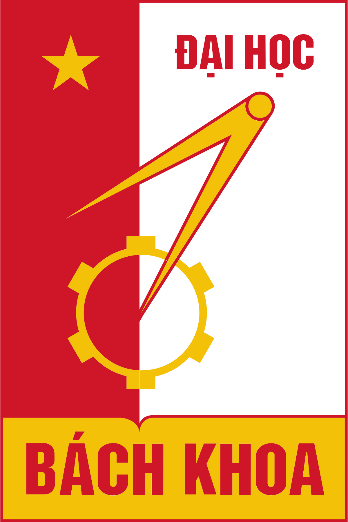
**ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỂN THÔNG**

****

**Báo cáo môn học: Nhập môn học máy và khai phá dữ liệu**

**Đề tài: Phân tích xu hướng quan tâm của người dùng về các video âm nhạc trên youtube năm 2021**

GVHD: PGS.TS. Thân Quang Khoát

Nhóm sinh viên:

|  |  |
| --- | --- |
| Họ và tên | MSSV |
| Phạm Thị Hương Quỳnh | 20183818 |
| Phạm Minh Hiệp | 20183738 |
| Phan Thị Lệ Hằng | 20183732 |
| Vũ Đình Hiếu | 20183745 |

Hà Nội, tháng 01 năm 2022

Mục lục

[1. Giới thiệu bài toán và ứng dụng 3](#_Toc92843475)

[a. Mô tả bài toán 3](#_Toc92843476)

[b. Ứng dụng 3](#_Toc92843477)

[2. Tổng quan dữ liệu 3](#_Toc92843478)

[a. Nguồn dữ liệu 3](#_Toc92843479)

[b. Mô tả dữ liệu 3](#_Toc92843480)

[c. Tiền xử lý dữ liệu: 4](#_Toc92843481)

[3. Giải quyết bài toán 5](#_Toc92843482)

[a. Phân tích dữ liệu: 5](#_Toc92843483)

[b. Mô hình và giải thuật áp dụng 10](#_Toc92843484)

[b.1. Các điểm lý thuyết chính 10](#_Toc92843485)

[c. Các thành phần chính trong mã nguồn 18](#_Toc92843486)

[4. Kết quả 20](#_Toc92843487)

[a. Các kết quả phân cụm 20](#_Toc92843488)

[b. Những suy luận rút ra được: 26](#_Toc92843489)

[5. Khó khăn và hướng phát triển tương lai 28](#_Toc92843490)

[a. Khó khăn: 28](#_Toc92843491)

[b. Hướng phát triển tương lai: 28](#_Toc92843492)

[6. Tài liệu tham khảo 29](#_Toc92843493)

# Phân công công việc

|  |  |
| --- | --- |
| **Họ và tên** | **Công việc** |
| Vũ Đình Hiếu | Phân tích các cụm dữ liệu, tìm hiểu thuật toán |
| Phạm Minh Hiệp | Tiền xử lý dữ liệu, phân tích dữ liệu sau khi tiền xử lý, tìm hiểu thuật toán |
| Phạm Thị Hương Quỳnh | Phân tích dữ liệu sau khi tiền xử lý, tìm hiểu thuật toán |
| Phan Thị Lệ Hằng | Crawl dữ liệu, tiền xử lý dữ liệu, tìm hiểu thuật toán |

# Giới thiệu bài toán và ứng dụng

## Mô tả bài toán

Phân tích xu hướng, quan tâm, sở thích của người dùng là một yếu tố quan trọng không chỉ đối với nhiều ngành kinh tế mà còn có ý nghĩa trong nghiên cứu xã hội. Bài toán phân tích xu hướng quan tâm đến các video âm nhạc của người dùng trên các nền tảng xã hội không nằm ngoại vòng quay ấy. Nếu như trước đây để phân tích sự yêu thích hay quan tâm của mọi người về sản phẩm, ta có thể dùng các phiếu câu hỏi khảo sát. Tuy nhiên, với số lượng lớn các video âm nhạc được phát hành trên các nền tảng số - cụ thể trong bài là Youtube, thì việc đưa ra các phiếu khảo sát là rất khó, đồng thời mỗi người dùng sẽ cảm thấy rất khó chịu nếu phải làm khảo sát quá nhiều. Vậy cần có một phương pháp khác có thể xử lý hiệu quả hơn việc này. Học máy cho ta một cách giải quyết đó là dùng các phương pháp phân cụm để tìm ra các nhóm video xu hướng mà mọi người đang quan tâm. Như vậy không cần phải các bài khảo sát nhàm chán, mà chỉ cần dựa trên những số liệu hay thông tin về các video được phổ biến ở một khoảng thời gian nào đó, thông qua phân cụm ta sẽ khám phá ra những xu hướng quan tâm của người dùng đồng thời cùng nhiều suy luận khác.

