là xếp hạng của người xem đối với phim đó ứng với tỉ lệ bên phải. một số ô màu trắng do người xem tương ứng đã không đánh giá bộ phim đó. Đây cũng là một vấn đề khi phân cụm trong thực tế, tập dữ liệu thực tế thường thưa thớt hoặc không có giá trị ở mỗi ô của tập dữ liệu. Điều này làm cho việc phân nhóm người xem trực tuyến theo rating của họ trở nên khó khăn hơn vì K-Means không phù hợp dùng khi các giá trị bị thiếu, điều này chứng tỏ dữ liệu càng lớn thì hiệu suất càng kém.

+Để tăng hiệu suất ta chỉ xét 1000 bộ phim thay vì hơn 9000 bộ phim trong tập dữ liệu

Input:

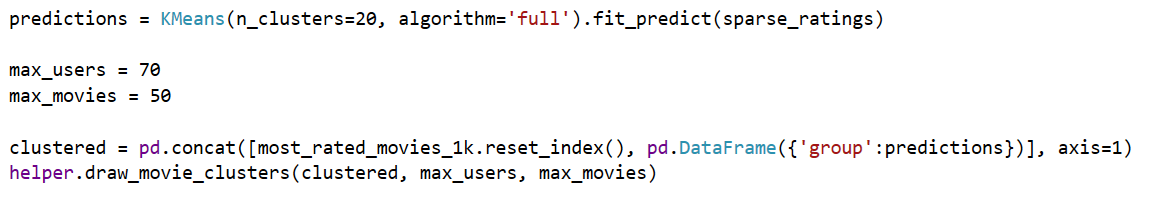


Để sklearn chạy k-mean clustering thành một tập dữ liệu có các giá trị bị thiếu như thế này, trước tiên chúng ta sẽ truyền nó sang kiểu sparse csr matrix được xác định trong thư viện SciPi.



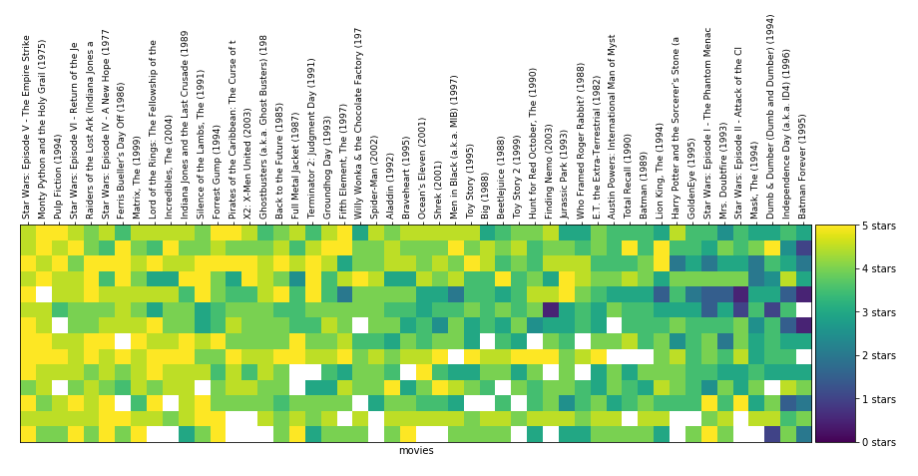
Với K-means, chúng ta phải xác định K: số lượng các cụm. Chọn tùy ý k = 20 (Tốt hơn là dùng Elbow Method nhưng sẽ mất khá nhiều thời gian cho dữ liệu lớn):

Để hình dung một số cụm này, ta sẽ vẽ từng cụm dưới dạng bản đồ nhiệt:

