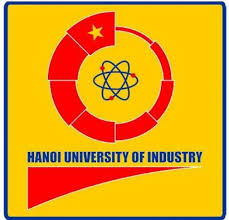
**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA: CÔNG NGHỆ THÔNG TIN**



**BÀI TẬP LỚN MÔN**

**TÍNH TOÁN SONG SONG VÀ PHÂN TÁN**

**ĐỀ TÀI: TÌM HIỂU THUẬT TOÁN SẮP XẾP SONG SONG( THUẬT TOÁN PHÂN CỤM K\_MEANS)**

**NHÓM: 10**

|  |  |
| --- | --- |
| **Giáo viên hướng dẫn:** | **Ths. Nguyễn Lan Anh** |
| **Sinh viên thực hiện:** | **Vũ Văn Hùng** |
|  | **Nguyễn Khắc Hiếu** |
|  | **Nguyễn Tùng Lâm** |
|  |  |

**Hà Nội, 10/2020**

# CHƯƠNG 1: TỔNG QUAN VỀ TÍNH TOÁN SONG SONG

1. **Tính toán song song**

Hiện nay, để giải quyết các bài toán lớn người ta thường nghĩ đến việc sử dụng các siêu máy tính hoặc việc kết hợp nhiều máy tính với nhau để tính toán. Tuy nhiên, với phương pháp lập trình cổ điển thì không thể nào phát triển được chương trình có thể tận dụng được sức mạnh của các hệ thống đó. Đó chính là lý do lập trình song song ra đời.

Lập trình song song là một công việc rất phức tạp so với lập trình tuần tự thông thường, người phát triển phải thực hiện một quá trình song song có khả năng tận dụng tối đa sức mạnh của hệ thống.

Tính toán song song là quá trình tính toán gồm những tiến trình được kích hoạt đồng thời và cùng tham gia tính toán giải quyết một công việc, và nói chung phải thực hiện trên các hệ thống đa bộ xử lý.

1. **Nguyên lý thiết kế tính toán song song**
   1. *Nguyên lý lập lịch*

Giảm tối thiểu các bộ xử lý sử dụng trong thuật toán sao cho thời gian tính taosn là không tang (xét theo khía cạnh độ phức tạp). Nghĩa là, nếu độ phức tạp tính toán của thuật toán là O(f(n)) thì thời gian thực hiện của chương trình có thể tăng khi số bộ xử lý giảm, và thời gian tính toán tổng thể tăng lên một hằng số nào đó – nhưng vẫn là O(f(n)).

* 1. *Nguyên lý hình ống*

Nguyên lý này được áp dụng khi bài toán xuất hiện một dây các thao tác {T1, T2,…,Tn}, tỏng đó Ti+1 thực hiện sau khi Ti kết thúc.

* 1. *Nguyên lý chia để trị*

Tức là chia bài toán thành những phần nhỏ hơn tương đối độc lập với nhau và giải quyết chúng một cách song song. Tạo ra một mô hình cây phân cấp để phân cấp quá trình truyền thông và tính toán.

+ Tăng tính song song so với mô hình trước

+ Thời gian chạy giảm từ O(n) xuống O(logn)

* 1. *Nguyên lý đồ thị phụ thuộc dữ liệu*

Phân tích mối quan hệ dữ liệu trong tính toán để xây dựng đồ thị phụ thuộc dữ liệu và dựa vào đó để xây dựng thuật toán song song.

* 1. *Nguyên lý điều khiển tương tranh*

Nếu hai tiến trình cùng muốn truy cập vào cùng một mục dữ liệu chia sẻ thì chúng phải tương tranh với nhau, nghĩa là chúng có thể cản trở lẫn nhau.

1. **Mục đích của tính toán song song**

* Nâng cao hiệu năng tính toán
* Tính toán các vấn đề khoa học rất phức tạp trong thế giới thực
* Khí quyển, trái đất, môi trường
* Vật lý ứng dụng, hạt nhân, hạt, áp suất cao, sự nóng chảy, lượng tử ánh sang
* Sinh học, công nghệ sinh học, di truyền học
* Hóa học, khoa học phân tử
* Địa chất, địa chấn học
* Công nghệ vũ trụ
* Kỹ thuật điện, thiết kế mạch, vi điện tử
* Khoa học máy tính, toán học