Bài toán sẽ có đầu vào sẽ chính là tập thông tin của các video phổ biến về âm nhạc được thu thập trên youtube trong phạm vi người dùng Việt Nam (hay các video âm nhạc được phát hành tại lãnh thổ Việt Nam). Đầu ra cần giải quyết là những suy luận, tìm kiếm mới về xu hướng quan tâm của người dùng về các video âm nhạc đó.

## Ứng dụng

Bài toán có thể ứng dụng trong một số nghiên cứu như:

* Phân tích tâm lý, xã hội của người dùng mạng xã hội
* Phân tích xu hướng, tìm ra những yếu tố mà mọi người quan tâm để phát triển các sản phẩm nghệ thuật, âm nhạc mới phù hợp với thị hiếu, …
* Phân tích xu hướng nghệ thuật mà mọi người quan tâm, từ đó đánh giá tiêu chuẩn thâm mỹ xã hội nói chung cũng như tiêu chuẩn âm nhạc nói riêng

# Tổng quan dữ liệu

## Nguồn dữ liệu

Dữ liệu bài toán được thu thập từ youtube với hai giai đoạn:  
 Giai đoạn 1: Sử dụng python selenium để thu thập đường dẫn của các video phổ biến trên youtube

Giai đoạn 2: Dựa trên các url lấy được của các video trên, sẽ tiến hành xử lý trích ra id của video và thông qua youtube api để lấy các thông tin tương ứng

## Mô tả dữ liệu

Dữ liệu thô sau khi thu thập về được lưu dưới dạng file csv

Số lượng bản ghi là 4962

Số cột ứng với số thuộc tính là 21

Với các thông tin cơ bản về các thuộc tính:

|  |  |
| --- | --- |
| *Thuộc tính* | *Ý nghĩa* |
| id | Mã id của video |
| publishAt | Thời gian đăng |
| Title | Tiêu đề video |
| description | Mô tả của video |
| channelTitle | Tiêu đề kênh đăng video |
| tags | Tag liên quan đến video |
| categoryId | Thể loại theo phân loại của youtube |
| defaultAudioLanguage | Ngôn ngữ của video |
| Duration | Thời lượng của video |
| dimension | Số chiều video (2D hay 3D) |
| definition | Định dạng của video |
| caption | Video có gắn dịch nghĩa ngôn ngữ |
| privacyStatus | Trạng thái bảo mật |
| embeddable | Cho phép nhúng video |
| madeoforKids | Video dành cho trẻ em |
| viewCount | Lượt xem của video |
| likeCount | Lượt yêu thích của video |
| commentCount | Lượt bình luận của video |
| dislikeCount | Lượt không thích của video |
| favouriteCount | ( youtube đã bỏ) |
| topicCategories | Chủ đề của video |

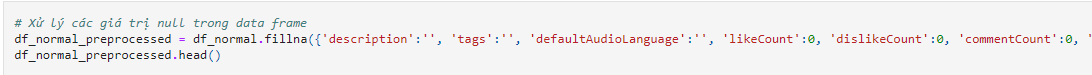
## Tiền xử lý dữ liệu:

Dữ liệu thu về cần tiền xử lý ở các điểm:

Lọc ra các video âm nhạc tức là categoryId là 10 – Music



Xử lý các giá trị null trong data frame



Xử lý topicCategory



Xử lý thời gian

Ảnh có chứa văn bản

Mô tả được tạo tự động

Xử lý các thuộc tính không nguyên, cụ thể là các thuộc tính dạng Boolean chuyển thành 0 và 1 để xử lý.

Ảnh có chứa văn bản

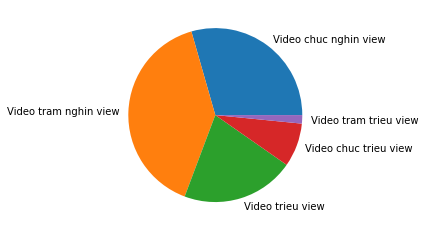
Mô tả được tạo tự động

# Giải quyết bài toán

## Phân tích dữ liệu:

Để phân tích xu hướng quan tâm của mọi người về các video âm nhạc thì một trong những tiêu chí dễ thấy nhất là thông qua số lượt xem của các sản phẩm này. Tuy nhiên mỗi video được youtube cung cấp thông tin có rất nhiều trường khác nhau. Vì vậy cần phân tích tương quan giữa các trường thông tin này để xác định đâu là yếu tố tác động đến lượt xem hay cho ra nhiều thông tin về xu hướng quan tâm của người dùng. Phần 3.1 sẽ tập trung phân tích thông qua các biểu đồ trực quan hóa dữ liệu để rút ra một số kết luận cơ bản để hỗ trợ giải quyết bài toán.

