**ĐẠI HỌC QUỐC GIA TP. HCM**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



Hướng dẫn gán nhãn:

Bộ dữ liệu Thông tin Video trên Youtube phục vụ cho bài toán phân loại video

## Giới thiệu

Bộ dữ liệu Thông tin các video trên Youtube phục vụ cho bài toán phân loại video được thu thập thông qua Extension Apps Script ([Apps Script | Google Developers](https://developers.google.com/apps-script)) của Google Sheets. Sử dụng script để lấy dữ liệu từ Youtube Data API v3 ([API Reference | YouTube Data API | Google Developers](https://developers.google.com/youtube/v3/docs)).

Hướng dẫn cách thu thập thông qua Apps Script của Google Sheet trong file đính kèm.

Bộ dữ liệu cung cấp thông tin về các video (phần lớn là) tiếng Việt trên Youtube theo 15 chủ đề. Các thông tin được thu thập bao gồm: Video ID, Tiêu đề (Title), Mô tả (Description), ID của kênh (Channel ID), Thời lượng của video (Duration/Length)

## Mục đích gán nhãn

Việc thu thập các video trên youtube còn nhiều video sai lệch so với các chủ đề từ khóa tìm kiếm ban đầu, vì vậy việc rà soát lại và thực hiện gán nhãn cho video là hết sức cần thiết để có được một bộ dữ liệu tốt. Việc gán nhãn lại giúp cho các video được gán với nhãn đúng giúp cho việc thực hiện phân loại, tìm kiếm, dự đoán sau này được chính xác hơn.

## Mục tiêu bộ dữ liệu.

Bộ dữ liệu Thông tin các Video trên Youtube phục vụ cho bài toán phân loại video giúp các mô hình máy học có được bộ dữ liệu tốt để thực hiện huấn luyện mô hình cho ra kết quả tốt, dự đoán chính xác hơn. Ngoài ra còn có thể dùng để phân tích các thông số của video để đưa ra những thể loại nào đang được thịnh hành, những kênh youtube đang có xu hướng làm về chủ đề nào, thời lượng tối ưu cho từng thể loại video,…

## Định nghĩa.

Các thuộc tính của bộ dữ liệu:

| Tên cột | Ý nghĩa | Kiểu dữ liệu | Giá trị |
| --- | --- | --- | --- |
| channelId | ID của channel. Chuỗi kí tự base64 có độ dài 24 ký tự | Text | “UCk3VxZ9lqDwfj2HJQq\_wBlw”, “UC2ztoti9gzl5airj4f6pvGA”,... |
| videoID | ID của video Youtube. Chuỗi kí tự base64 có độ dài 11 ký tự | Text | “3x-peJ4ka-g”, “BnTZ0gVGrnQ”,.... |
| title | Tiêu đề của video | Text |  |
| Description | Mô tả của video | Text |  |
| viewCount | Số lượt xem của video | Numeric |  |
| likeCount | Số lượt thích của video | Numeric |  |
| Duration | Thời lượng video (đối với những video đang được phát trực tiếp trong lúc thu thập, thời lượng = 0) | Time(hh:mm:ss) | 0:42:57, 0:10:00,... |

Ngoài các thông tin thu thập được, cột label được thêm vào bộ dữ liệu cho biết nhãn của video. Miền giá trị label gồm 15 giá trị, là các chủ đề mà Youtube đã phân loại sẵn:

* Film & Animation (Phim & Hoạt hình)
* Autos & Vehicles (Ô tô & Xe cộ)
* Music (Âm nhạc)
* Pets & Animals (Thú cưng & Động vật)
* Sports (Thể thao)
* Travel & Events (Du lịch & Sự kiện)
* Gaming (Chơi game)
* People & Blog (Con người & Blog)
* Comedy (Hài kịch)
* Entertainment (Giải trí)
* News & Politics (Tin tức & Chính trị)
* How-to & Style (Làm thế nào & Phong cách)
* Educations (Giáo dục)
* Science & Technology (Khoa học & Công nghệ)
* Nonprofits & Activism (Phi lợi nhuận & Hoạt động).

## 

## Yêu cầu dán nhãn:

**Nhãn:**

Tất cả 15 nhãn được ghi bằng tiếng Anh theo phân loại từ Youtube. Mỗi nhãn là một thể loại chính ứng với video được dán nhãn. Mô tả chi tiết các nhãn:

Định nghĩa chi tiết của 15 nhãn:

**Film & Animation:** Gồm 2 lĩnh vực là Phim ảnh và Hoạt hình. Phim ảnh bao gồm phim, trailer phim, nhạc phim, các cảnh quay trong phim, hậu trường phim, và những video có nội dung liên quan. Hoạt hình bao gồm các video hoạt hình ngắn, chuyện ngắn, video kể chuyện và những video hoạt ảnh khác liên quan. Danh mục này thường được sử dụng bởi những thương hiệu lớn, cũng như các nhà sản xuất, liên quan trực tiếp đến ngành công nghiệp điện ảnh và hoạt hình.

**Autos & Vehicles:** Danh mục Ô tô và xe gồm những video liên quan đến ô tô, công nghệ ô tô, các loại xe,... Ngoài ra còn có các video dạng review, mẹo, giải pháp, ra mắt, và các dạng video khác có liên quan đến chủ đề.

**Music:** Bao gồm tất cả các thể loại nhạc, bài hát, và những video liên quan. Có thể bao gồm nhạc cụ và các hướng dẫn thanh nhạc. Đây là danh mục phổ biến nhất, có số lượng lớn người đăng kí, theo dõi.

**Pets & Animals:** Bao gồm các video về thú cưng, động vật, thức ăn cho động vật, mẹo chăm sóc, cũng như các sản phẩm liên quan đến động vật, các video vui nhộn, và các video khác liên quan.

**Sports:** Danh mục thể thao bao gồm các video về thể thao, thiết bị thể thao, các mẹo, các khoảnh khắc thú vị, các thống kê, hoặc các hướng dẫn tập thể thao, hoặc các hoạt động khác có liên quan.

**Travel & Events:** Danh mục bao gồm các mẹo vặt du lịch, địa điểm du lịch, các nơi nghỉ dưỡng, sự kiện, tổ chức các sự kiện, và các video khác có liên quan.

**Gaming:** Danh mục trò chơi bao gồm các thông tin về game, các mẹo, các đánh giá, và các video chơi game, cũng như các video khác liên quan.

**People & Blogs:** Gồm các video cho con người, lối sống, người nổi tiếng, các vấn đề về con người, các blog phổ biến, hoặc các trang thông tin, đánh giá liên quan đến con người và blogs.

**Comedy:** Gồm các video hài kịch, có thể bao gồm những bài nói, câu chuyện ngắn, phim hoạt hình, hoặc các video vui nhộn. Hay nói cách khác, danh mục này chứa các video gây cười.

**Entertainment:** Giải trí bao gồm nhiều chủ đề, từ nhảy, kịch, chuyện kể,... Kiểm tra các danh mục có liên quan (phim ảnh, âm nhạc, hài kịch) trước khi xếp vào danh mục giải trí.

**News & Politics:** Danh mục này bao gồm các video tin tức ngắn, các video thời sự, các tin chính trị, và các video liên quan.

**Howto & Style:** Gồm các video về cách làm, và các video liên quan đến thời trang.

**Education:** Danh mục giáo dục bao gồm các hướng dẫn,các chỉ dẫn mang kiến thức, các phương pháp học, lớp học, và các video giàu thông tin khác.

**Science & Technology:**

Bao gồm các video thuộc lĩnh vực khoa học, công nghệ. Ví dụ như các cải tiến, các sự thật, các công nghệ tương lai, công nghệ di động, máy tính,...

**Nonprofits & Activism:**

Gồm các hoạt động phi lợi nhuận và và các chiến dịch quảng bá.

* 1. Trường hợp 1: thông qua tiêu đề có thể xác định được thể loại của video dễ dàng thông qua từ khóa và ngữ cảnh.

**Bản tin** Dự báo Thời tiết đêm 10, ngày 11/5/2022 -> **News**

CẶP **TRƯỢT BĂNG NGHỆ THUẬT** NHẬT BẢN TẠI **THẾ VẬN HỘI** MÙA ĐÔNG BẮC KINH 2022 -> **Sports**

**Lạp Sườn Cao Bằng** - Bản tin **đặc sản vùng miền** của Đài Truyền Hình Nhân Dân 098.339.3412 -> **Travel & Events**

* 1. Trường hợp 2: Qua tiêu đề chưa thể xác định được nhãn thì ta sẽ dựa vào phần Description.

Title: CIP Online : **Cách Sử Dụng Giới Từ** Cho **Phương Tiện Giao Thông** -> “Auto & Vehicles” hoặc “Education”

Mô tả: Chào mừng các bạn đến với chương trình **học tiếng Anh** Online cùng với trường Anh ngữ CIP. Mình là Thảo. Hôm nay mình sẽ ... -> Education

* 1. Trường hợp 3: Trong trường hợp tiêu đề lẫn nội dung của video mang 2 chủ đề khác nhau sẽ chọn chủ đề chính và quan trọng nhất của video.

Ví dụ: **Phương tiện giao thông**/BÉ NGHE NHÌN VÀ ĐỌC TÊN CÁC LOẠI XE QUEN THUỘC TRÊN ĐƯỜNG PHỐ/**nhạc** vui nhộn.

Video trên bao gồm 3 chủ đề Auto & Vehicle, Education, Music nhưng mục đích của video hướng đến giáo dục trẻ em nên nhãn được chọn là Education.

* 1. Trường hợp 4: Thông qua title và description vân chưa xác định được thì sẽ sử dụng VideoId để lấy đường link đến video trên Youtube và người gán nhãn sẽ xem video và quyết định nhãn

## Quy trình gán:

1. Mỗi người dán nhãn truy cập theo đường dẫn tới thư mục chứa sheet dữ liệu và guideline của mình.
2. Thực hiện gán trực tiếp trên Google Sheet
3. Chỉ chọn một trong các chủ đề đã có tại cột **label**