1. **Ưu điểm của tính toán song song**

* Tiết kiệm thời gian hoặc tài chính:
* Khi chi phí nhiều tài nguyên cho một công việc nào đó sẽ rút ngắn được thời gian hoàn thành công việc đó, và tiết kiệm các chi phí tiềm năng
* Phần cứng song song có thể được xây dựng với giá thành thấp và dễ mua
* Giải quyết được các vấn đề lớn hơn phức tạp hơn:
* Nhiều vấn đề quá lớn quá phức tạp đến nỗi nó không thực tế hoặc không có khả năng giải quyết chúng trên máy tính đơn, đặc biệt là bộ nhớ máy tính lại bị hạn chế. Chẳng hạn các vấn đề:
* Dự báo thời tiết, bão, động đất, sóng thần, mô hình sinh thái
* Máy dò tìm Web/CSDL để xử lý hàng triệu giao dịch trên một giây
* Cung cấp tính đồng thời:
* Tài nguyên của một máy tính đơn nguyên chỉ có khả năng thực hiện một lệnh tại một thời điểm
* Tài nguyên của máy tính đa bộ xử lý có thể thực hiện nhiều lệnh đồng thời
* Sử dụng tài nguyên phi cục bộ:

* Sử dụng tài nguyên tính toán trên một mạng diện rộng hoặc thậm chí internet khi tài nguyên nơi tính toán tài nguyên đang khan hiếm, không đủ

# CHƯƠNG 2: THUẬT TOÁN K-Means

1. **Tìm hiểu về thuật toán K-Means**

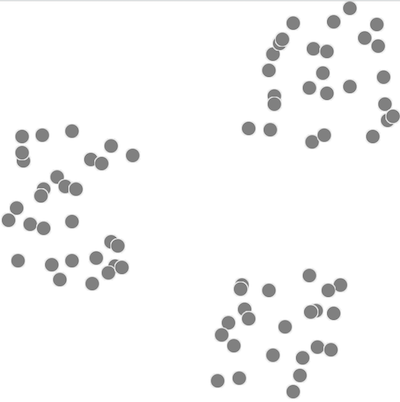
# Phương pháp k-means là gì

Ngắn gọn phương pháp k-means là 1 trong số những thuật toán phân 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.

Mục đích là phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

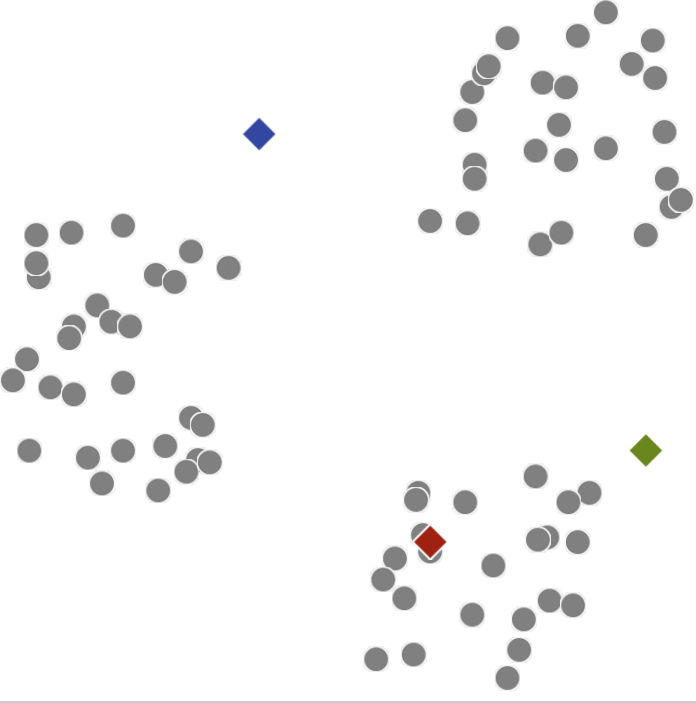
# Phân tích chi tiết

Đầu tiên là chuẩn bị dữ liệu cần phân cụm. Tiếp theo quyết định số lượng cụm (cluster) cần phân chia. Ở ví dụ này mình thử chọn số cluster là 3. Ở đây data được thể hiện dưới dạng các điểm cho dễ quan sát. Cự ly của các dữ liệu được hiểu là độ dài đoạn thẳng nối giữa 2 điểm với nhau.