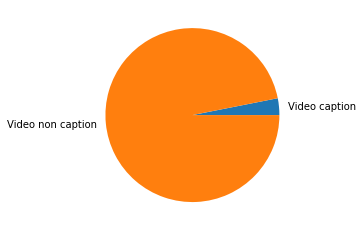
Dưới đây là biểu đồ cho thấy sự phân phối tỉ lệ các video với số lượt xem được phân loại theo các mức tương ứng.



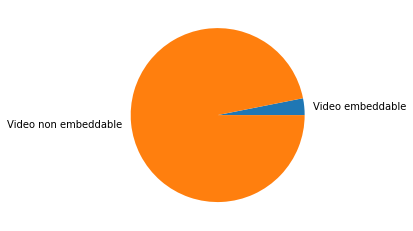
Như vậy qua các biểu đồ cho trong số tập dữ liệu, chiếm phần nhiều là các video ở mức trăm nghìn lượt xem, và chục nghìn lượt xem, còn lại thì càng ít các video có lượt xem cao ở mức hàng triệu lượt xem.

Tiếp tục xét tỉ lệ số các video âm nhạc với các thuộc tính khác:

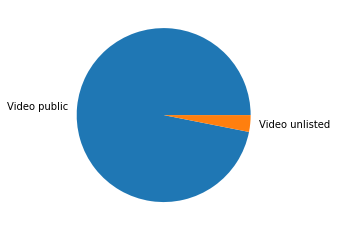
Tỉ lệ số video với thuộc tính caption:



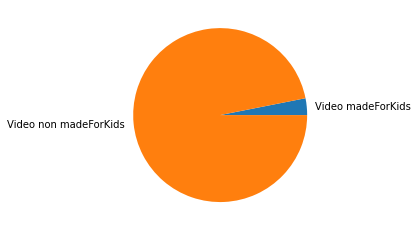
Tỉ lệ số video với thuộc tính embeddable:



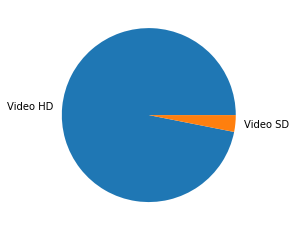
Tỉ lệ số video với thuộc tính privacyStatus:



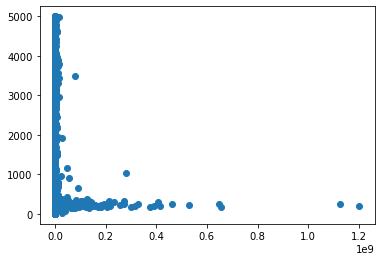
Tỉ lệ số video với thuộc tính madeForKids:

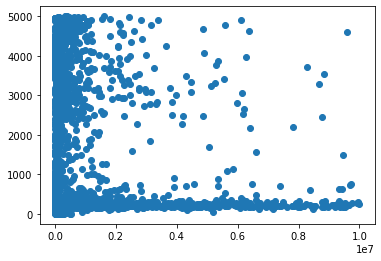


Tỉ lệ số video với thuộc tính definition:



Xét tương quan giữa lượt xem các video với thời lượng của video:





Như vậy qua phân tích sơ bộ quan hệ của các thuộc tính thông tin trong mỗi video ta rút ra một số kết luận sau:

* Các video mà mọi người thường xem có định dạng HD, không có caption, không tạo cho trẻ em, không nhúng được và là các video public
* Xu hướng lượt xem của mọi người ở các sản phẩm âm nhạc thường ở mức triệu lượt xem trở xuống, còn các video có lượt xem cao hơn thì sẽ ít. Đồng thời các video có lượt xem cao nhất thường tập trung ở vùng video có độ dài ngắn tức khoảng 400 giây hay tầm 5 phút. Đây cũng là thời lượng phổ biến của các sản phẩm ca nhạc, mv hiện nay. Còn vùng có thời lượng dài hơn, tập trung cũng nhiều video và thời lượng thường tập trung ở vùng triệu lượt xem trở xuống. Qua đó ta có một giả thiết rằng, người xem Việt Nam thường có xu hướng nghe lại các bài đã phát hành dưới dạng tuyển tập hay album vì phân bố cho thấy các video có thời lượng dài thường tập trung ở mảng hàng nghìn tới triệu lượt xem và đây là vùng lượt xem chủ yếu của các sản phẩm này.

Tuy nhiên nếu chỉ kết luận vậy thì không đủ để chúng ta phân tích nhiều hơn về thị trường âm nhạc, mà cần phải xem xét thêm cả thể loại nhạc của các video đó. Nhưng một vấn đề gặp phải là các video của youtube có một trường thông tin là topicCategories, có liệt kê một số chủ đề liên quan đến video nhưng thông tin đó chưa thực sự hữu ích để phân loại vì: phần lớn cho thấy ở các thuộc tính này (như caption, embeddable, madeForKid, … ) số lượng video nằm đa số hẳn về một bên nên chưa thể kết luận nhiều về xu hướng ở đây.

Nhưng, một điều đặc biệt ở các video âm nhạc trên youtube nói chung hay Việt Nam nói riêng, đó chính là chủ đề âm nhạc mà mọi người quan tâm còn nằm ở chính tiêu đề. Vì khi xem một video âm nhạc, thì tiêu đề sẽ là phần đầu tiên mà mọi người nhìn thấy, và tự họ có thể xác định nhanh chóng thể loại âm nhạc đó là gì mà không cần phải đào sâu xuống trường thuộc tính topicCategories. Vì vậy chúng ta cần phải phân cụm tiêu đề các video này để tìm ra các nhóm tiêu đề được quan tâm và phân tích các chúng để hiểu về thể loại cũng như sự quan tâm của mọi người đối với các sản phẩm này. Tuy nhiên việc phân cụm dữ liệu văn bản như tiêu đề với số lượng lớn như vậy thì không thể hoàn thành tốt bằng phương pháp thủ công thông thường được, mà cần có các công cụ của học máy. Cụ thể áp dụng ở bài toán này là phương pháp phân cụm K-means

## Mô hình và giải thuật áp dụng

Bài toán sử dụng mô hình phân cụm K-mean

### b.1. Các điểm lý thuyết chính

* Bài toán phân cụm

Đây là bài toán học không giám sát (unsupervised learning)

Có nhiều loại thuật toán phân cụm kể đến như: Kmeans, Agglomerative clustering, Spectral clustering, Gaussian mixtures,…

Nhóm tiến hành thử nghiệm bài toán với các thuật toán lần lượt là Kmeans, Agglomerative clustering, Spectral clustering, Gaussian mixtures và thu được một số kết quả gần tương tự nhau nên sẽ chỉ lựa chọn thuật toán Kmeans để nhận xét sau cùng. Dưới đây là hình ảnh kết quả của 3 phương trên, với kết quả của Kmeans sẽ được trình bày ở phần sau của báo cáo.

Ảnh kết quả chạy của thuật toán Spectral clustering (15 cụm)

|  |  |
| --- | --- |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động |  |

Ảnh kết quả chạy của thuật toán Agglomerative clustering (k = 13)

|  |  |
| --- | --- |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động |  |

Ảnh kết quả chạy của thuật toán Guassian mixtures (13 cluster):

|  |  |
| --- | --- |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động | Ảnh có chứa văn bản  Mô tả được tạo tự động |
| Ảnh có chứa văn bản  Mô tả được tạo tự động |  |

* Lý thuyết K-means

Phương pháp phân cụm K-means là một trong những phương pháp phổ biến nhất cho bài toán phân cụm dựa trên việc việc chia thành các phần. Trong phương pháp này, dữ liệu được biểu diễn dưới dạng tập các vector n chiều trong không gian Euclidean. Phương pháp K-means sẽ phân tập dữ liệu D ban đầu thành K cụm cluster:

Mỗi cụm có một điểm trung tâm là centroid.

K là một số cố định cho trước.

Các bước chính trong phương pháp này bao gồm:

Đầu vào: Tập dữ liệu D gồm các vector n chiều, số cụm K và phép đo khoảng cách d (x, y)

Bước khởi tạo: Chọn ngẫu nhiên K phần tử của D làm các centroid khởi tạo

Lặp lại các bước sau đến khi hội tụ:

Bước 1: Với mỗi phần tử, gán phần tử đó vào cụm có centroid gần mình nhất

Bước 2: Với mỗi cụm, tính toán và cập nhật lại centroid mới

Các bước của phương pháp sẽ hội tụ nếu thỏa mãn một trong 3 điều kiện sau:

Rất ít các phần tử thay đổi sang cụm khác

Các tâm cụm thay đổi không đáng kế

Tổng khoảng cách bình phương từ các phần tử đến tâm cụm tương ứng thay đổi không đáng kể

Như vậy ở thuật toán, có một số điểm cần lưu ý đó là:

* Số cụm cần chọn phù hợp để các cụm được phân ra đúng và hợp lý (Bài toán được nhóm sử dụng phương pháp Elbow để giải quyết vấn đề này)
* Các điểm centroid khởi tạo cần được lựa chọn tốt để phân cụm đạt được hiệu quả (Bài toán được nhóm sử dụng phương pháp K-means++ để giải quyết vấn đề này)
* Sử dụng phép đo khoảng cách giữa các phần tử phù hợp (Nhóm lựa chọn trong bài toán sử dụng phép đo khoảng cách Euclid)
* Phương pháp Elbow

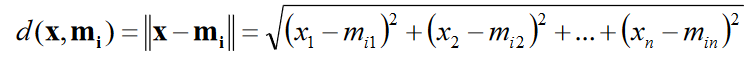
Phương pháp Elbow được dùng để xác định số cụm tối ưu cho phương pháp phân cụm K-means. Phương pháp hướng tới vẽ các giá trị hàm chi phí tạo bởi các giá trị k khác nhau. Ở đây ta đánh giá thông qua tổng khoảng cách bình phương từ các phần tử đến tâm cụm tương ứng. Khi k tăng, giá trị này giảm dần, nên đường elbow là một đường cong giảm dần. Vì điều kiện hội tụ của K-means là tổng khoảng cách bình phương này thay đổi không đáng kể nên dựa trên đường cong này ta sẽ xác định điểm gãy, tức độ dốc giảm đột ngột. Vị trí tương ứng đó là số k tối ưu.

* Phương pháp K-means++

Phương pháp Kmeans++, cho phép chúng ta khởi tạo các centroids đầu tiên với các bước như sau:

* Chọn ngẫu nhiên centroid thứ 1 là m1
* Chọn centroid thứ 2 sao cho centroid này xa m1 nhất
* …
* Chọn centroid thứ i sao cho centroid này xa tập {m1, m2, …, mi-1}
* ….
* Phép đo khoảng cách Euclid

Công thức tính khoảng cách Euclid:



* Xử lý dữ liệu cho K-means

Ở phần lý thuyết phía trên ta thấy: đầu vào của phương pháp K-means là tập các vector n chiều, mà tập đầu vào của bài toán thực ở đây là tập các tiêu đề, có kiểu dữ liệu là text. Vì vậy, dữ liệu cần được tiền xử lý về dạng các vector để làm đầu vào cho phương pháp K-means. Nhóm sử dụng phương pháp chuyển đổi dữ liệu text về dạng các vector IF-TDF.

* Vector IF-TDF

TF-IDF (Term Frequency – Inverse Document Frequency) là 1 kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. Một vài biến thể của tf-idf thường được sử dụng trong các hệ thống tìm kiếm như một công cụ chính để đánh giá và sắp xếp văn bản dựa vào truy vấn của người dùng. Tf-idf cũng được sử dụng để lọc những từ stopwords trong các bài toán như tóm tắt văn bản và phân loại văn bản.

TF: Term Frequency (Tần suất xuất hiện của từ) là số lần từ xuất hiện trong văn bản. Vì các văn bản có thể có độ dài ngắn khác nhau nên một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn là một văn bản ngắn. Như vậy, term frequency thường được chia cho độ dài văn bản( tổng số từ trong một văn bản).

Ảnh có chứa văn bản

Mô tả được tạo tự động

IDF: Inverse Document Frequency (Nghịch đảo tần suất của văn bản), giúp đánh giá tầm quan trọng của một từ. Khi tính toán TF, tất cả các từ được coi như có độ quan trọng bằng nhau. Nhưng một số từ như “is”, “of” và “that” thường xuất hiện rất nhiều lần nhưng độ quan trọng là không cao. Như thế chúng ta cần giảm độ quan trọng của những từ này xuống.

Ảnh có chứa văn bản

Mô tả được tạo tự động

Trong đó:

* idf (t, D): giá trị idf của từ t trong tập văn bản
* |D|: Tổng số văn bản trong tập D
* | {d ∈ D: t ∈ d} |: thể hiện số văn bản trong tập D có chứa từ t

Cơ số logarit trong công thức này không thay đổi giá trị idf của từ mà chỉ thu hẹp khoảng giá trị của từ đó. Vì thay đổi cơ số sẽ dẫn đến việc giá trị của các từ thay đổi bởi một số nhất định và tỷ lệ giữa các trọng lượng với nhau sẽ không thay đổi. (nói cách khác, thay đổi cơ số sẽ không ảnh hưởng đến tỷ lệ giữa các giá trị IDF). Việc sử dụng logarit nhằm giúp giá trị tf-idf của một từ nhỏ hơn, do chúng ta có công thức tính tf-idf của một từ trong 1 văn bản là tích của tf và idf của từ đó.

Cụ thể, chúng ta có công thức tính tf-idf hoàn chỉnh như sau: tfidf (t, d, D) = tf (t, d) x idf (t, D)

Khi đó:

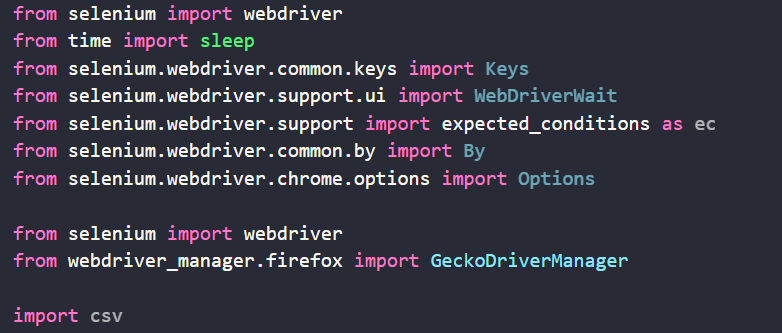
Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).

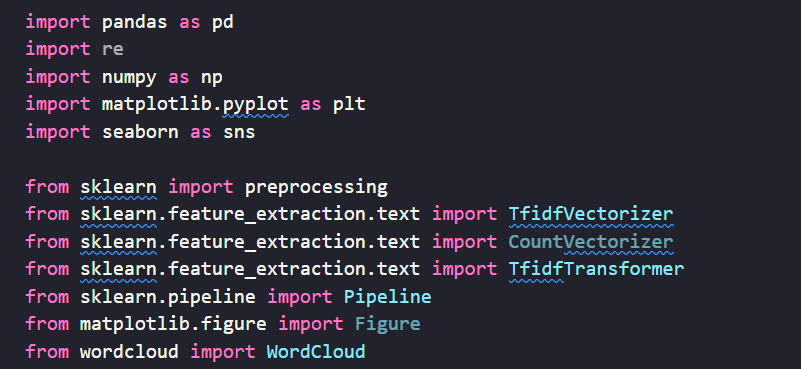
## Các thành phần chính trong mã nguồn

Crawl dữ liệu: các file: [scraping\_info\_data\_youtube.py](https://github.com/vudinhhieu20/Introduction-to-Machine-Learning/blob/main/scraping_info_data_youtube.py), [Analysis\_Youtube.ipynb](https://github.com/vudinhhieu20/Introduction-to-Machine-Learning/blob/main/Analysis_Youtube.ipynb)

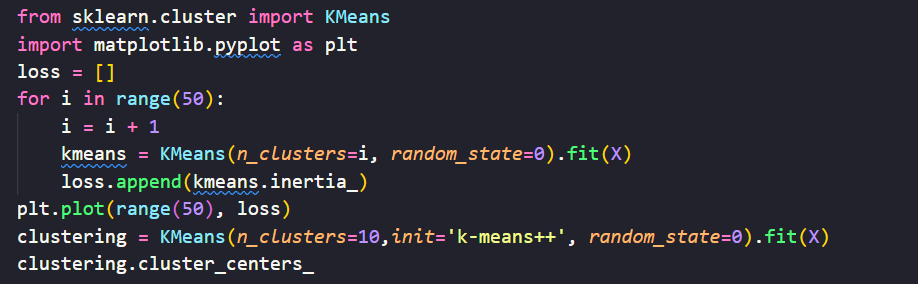
Tiền xử lý dữ liệu: ở file: [youtube\_music\_analysis.ipynb](https://github.com/vudinhhieu20/Introduction-to-Machine-Learning/blob/main/youtube_music_analysis.ipynb)

Các thư viện và gói cần thiết:

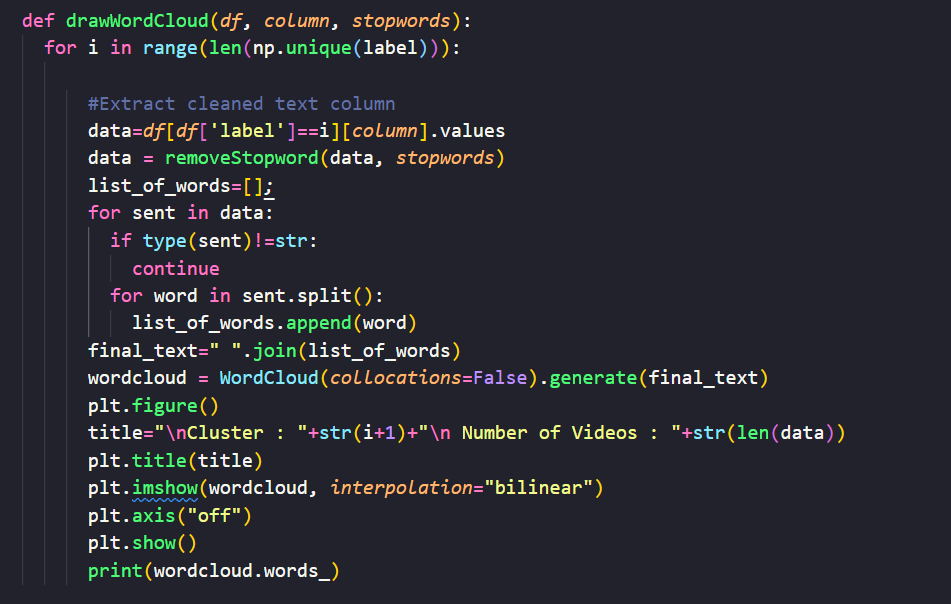




Thực hiện các phương pháp phân cụm: ở file: [youtube\_music\_analysis.ipynb](https://github.com/vudinhhieu20/Introduction-to-Machine-Learning/blob/main/youtube_music_analysis.ipynb)