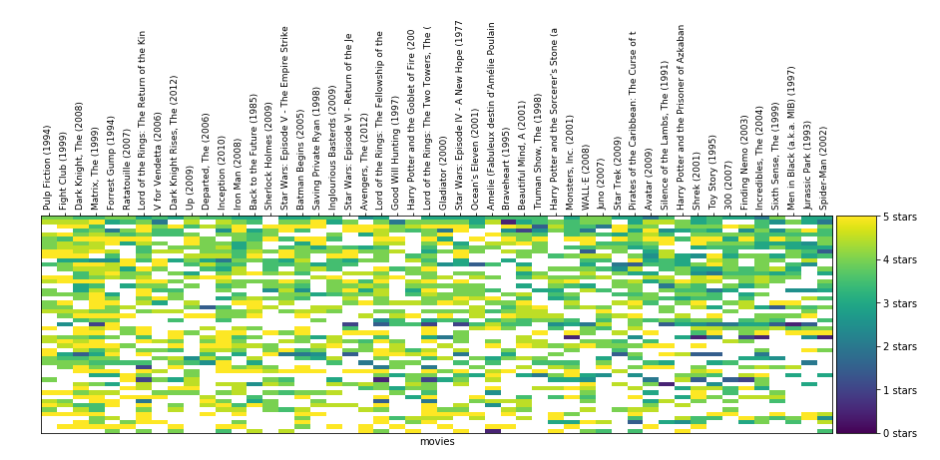


Output

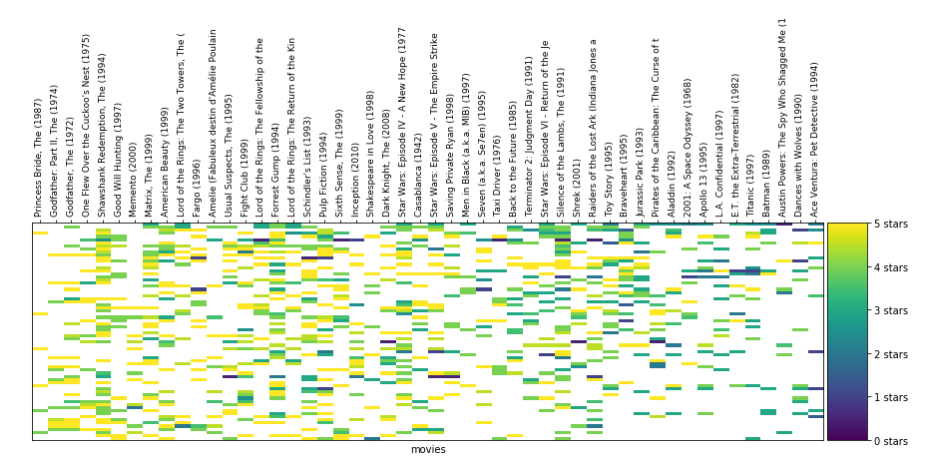
Cụm #1:



Cụm 11:



Cụm 18:



Từ các cụm ta thấy:

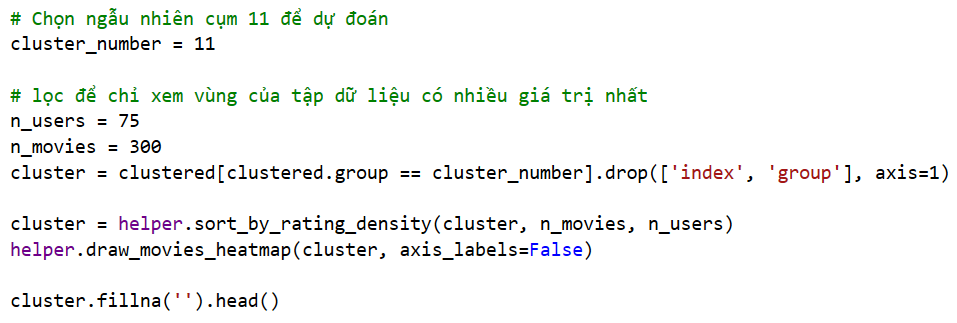
Các xếp hạng trong một cụm càng giống nhau, thì càng có thể tìm thấy nhiều đường thẳng đứng có màu sắc tương tự trong cụm đó.

Một số cụm thưa thớt hơn các cụm khác, chứa những người có thể xem và xếp hạng phim ít hơn các cụm khác.

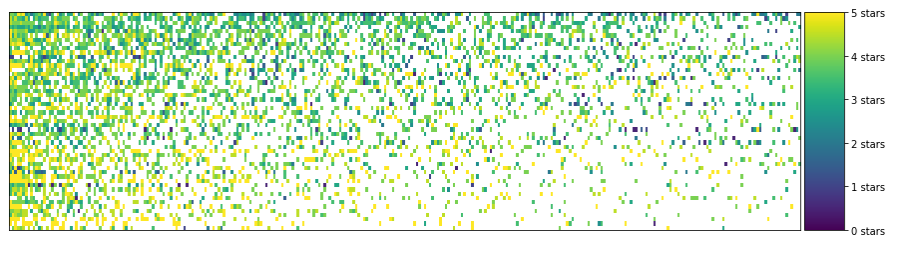
Một số cụm chủ yếu có màu vàng và tập hợp những người thực sự yêu thích một nhóm phim nhất định. Các cụm khác chủ yếu có màu xanh lá cây hoặc xanh nước biển có nghĩa là chúng chứa những người đánh giá phim 2-3 sao vì phim chưa hay lắm.