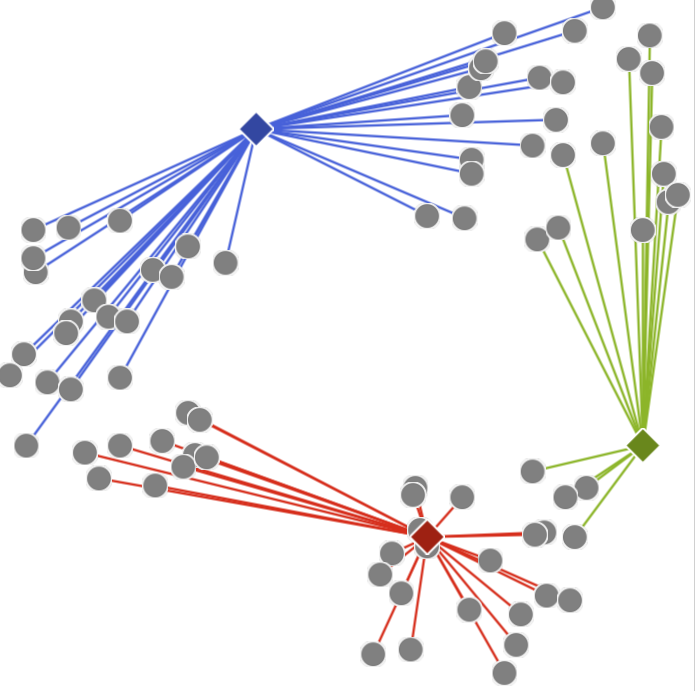
### **Bước 2：**

Chọn ngẫu nhiên 3 điểm làm điểm trung tâm của cluster.



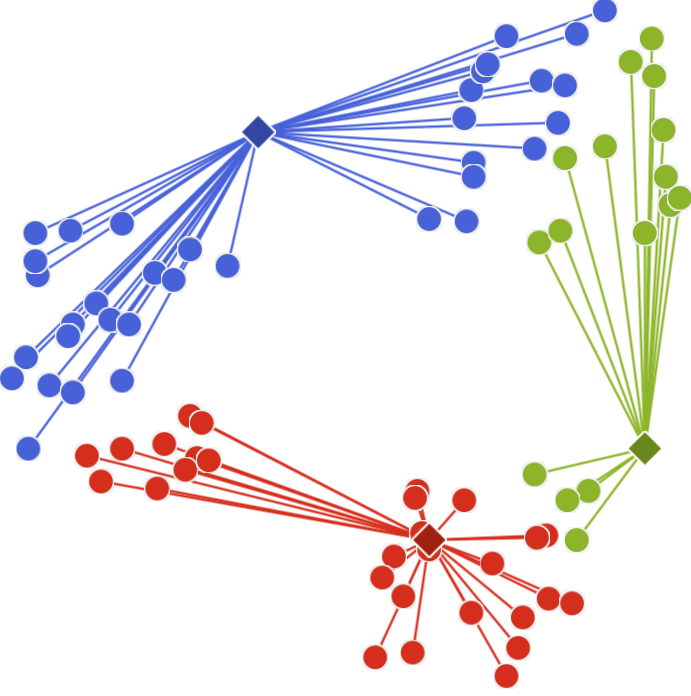
### **Bước 3:**

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.



### **Bước 4:**

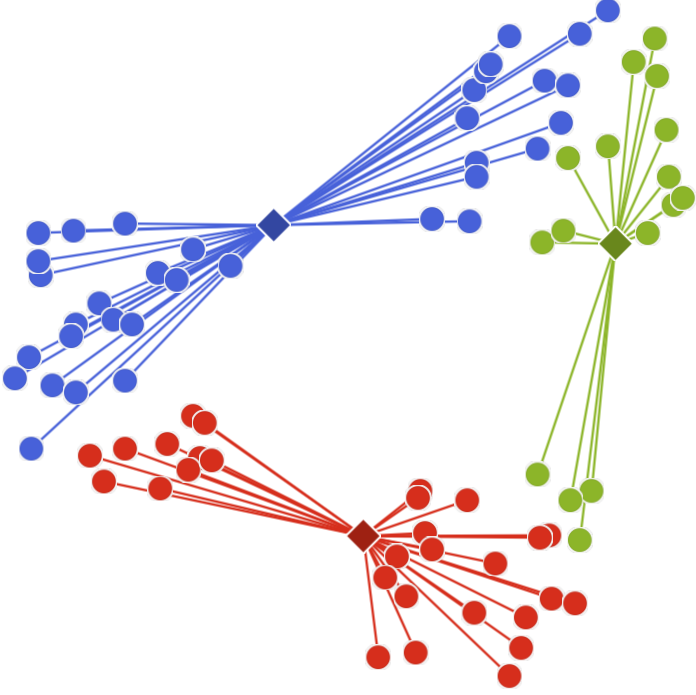
Từ bước tính toán trên, tiến hành phân loại các điểm về các cluster đã quyết định(cluster gần nó nhất). Vậy là đã phân ra được 3 cụm.



Quá dễ dàng nhỉ. Tuy nhiên nhìn vào hình trên bạn có thể nhận ra ngay các cụm dữ liệu này chưa phải là chuẩn nhất. Chính xác là do điểm trung tâm cụm được chọn chưa chính xác. Nên bước 5 sẽ được tiến hành.

### **Bước 5:**

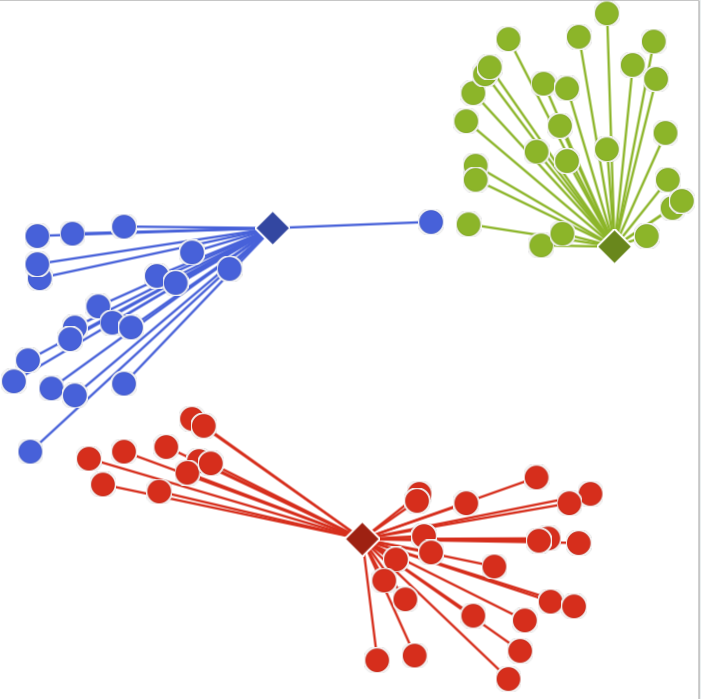
Bước trên chúng ta đã thu được 3 cụm, bây giờ tiến hành tính trọng tâm của các điểm dữ liệu của từng cụm. Sau đó di chuyển điểm trung tâm của cụm sang vị trí vừa tính được.



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.

### **Bước 6:**

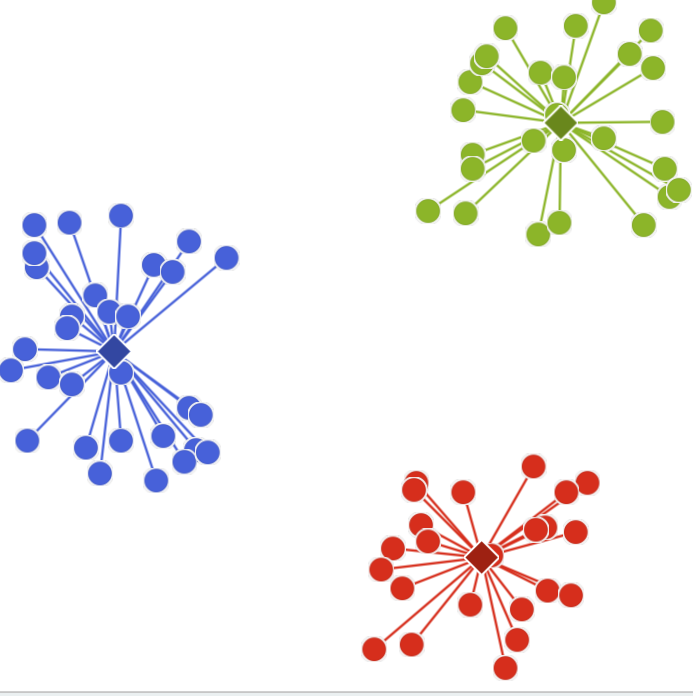
Một lần nữa tiến hành bước 3, tính toán lại khoảng các các điểm đến các điểm trung tâm. Sau đó phân loại lại các điểm dữ liệu về các cụm.



### **Bước 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.



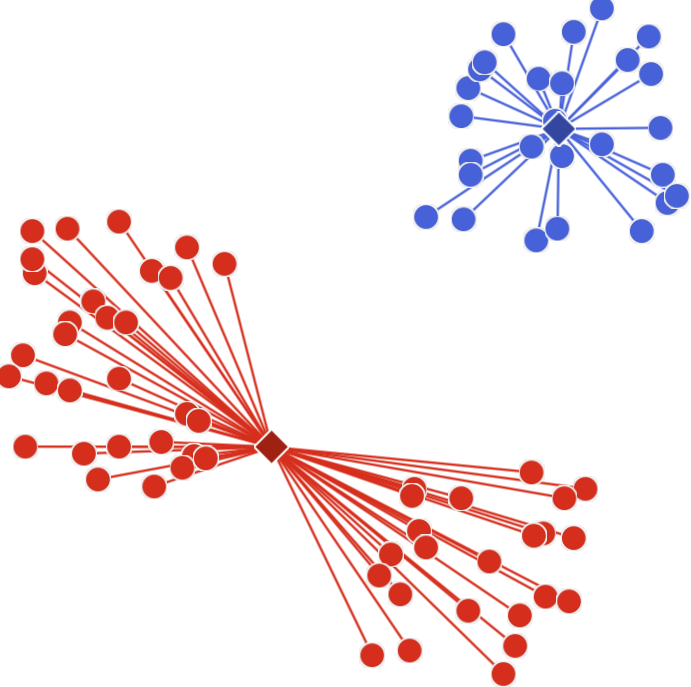
Với ví dụ này thì với với lần lặp thứ 4, thuật toán đã phân cụm thành công.

# Những lưu ý khi sử dụng phương pháp k-means

Trước khi sử dụng phương pháp này, chúng ta phải quyết định trước số lượng cluster, tuy nhiên trong quá trình tính toán số lượng cluster có thể khác với số lượng cluster mình dự đoán nên kết quả sẽ không chính xác.

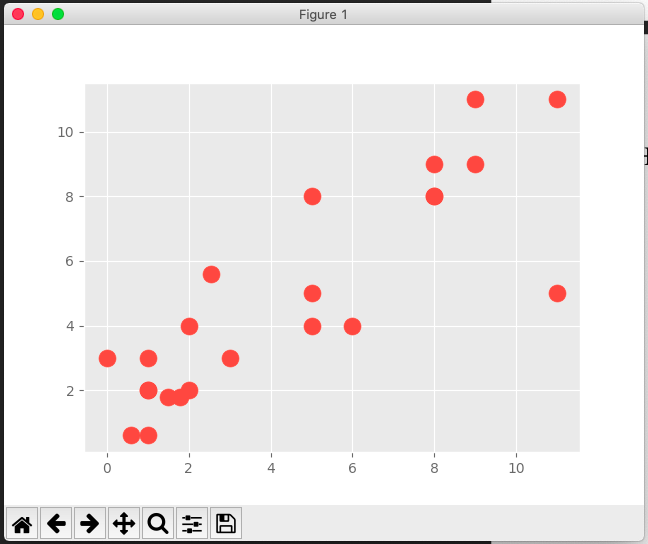
Vì vậy để giải quyết vấn đề này, để có thể chọn ra số lượng cluster thích hợp thì cần phải phân tích dữ liệu cẩn thận, chạy thử k-means với nhiều biến số số lượng cluster.

Cùng ví dụ trên nếu thay số lượng cluster thành 2, kết quả phân loại sẽ thành ra như sau:

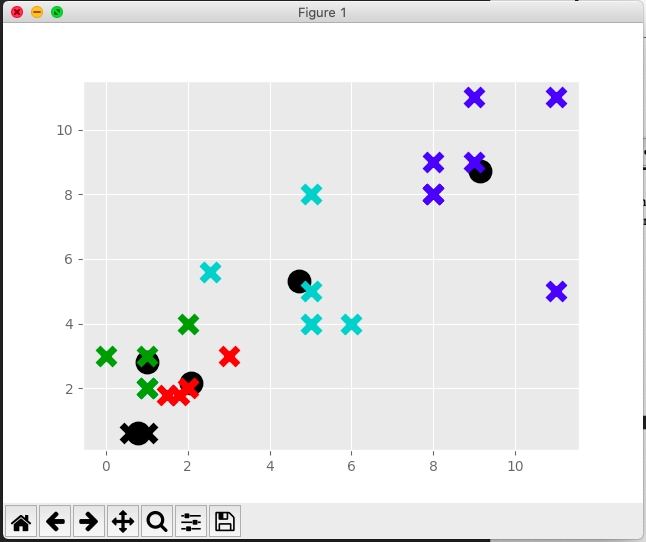


# CHƯƠNG 3: TÍNH TOÁN TUẦN TỰ THUẬT TOÁN K-Means

* Hiển thị tập data set lên đồ thị trước khi sử dụng K-Means để phân cụm



* Phân cụm dữ liệu sau khi sử dụng K-Means với số lượng cụm (Cluster) k=5 (phân tập dữ liệu thành 5 cụm)



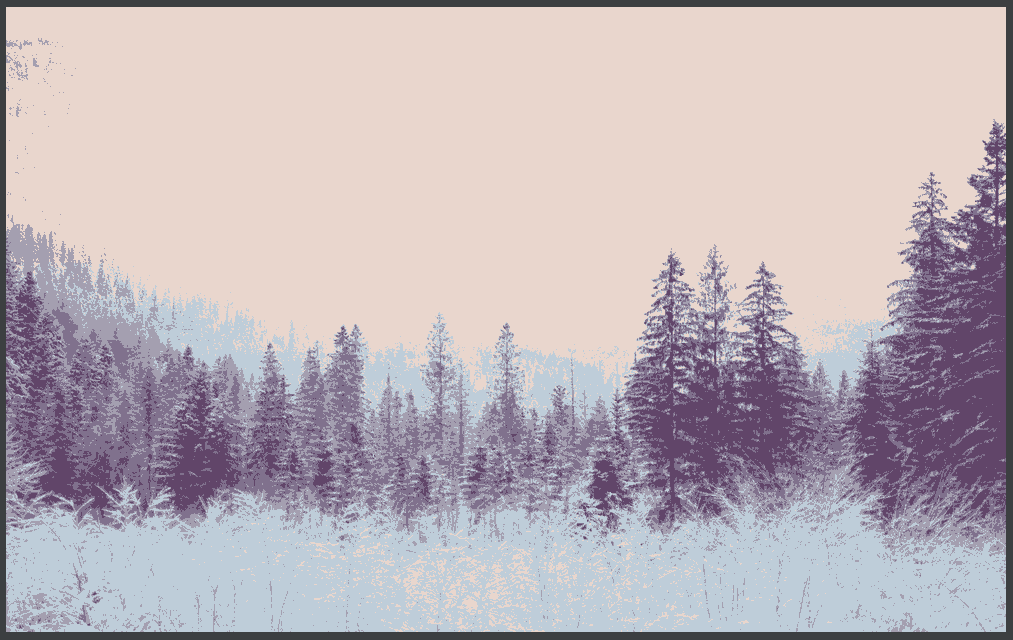
# CHƯƠNG 4: TÍNH TOÁN SONG SONG THUẬT TOÁN K-Means

* Ứng dụng thuật toán K-Means để phân đoạn ảnh (nén ảnh)

Ảnh ban đầu khi chưa nén:



Ảnh sau khi nén với số cụm Cluster (k = 5)



**Tài liệu tham khảo**

1. Giáo trình Xử lý song song và phân tán
2. https://github.com/ishanhan/parallel-implementation-of- kmeans?fbclid=IwAR0EazU-brzm3rkoDq5Iw\_mg3EdeHsJ-Ose8QezHbwuoin37QydcAgLsx6g
3. https://codetudau.com/de-dang-hieu-phuong-phap-k-means-qua-hinh-ve/index.html?fbclid=IwAR0tW\_2PjLpZabjnBc26QJylfxVQDbWWEezHuv\_FY31SzjWAHX61CjX5zKg