Mô phỏng kết quả: ở file: [youtube\_music\_analysis.ipynb](https://github.com/vudinhhieu20/Introduction-to-Machine-Learning/blob/main/youtube_music_analysis.ipynb)

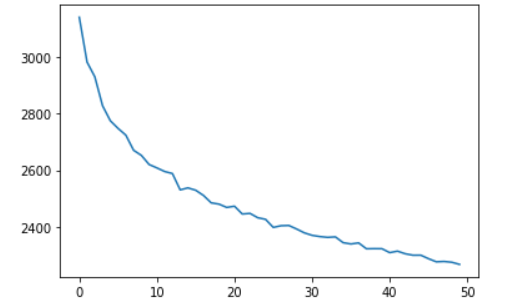


# Kết quả

## Các kết quả phân cụm

Lựa chọn số cụm dựa trên phương pháp elbow:

Dựa theo phương pháp, ta lựa chọn số cụm là 10



Phân cụm theo tiêu đề video

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Xem xét tỉ lệ các video trong các cụm theo số lượt xem

|  |  |
| --- | --- |
| 1 | 2 |
| 3 | 4 |
| 5 | 6 |
| 7 | 8 |
| 9 | 10 |

Đánh giá số lượt xem, lượt yêu thích và tổng thời gian chiếu:



Nhận xét điểm bất thường

Cụm số 7 đang có một sự bất thường là hiện tại cụm này cho thấy tập trung nhóm nhạc tình khúc bolero. Và phân phối các video trong này có nhiều video có lượt xem lớn tới các mức trăm triệu view, và tổng số lượng cụm đột biến với hơn 1 nghìn video. Điều này gây sự chú ý không chỉ bởi số lượng mà trên thực tế nếu nhóm chủ yếu là tình khúc bolero thì sẽ mâu thuẫn với thực tế rất ít các video ca nhạc về thể loại nhạc này có lượt xem ở mức triệu lượt xem đổ lên. Vì vậy cần phân cụm nhóm này thêm.

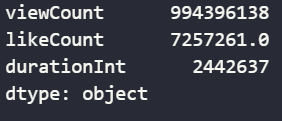
Phân cụm tiếp cụm bất thường

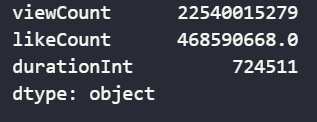
|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Xem xét tỉ lệ các video trong các cụm theo số lượt xem các cụm

|  |  |
| --- | --- |
| 1 | 2 |
| 3 | 4 |
| 5 |  |

Đánh giá số lượng lượt xem, lượt yêu thích và tổng thời gian video:





## Những suy luận rút ra được:

* Xét các 10 cụm phân loại đầu tiên ta có:

Ta thấy các thể loại chủ yếu của từng cụm là:

* Cụm 1: nhạc nhẹ nhàng chill lofi mix buồn
* Cụm 2: liên khúc nhạc trữ tình nhạc vàng bolero
* Cụm 3: liên khúc nhạc trẻ mới hay nhất
* Cụm 4: nhạc hay nhất bảng xếp hạng top
* Cụm 5: nhạc tiktok edm
* Cụm 6: liên khúc nhạc trẻ hay nhất
* Cụm 7: tình khúc nhạc sống bolero
* Cụm 8: nhạc remix nonstop
* Cụm 9: nhạc thiếu nhi
* Cụm 10: nhạc trẻ remix

Từ đây ta thấy một số nhận xét sau:

* Nhóm cụm số 7 đang có số lượt xem cao nhất, thứ 2 là nhóm nhạc thiếu nhi thứ 3 là nhóm nhạc trẻ
* Về tổng thời gian chiếu thì nhóm số 6 liên khúc nhạc trẻ cao nhất, thứ 2 là nhóm 7 nhạc tình khúc bolero thứ 3 là nhóm 2 liên khúc nhạc vàng trữ tình
* Như vậy cho thấy mọi người hiện nay thường quan tâm đến các nhóm thể loại là nhạc trẻ; nhạc trữ tình, nhạc vàng, bolero; nhạc tiktok, edm; nhạc lofi nhẹ nhàng; nhạc thiếu nhi; nhạc remix
* Đối tượng người xem là trẻ em rất lớn và các nội dung về nhạc thiếu nhi cũng có lượt xem cao. Điều này cho thấy việc âm nhạc trên youtube đang trở thành một món ăn tinh thần mới cho các em. Lý giải cho việc này có thể nhận thấy hai điều:
  + Đồ họa các video âm nhạc thu hút các em nhỏ cả về nghe lẫn nhìn
  + Bối cảnh đại dịch các em nhỏ ở nhà nhiều hơn, thì các video này là một biện pháp giúp các phụ huynh chăm các em
* Đối tượng người nghe Việt Nam thường thích nghe những video âm nhạc có từ nhất, hay top. Bởi nhiều người nghe không có thời gian, mà họ muốn tìm nhanh những bài hát đang nổi để bắt kịp xu thế, hay cũng thể lý giải bởi sự thích cái nhất trong tâm lý.
* Đối tượng người nghe Việt Nam nghe nhiều nhạc có giai điệu nhẹ nhàng, trữ tình. Lý giải điều này nằm ở chỗ, bản tính hiền lành, chất phác của chính người dân Việt; bên cạnh đó là mọi người thường nghe nhạc để thư giãn và họ chọn giai điệu nhẹ nhàng
* Đối tượng người trẻ nghe nhạc đang có số lượng lớn và thời lượng chiếu nhiều phản ánh đối tượng chính đang tham gia thị trường này
* Mọi người quan tâm nhiều thể loại liên khúc, tuyển tập những bài hát của cùng một thể loại nào đó
* Có bộ phận người nghe tập trung vào dòng nhạc có giai điệu sôi động như nhạc remix, nostop, nhạc edm. Nhưng đặc biệt ở chỗ nhạc remix của chúng ta có đặc điểm là sự biến tấu giai điệu trên nhiều thể loại khác như kết hợp remix với nhạc trẻ.
* Xét các cụm phân tích từ cụm 7 bất thường ở trên:

Ta thấy các thể loại chủ yếu của từng cụm là:

* Cụm 1: Chủ đề thiên về các chương trình âm nhạc của truyền hình vĩnh long, thể loại bolero
* Cụm 2: nhạc sống, vn365
* Cụm 3: Nhiều chủ đề chưa phân loại được
* Cụm 4: Thể loại bolero, giọng để đời
* Cụm 5: Bolero

Từ đây ta có thể rút ra một số nhận xét sau:

- Các cụm chủ yếu thiên về hai hướng về trữ tình bolero và nhóm còn lại, trong đó hướng bolero có thời lượng nhiều nhất

- Cụm nhiều chủ đề chưa phân loại đang chiếm nhiều video nhất, lượng xem cao nhất nhưng thời lượng thì ít hơn đáng kể

- Trong nhóm bolero ta thấy có một bộ phận người nghe quan tâm đến cụm ‘giọng để đời’, điều này cho thấy mọi người quan tâm đến người hát trước, đến giọng ca và ít quan tâm đến thể loại

- Nhóm các cụm chủ để chưa phân loại được có thể lý giải là nhóm các mv ca nhạc. Vì nhóm này số lượt xem thường cao và thời gian ngắn, đồng thời khi nghe các video này thì người nghe tập trung vào người trình bày. Còn bản thân tiêu đề có thể không nói lên được thể loại nhạc nhưng người nghe họ xác định thể loại hay quyết định nghe phụ thuộc và người hát. Điều này phù hợp với nhận xét ngay trên khi nhiều người quan tâm đến cụm ‘giọng ca để đời’.

# 5. Khó khăn và hướng phát triển tương lai

## a. Khó khăn:

- Tìm kiếm được nguồn dữ liệu toàn diện bao gồm tối đa các video âm nhạc mà mọi người quan tâm và nghe

- Nhiều thuộc tính gán cho các video chưa được đúng đắn và cụ thể

- Tiền xử lý, tìm ra các mối quan hệ ẩn phản ánh xu hướng người nghe

- Khi phân tích thể loại được mọi người yêu thích, gặp phải một vấn đề đó là khi phân tích theo tiêu đề thì rất nhiều video ca nhạc có tiêu đề hay tên bài hát không thể hiện trực tiếp thể loại âm nhạc của sản phẩm. Mà người nghe xác định thể loại nhạc dựa trên người trình diễn. Đôi khi yếu tố quyết định mọi người nghe nhạc lại nằm ở yếu tố người trình bày. Nên việc phân cụm ở đây gặp nhiều khó khăn. Vì vậy để phân tích tốt hơn cần cải tiến phương pháp để lọc người hát ở những video này.

## b. Hướng phát triển tương lai:

- Cần phải kết hợp với bài toán phân loại để phân loại thể loại âm nhạc của video, vì hiện nay việc tìm ra thể loại hay xu hướng quan tâm của mọi người dựa trên tiêu đề cho thấy những hạn chế ở nhiều video như mv ca nhạc. Vậy để biết thể loại mà mọi người quan tâm thì trước hết bản thân các video cần được gán thể loại rõ ràng và đúng đắn.

1. Tài liệu tham khảo

[1] *Tập bài giảng nhập môn học máy và khai phá dữ liệu - PGS.TS. Thân Quang Khoát*

[2] *https://developers.google.com/youtube/v3/docs*

[3] *https://scikit-learn.org/stable/modules/clustering.html*

[4] *https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html*

[5] *https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6API Reference | YouTube Data API | Google Developers*

[6][*https://scikit-learn.org/stable/auto\_examples/cluster/plot\_kmeans\_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py*](https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py)

[7] *https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods*