**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG CƠ SỞ TẠI TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN 2**

**🙡🕮🙣**



**ĐỒ ÁN MÔN PHÁT TRIỂN CÁC HỆ THỐNG THÔNG MINH**

Đề tài: Thiết kế Website bán sách

(đề xuất sách dựa trên lịch sử sản phẩm đã xem)

**Giảng viên phụ trách:** Thầy Nguyễn Ngọc Duy

**Lớp:** D21CQCNPM01-N

**Sinh viên thực hiện:**

1. Nguyễn Ngọc Thiên Phúc – N21DCCN066
2. Trần Thị Thùy Ngân – N21DCCN055
3. Trần Song Nguyên – N21DCCN058

TP. Hồ Chí Minh, ngày 16 tháng 10 năm 2024

# LỜI CÁM ƠN:

Trong thời đại số, nhu cầu đọc sách ngày càng tăng. Để đáp ứng nhu cầu đó, nhóm em đã nghiên cứu và phát triển một hệ thống khuyến nghị sách thông minh không chỉ dựa trên lịch sử duyệt sách mà còn kết hợp với các yếu tố khác như đánh giá của người dùng, thông tin về sách và sở thích cá nhân để đưa ra những gợi ý chính xác hơn, nhằm giúp người dùng dễ dàng khám phá những cuốn sách phù hợp với sở thích của mình. Chúng em đã cố gắng vận dụng những kiến thức đã học trên lớp cũng như trong thực tế để hoàn thành tốt đề tài và hứa hẹn sẽ mang đến trải nghiệm mua sách mới mẻ và thú vị hơn.

Chúng em xin chân thành cảm ơn Thầy Nguyễn Ngọc Duy đã tận tình chỉ bảo, giúp đỡ và giải đáp các vướng mắc để chúng em có thể hoàn thành đề tài một cách tốt đẹp. Do kiến thức còn hạn chế nên trong quá trình thực hiện đề tài nghiên cứu sẽ không thể tránh khỏi những sai sót, kính mong thầy chỉ dẫn, bỏ qua và giúp đỡ nhóm em. Chúng em rất mong nhận được sự đóng góp của thầy và các bạn để nội dung đề tài này ngày càng hoàn thiện hơn.

# MỤC LỤC:

[LỜI CÁM ƠN: 2](#_Toc180402200)

[Chương I: Giới thiệu đề tài: 4](#_Toc180402201)

[1. Tên đề tài: 4](#_Toc180402202)

[2. Lý do chọn đề tài: 4](#_Toc180402203)

[3. Mục tiêu nghiên cứu: 4](#_Toc180402204)

[4. Phương pháp nghiên cứu: 5](#_Toc180402205)

# Chương I: Giới thiệu đề tài

## Tên đề tài:

Thiết kế website bán sách tích hợp trí thông minh tự động gợi ý sách dựa trên lịch sử sách đã xem.

## Lý do chọn đề tài:

* **Nhu cầu thực tiễn**: Ngày nay, các trang web thương mại điện tử sử dụng các hệ thống gợi ý để cải thiện trải nghiệm mua sắm của người dùng. Đối với sách, việc gợi ý dựa trên lịch sử sách đã xem giúp người dùng dễ dàng khám phá thêm những cuốn sách mới phù hợp với sở thích cá nhân.
* **Cá nhân hóa trải nghiệm người dùng**: Hệ thống gợi ý dựa trên trí tuệ nhân tạo có thể phân tích hành vi người dùng, từ đó đưa ra các gợi ý sách phù hợp, tạo cảm giác cá nhân hóa và thu hút người dùng quay lại trang web thường xuyên hơn.
* **Tăng doanh số bán hàng**: Việc tự động gợi ý sách liên quan hoặc sách cùng chủ đề có thể khuyến khích người dùng mua thêm sách, từ đó tăng doanh thu cho cửa hàng.
* **Xu hướng phát triển công nghệ**: Trí tuệ nhân tạo và máy học (machine learning) đang trở thành những công nghệ hàng đầu trong nhiều lĩnh vực, đặc biệt là thương mại điện tử. Việc áp dụng AI vào một trang web bán sách không chỉ đáp ứng xu hướng mà còn tạo điều kiện cho bạn tìm hiểu và phát triển kỹ năng trong lĩnh vực này.
* **Tiện ích cho người dùng**: Người dùng không cần phải tìm kiếm quá nhiều mà có thể nhận được gợi ý chính xác về sách họ quan tâm dựa trên lịch sử xem, giúp tiết kiệm thời gian và công sức trong quá trình tìm kiếm.
* **Thị trường tiềm năng**: Sách luôn là một lĩnh vực có nhu cầu lớn, đặc biệt là với sự phát triển của các nền tảng học tập và tự học trực tuyến. Một website bán sách thông minh sẽ có tiềm năng phát triển rộng lớn.

## Mục tiêu nghiên cứu:

* Tìm hiểu hệ thống khuyến nghị sản phẩm phù hợp với mô hình kinh doanh sách trên website.
* Nghiên cứu về thuật toán K-Means và mô hình học máy không giám sát (unsupervised learning).
* Thu thập, tìm hiểu, phân tích tài liệu, thông tin liên quan đến đồ án.
* Ứng dụng của thuật toán K-Means vào việc phân các cụm các sản phẩm tương đồng phục vụ cho hệ thống khuyến nghị.
* Cài đặt thử nghiệm phương pháp và đánh giá kết quả.

## Phương pháp nghiên cứu:

**Phương pháp nghiên cứu lý thuyết:**

* **Nghiên cứu, thu thập thông tin về mô hình học máy không giám sát, thuật toán K-Means và hệ thống khuyến nghị sách trong thương mại điện tử:** Tìm hiểu cơ sở lý thuyết về các thuật toán học máy không giám sát, đặc biệt là K-Means, cùng với nguyên lý hoạt động của hệ thống khuyến nghị trong các nền tảng thương mại điện tử. Từ đó, đưa ra các phương pháp phù hợp để ứng dụng vào bài toán đề xuất sách.
* **Nghiên cứu ứng dụng K-Means vào giải quyết các bài toán phân cụm sản phẩm:** Tìm hiểu cách thức thuật toán K-Means được sử dụng trong việc phân cụm sản phẩm dựa trên các đặc điểm, nhằm giúp hệ thống gợi ý sách phù hợp hơn dựa trên lịch sử duyệt sách của người dùng.
* **Nghiên cứu ứng dụng deep learning vào giải quyết bài toán hệ thống khuyến nghị sản phẩm trong thương mại điện tử:** Khảo sát các mô hình deep learning, đặc biệt các mô hình học sâu như neural collaborative filtering (NCF) hoặc các kỹ thuật embedding, trong việc cải thiện hiệu quả hệ thống gợi ý thông minh trong thương mại điện tử.
* **Tham khảo ý kiến từ các cộng đồng uy tín:** Tham khảo các nghiên cứu, kinh nghiệm từ những cộng đồng chuyên ngành như Kaggle, Stack Overflow, hoặc các diễn đàn học thuật để cập nhật các xu hướng mới nhất trong lĩnh vực hệ thống khuyến nghị và thuật toán học máy.

**Phương pháp nghiên cứu thực nghiệm:**

* **Xây dựng một website bán sách đơn giản áp dụng mô hình thử nghiệm trên bộ dữ liệu tự xây dựng và đưa hệ thống khuyến nghị sản phẩm thông minh vào website:** Phát triển một trang web thử nghiệm với cơ sở dữ liệu sản phẩm tự xây dựng, trong đó ứng dụng mô hình K-Means và deep learning để gợi ý sản phẩm. Qua quá trình xây dựng, tiến hành đánh giá độ chính xác và hiệu quả của hệ thống khuyến nghị.
* **Dựa vào kết quả để phân tích, đánh giá hệ thống khuyến nghị thông minh:** Thu thập và phân tích dữ liệu từ quá trình thử nghiệm, đánh giá các tiêu chí như độ chính xác của đề xuất, sự hài lòng của người dùng, khả năng mở rộng và tính thực tế của hệ thống khi triển khai trong thương mại điện tử.

# Chương II: Cơ sở lý thuyết và công nghệ

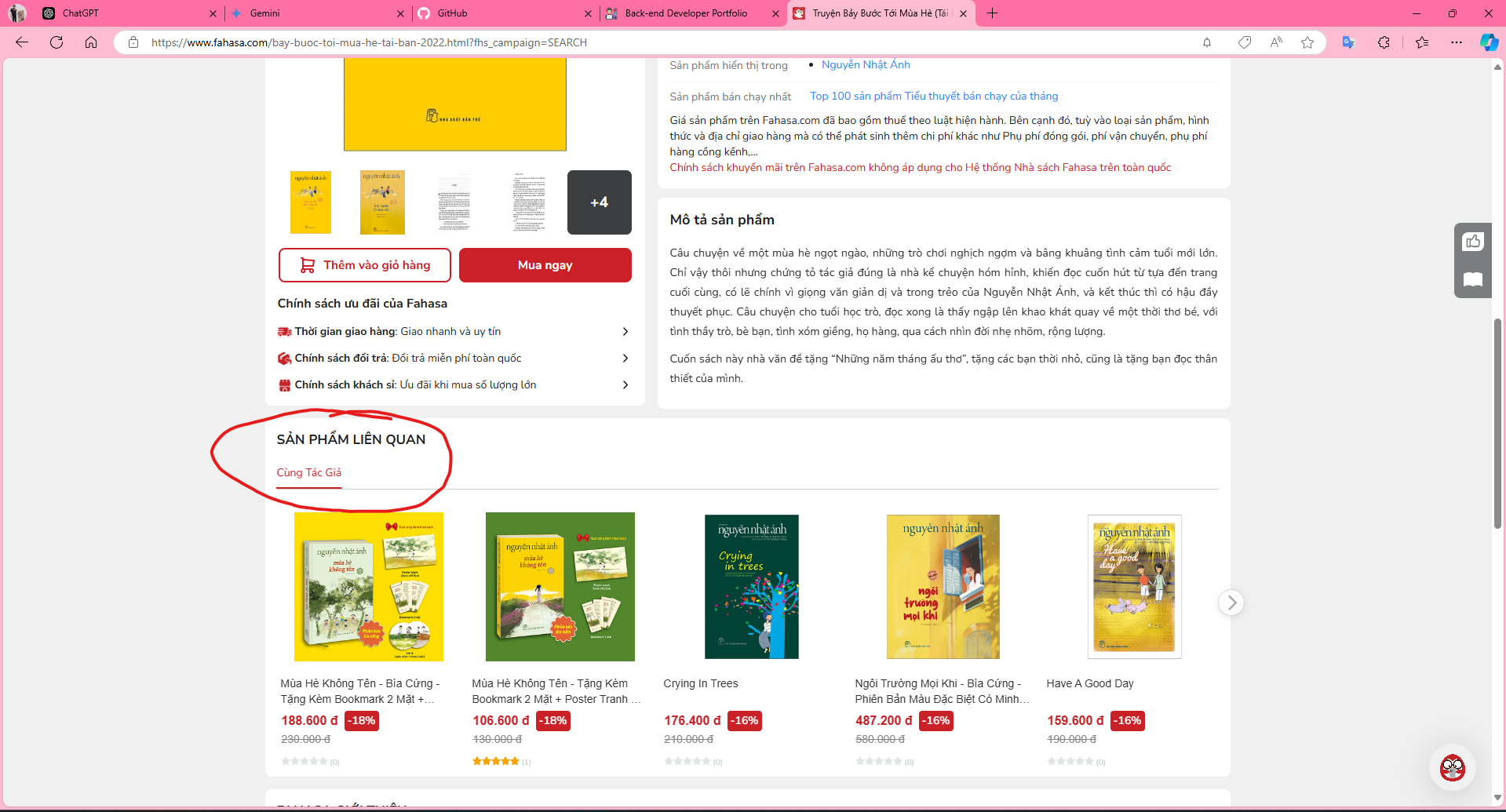
## Giới thiệu về học máy:

**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.

## Kỹ thuật xây dựng hệ thống khuyến nghị dựa trên kỹ thuật không cá nhân hóa:

Trong nhóm các kỹ thuật này, do tính đơn giản và dễ triển khai, chúng thường được tích hợp vào nhiều loại website và hệ thống khác nhau, bao gồm cả các trang thương mại điện tử, trang tin tức, hay trang giải trí. Ví dụ, trong các hệ thống bán hàng trực tuyến, thường có các gợi ý sản phẩm dựa trên các tiêu chí như: sản phẩm được xem/mua/bình luận nhiều nhất; sản phẩm mới nhất; sản phẩm cùng loại hoặc cùng nhà sản xuất; và các sản phẩm thường được mua kèm nhau. Một ví dụ điển hình là việc áp dụng thành công luật kết hợp (chẳng hạn như thuật toán Content-based Filtering) của Fahasa để xác định các sản sản phẩm cùng tác giả, như minh họa trong hình Hình 1.



Hình 1: Gợi ý sách cùng tác giả

Ta có thể thấy bất lợi của các phương pháp này là không cá nhân hóa cho từng người dùng, nghĩa là tất cả các người dùng đều được gợi ý giống nhau khi chọn cùng sản phẩm.

## Machine learning trong hệ thống khuyến nghị:

Machine learning là một trong những hướng nghiên cứu phát triển mạnh mẽ của lĩnh vực trí tuệ nhân tạo trong những năm gần đây. Các mô hình Machine learning đã mang lại những đột phá quan trọng trong việc giải quyết các bài toán phức tạp. Machine learning dựa trên một số ý tưởng từ hoạt động của não bộ, với khả năng tiếp thu nhiều tầng biểu diễn dữ liệu, từ cụ thể đến trừu tượng, giúp làm rõ ý nghĩa của các loại dữ liệu khác nhau. Trong bối cảnh thương mại điện tử phát triển nhanh chóng, nguồn dữ liệu khổng lồ được tạo ra gây ra thách thức lớn trong việc phân tích và xử lý dữ liệu để xây dựng các hệ thống khuyến nghị sản phẩm. Do đó, việc ứng dụng Machine learning để phát triển hệ thống khuyến nghị sản phẩm trong thương mại điện tử là một hướng nghiên cứu hợp lý và cần thiết ở thời điểm hiện tại.

## Thuật toán K-Means Clustering:

### Giới thiệu:

K-Means Clustering là một thuật toán học máy không giám sát (unsupervised learning) phổ biến, được sử dụng để phân nhóm dữ liệu thành các cụm (clusters) dựa trên đặc điểm tương đồng. Thuật toán hoạt động bằng cách chia tập dữ liệu thành K cụm khác nhau, với mỗi cụm được đại diện bởi một trung tâm cụm (centroid). Quá trình này lặp đi lặp lại cho đến khi các điểm dữ liệu trong cùng một cụm có khoảng cách nhỏ nhất đến trung tâm cụm của nó, trong khi khoảng cách giữa các cụm khác nhau là lớn nhất. K-Means thường được ứng dụng trong phân tích dữ liệu, nhận dạng mẫu, và các bài toán phân nhóm như phân loại khách hàng hoặc sản phẩm.

### Ý tưởng của bài toán K-Means Clustering:

Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm *ở gần nhau trong một không gian nào đó* (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình 2 bên dưới là 1 ví dụ về 3 cụm dữ liệu (gọi tắt là *cluster)*

Chart, scatter chart

Description automatically generated

Hình 2: Bài toán với 3 clusters

Giả sử mỗi cluster có một điểm đại diện (center) hinh tròn. Và những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó. Một cách đơn giản nhất, xét một điểm bất kỳ, ta xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó.

**Tóm tắt thuật toán:**

**Input:** Dữ liệu **X** và số lượng cluster cần tìm **K**.

**Output:**Các center **M** và label vector cho từng điểm dữ liệu **Y**.

**Các bước thực hiện:**

1. Chọn K điểm bất kỳ làm các center ban đầu.
2. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
5. Quay lại bước 2.

Chúng ta có thể đảm bảo rằng thuật toán sẽ dừng lại sau một số hữu hạn vòng lặp. Thật vậy, vì hàm mất mát là một số dương và sau mỗi bước 2 hoặc 3, giá trị của hàm mất mát bị giảm đi. Theo kiến thức về dãy số trong chương trình cấp 3: *nếu một dãy số giảm và bị chặn dưới thì nó hội tụ!* Hơn nữa, số lượng cách phân nhóm cho toàn bộ dữ liệu là hữu hạn nên đến một lúc nào đó, hàm mất mát sẽ không thể thay đổi, và chúng ta có thể dừng thuật toán tại đây.

### Các hàm số cần thiết cho K-Means Clustering:

Viết các hàm:

1. kmeans\_init\_centers để khởi tạo các centers ban đầu.
2. kmeans\_asign\_labels để gán nhán mới cho các điểm khi biết các centers.
3. kmeans\_update\_centers để cập nhật các centers mới dữa trên dữ liệu vừa được gán nhãn.
4. has\_converged để kiểm tra điều kiện dừng của thuật toán.

**def** **kmeans\_init\_centers**(X, k):

*# randomly pick k rows of X as initial centers*

**return** X[np.random.choice(X.shape[0], k, replace**=**False)]

**def** **kmeans\_assign\_labels**(X, centers):

*# calculate pairwise distances btw data and centers*

D **=** cdist(X, centers)

*# return index of the closest center*

**return** np.argmin(D, axis **=** 1)

**def** **kmeans\_update\_centers**(X, labels, K):

centers **=** np.zeros((K, X.shape[1]))

**for** k **in** range(K):

*# collect all points assigned to the k-th cluster*

Xk **=** X[labels **==** k, :]

*# take average*

centers[k,:] **=** np.mean(Xk, axis **=** 0)

**return** centers

**def** **has\_converged**(centers, new\_centers):

*# return True if two sets of centers are the same*

**return** (set([tuple(a) **for** a **in** centers]) **==**

set([tuple(a) **for** a **in** new\_centers]))

Phần chính của K-Means Clustering:

**def** **kmeans**(X, K):

centers **=** [kmeans\_init\_centers(X, K)]

labels **=** []

it **=** 0

**while** True:

labels.append(kmeans\_assign\_labels(X, centers[**-**1]))

new\_centers **=** kmeans\_update\_centers(X, labels[**-**1], K)

**if** has\_converged(centers[**-**1], new\_centers):

**break**

centers.append(new\_centers)

it **+=** 1

**return** (centers, labels, it)

Áp dụng thuật toán vừa viết vào dữ liệu ban đầu, hiển thị kết quả cuối cùng.

(centers, labels, it) **=** kmeans(X, K)

**print**('Centers found by our algorithm:')

**print**(centers[**-**1])

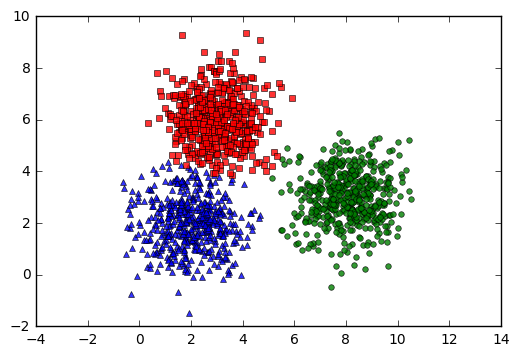
kmeans\_display(X, labels[**-**1])

Centers found by our algorithm:

[[ 1.97563391 2.01568065]

[ 8.03643517 3.02468432]

[ 2.99084705 6.04196062]]



Từ kết quả này chúng ta thấy rằng thuật toán K-Means Clustering làm việc khá thành công, các centers tìm được khá gần với kỳ vọng ban đầu. Các điểm thuộc cùng một cluster hầu như được phân vào cùng một cluster (trừ một số điểm màu đỏ ban đầu đã bị phân nhầm vào cluster màu xanh da trời, nhưng tỉ lệ là nhỏ và có thể chấp nhận được).

### Ưu và nhược điểm của thuật toán K-means:

**Ưu điểm:**

* **Dễ hiểu và dễ triển khai:** K-Means là một thuật toán tương đối đơn giản và dễ hiểu. Việc triển khai và lập trình cũng khá dễ dàng với các thư viện phổ biến như Scikit-learn trong Python.
* **Hiệu suất tính toán cao:** K-Means có khả năng chạy nhanh và xử lý tốt trên các tập dữ liệu lớn, nhờ vào tính toán nhanh chóng giữa các điểm dữ liệu và trung tâm cụm (centroids). Điều này khiến nó trở nên hiệu quả cho các bài toán yêu cầu thời gian xử lý nhanh.
* **Khả năng mở rộng:** K-Means dễ dàng mở rộng để làm việc với các tập dữ liệu lớn và có thể áp dụng với nhiều loại dữ liệu khác nhau, từ dữ liệu số đến dữ liệu dạng vector.
* **Tối ưu hóa tốt cho các bài toán phân cụm:** Khi dữ liệu có các cụm rõ ràng và được phân phối một cách đồng đều, K-Means có thể tìm được các cụm rất chính xác và giúp phân loại tốt.

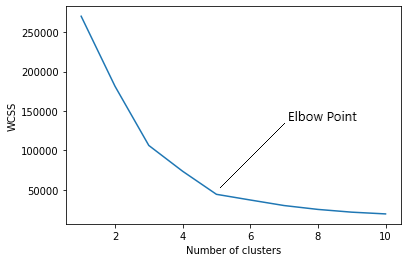
**Nhược điểm:**

* **Phụ thuộc vào số lượng cụm (K):** Một trong những điểm yếu chính của K-Means là bạn phải xác định trước số lượng cụm (K), trong khi trong nhiều bài toán thực tế, số lượng cụm không phải lúc nào cũng rõ ràng. Việc chọn sai số cụm có thể dẫn đến kết quả không chính xác.
* **Nhạy cảm với vị trí khởi tạo:** K-Means phụ thuộc vào việc khởi tạo ngẫu nhiên vị trí của các trung tâm cụm (centroids) ban đầu. Nếu vị trí khởi tạo không tốt, nó có thể dẫn đến các cụm không tối ưu hoặc kết quả không ổn định (các lần chạy khác nhau cho kết quả khác nhau).
* **Không phù hợp với dữ liệu có dạng cụm không cầu (non-spherical clusters):** K-Means hoạt động tốt khi các cụm có hình cầu (spherical clusters) và kích thước tương đối đồng đều. Tuy nhiên, nếu dữ liệu có dạng cụm phức tạp hoặc kích thước cụm khác nhau, K-Means có thể không phân nhóm chính xác.
* **Nhạy cảm với ngoại lệ (outliers):** K-Means rất nhạy cảm với các outliers (dữ liệu ngoại lệ), vì những điểm này có thể làm lệch trung tâm cụm và ảnh hưởng đến việc phân nhóm. Việc này có thể dẫn đến các cụm không chính xác.
* **Không đảm bảo tìm được giải pháp tối ưu:** K-Means có thể mắc kẹt trong minimum cục bộ (local minima), khiến kết quả không phải lúc nào cũng là giải pháp tối ưu toàn cục.

## Elbow trong việc phân cụm:

Trong phương pháp Elbow, thay đổi số lượng cụm (K) từ 1 – 10. Đối với mỗi giá trị của K, chúng tôi tính toán WCSS (Trung bình cộng khoảng cách trong cụm).

WCSS là tổng bình phương khoảng cách giữa mỗi điểm và trọng tâm trong một cụm. Sau đó vẽ đồ thị WCSS với giá trị K, đồ thị trông giống như một Khuỷu tay. Khi số cụm tăng lên, giá trị WCSS sẽ bắt đầu giảm .Giá trị WCSS lớn nhất khi K = 1. Khi phân tích đồ thị ta thấy đồ thị sẽ thay đổi nhanh chóng tại một điểm và do đó tạo ra hình dạng khuỷu tay. Từ thời điểm này, đồ thị bắt đầu di chuyển gần như song song với trục X. Giá trị K tương ứng với điểm này là giá trị K tối ưu hoặc số cụm tối ưu.



Hình 3: Đồ thị Elbow

## Đánh giá mức độ hiệu quả của hệ thống đề xuất sách thông minh:

**Ý tưởng:** Mỗi khi người dùng truy cập trang web sẽ lưu lại thời gian và ID của session. Khi người dùng click vào xem những sản phẩm ở danh mục được hệ thống khuyến nghị, lúc này những sản phẩm đó sẽ được lưu lại ở cơ sở dữ liệu nhằm phục vụ cho chức năng gợi ý sản phẩm, nâng cấp và phát triển hệ thống sau này.

**Công thức đánh giá:**

## Công nghệ sử dụng:

* **Front-end:** HTML5, CSS3, JS, Agular, Bootstrap
* **Back-end:** Java (Spring boot), Pythong (FastAPI)
* **Model:** Python
* **Database:** SQL Server

# Chương III: Thiết kế hệ thống

## 1.