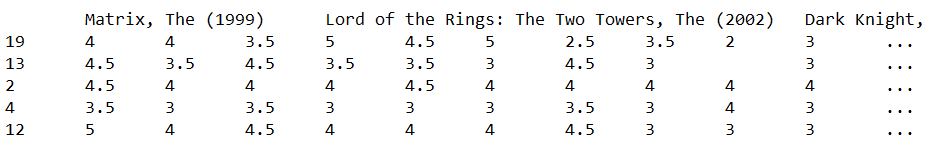
+Chọn một cụm và một người xem cụ thể để dự đoán:

Input:



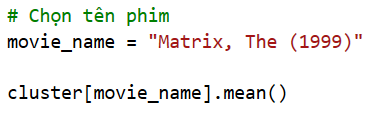
Output:





Chọn một bộ phim (có chứa ô trống). Nó trống vì người xem đó không đánh giá bộ phim đó. Chúng ta có thể đoán được liệu người đó có muốn xem bộ phim này hay không ? Vì người này nằm trong một nhóm người xem có vẻ giống nhau, ta có thể lấy trung bình số ratings cho bộ phim đó trong nhóm này và đó sẽ là một dự đoán hợp lý về mức độ người xem này sẽ thích bộ phim.

Input:



Output:



+Gợi ý phim cho người xem:

Ta đã sử dụng K-Means để phân cụm người xem theo xếp hạng của họ. Nhóm người xem có xếp hạng tương tự và do đó nhìn chung có gu xem phim tương tự. Dựa trên điều này, khi một người dùng không có xếp hạng cho một bộ phim nhất định, ta sẽ tính trung bình xếp hạng của tất cả những người dùng khác trong cụm và đó là dự đoán của ta xem người này có thích bộ phim này không.

Sử dụng logic này, Ta tính điểm trung bình trong cụm này cho mọi bộ phim, ta sẽ hiểu được cảm nhận của 'cụm vị giác' này về mỗi bộ phim trong tập dữ liệu.

Input:



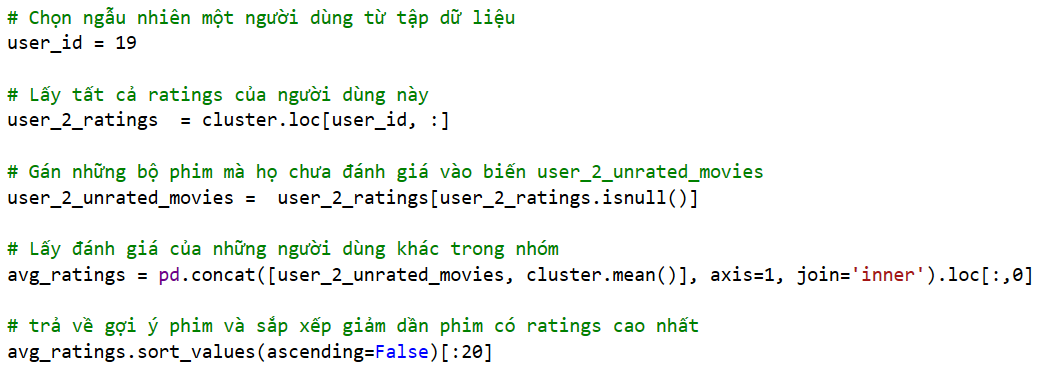
+Output: 

Ta có thể sử dụng như một công cụ đề xuất cho phép người dùng khám phá những bộ phim mà họ có thể sẽ thích.

Khi người dùng đăng nhập vào ứng dụng, ta có thể hiển thị cho họ các đề xuất phù hợp với sở thích của họ. Công thức cho các đề xuất này là chọn các phim được xếp hạng cao nhất của cụm mà người dùng chưa xếp hạng.

+Chọn một người dùng ngẫu nhiên để thực hiện:

Input:



Output:

