Báo cáo: Thu thập và tiền xử lí dữ liệu

I. Đặt vấn đề

Youtube là một nền tảng mạng xã hội chia sẽ video trực tuyến lớn nhất thế giới, có trụ sở chính tại Mỹ. Lượng dữ liệu trên Youtube cực kì lớn bao gồm nhiều dạng thông tin dưới hình thức văn bản, hình ảnh, video,…Tuy nhiên, để có thể xử lí, phân tích hay thống kê các thông tin này chúng ta cần phải thực hiện một quy trình thu thập và xử lí dữ liệu để có được một bộ dữ liệu “sạch” để các công việc có thể thực hiện một cách chính xác. Vì vậy nhóm quyết định thực hiện đồ án môn học Thu thập và tiền xử lí dữ liệu với bộ dữ liệu Thông tin các Video trên Youtube và Xây dựng mô hình phân loại Video.

II. Giới thiệu bộ dữ liệu

Bộ dữ liệu Thông tin các Video trên Youtube bao gồm 9 thuộc tính với 10734 dòng dữ liệu được thu thập tự động thông qua YouTube API V3 với script được viết bằng Extension App Script Google. 9 thuộc tính của bao gồm:

· ID: ID của Video trên Youtube

· Title: Tiêu đề của Video trên Youtube

· Description: Phân mô tả của Video

· Channe ID: ID của kênh đăng tải Video

· Channel Subriber count: Số người đăng kí của kênh đăng tải Video

· Channel view count: Số lượng lượt xem trên kênh

· View count: Số lượt xem của Video.

· Like count: Số like trên kênh

· Duration: Thời lượng Video

Bộ dữ liệu hướng đến các Video có ngôn ngữ tiếng Việt hoặc có lượng người xem/liên quan đến Việt Nam. Các Video trong bộ dữ liệu thu thập được hướng đến 15 chủ đề được phần loại sẵn theo Youtube Category bao gồm:

· Film & Animation

· Autos & Verhicles

· Music

· Pet & Animal

· Sport

· Travel & Events

· Gaming

· People & Blogs

· Comedy

· Entertaiment

· News & Politics

· Howto & Style

· Education

· Science & Technology

· Nonprofits & Activism

Đồng thời các chủ đề trên cũng sẽ là nhãn của các dòng dữ liệu. Mỗi điểm dữ liệu sẽ mang một nhãn ứng với 1 trong 15 chủ đề ở trên.

III. Phương thức thu thập

Sau khi xác định được nội dung và đối tượng cần thu thập mà bộ dữ liệu hướng đến, nhóm tiếp tục tiến hành giai đoạn thu thập dữ liệu. Nhóm sử dụng API để lấy thông tin từ Youtube, ở đây nhóm chọn Youtube API V3 để lấy thông tin cùng với Script được viết trên Google App Script(Extension được tích hợp sắn trên Google Sheet) và lưu lại Raw Data trên Google Sheet. Bằng cách đưa ra các từ khóa liên quan đến chủ đề, Youtube API V3 sẽ trả về các video liên quan đến chủ đề được tìm thấy thông qua các từ khóa.

IV. Tiền xử lí

Sau khi thu thập được dữ liệu Raw, chúng ta cần phải thực hiện tiền xử lí để dữ liệu về đúng được định dạng và đủ tiêu chuẩn một bộ tidy data. Sau khi thu thập, nhiều dòng dữ liệu bị null và còn có thuộc tính chưa đưa về đúng định dạng mong muốn, vì vậy ở bước này ta sẽ thực hiện xử lí các giá trị này.

Trước tiên, ta cần gộp 15 file ứng với 15 chủ đề lại thành một file. Vì sau khi thu thập bằng cách đưa ra từ khóa, ứng với mỗi chủ đề sẽ có 1 file raw data riêng vì vậy ta cần ghép các file này lại.

Tiếp theo cần loại bỏ các dòng dữ liệu bị trống cũng như các dòng dữ liệu bị trùng thông qua VideoID để đảm bảo các video không bị xuất hiện trùng lặp.

Sau khi xử lí các dòng dữ liệu trùng, ta tiếp tục xử lí về kiểu dữ liệu, đảm bảo một số thuộc tính ở đúng định dạng numeric như Channel view count, Channel Subriber count, Like count. Thứ 2 là ở định dạng của cột Duration, API trả về định dạng ở cột này theo chuẩn ISO-8601 có dạng PTHMS nên chúng ta sẽ đưa về định dạng h:m:s.

V. Gán nhãn

1. Quy trình gán nhãn

Sau quá trình tiền xử lí, bộ dữ liệu của chúng ta chỉ còn thiếu ở cột nhãn. Nhóm thực hiện việc dán nhãn qua 2 giai đoạn: dán nhãn đánh giá độ đồng thuận và dán nhãn diện rộng.

Đầu tiên, nhóm sẽ xây dựng bộ hướng dẫn dán nhãn Guideline. Guideline sẽ trình bày các quy tắc để dán được nhãn cho bộ dữ liệu. Các nhãn dữ liệu được giải thích rõ ràng:

**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.

Sau khi giải thích rõ các nhãn sẽ là các quy tắc để xác định được nhãn:

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.

**VD: 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**

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

Title: CIP Online :

**VD**: **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

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.

**VD**: **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.

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

Sau khi đọc kĩ và được giải thích rõ ràng, 2 người sẽ tiến hành dán nhãn độc lập với nhau. Khi 2 người dán nhãn hoặc tất việc dán nhãn, ta sẽ tiến hành đánh giá độ đồng thuận để

2. Đánh giá độ đồng thuận:

Sau khi gán nhãn, 2 bộ dữ liệu được dán nhãn sẽ được đem đi tính độ đồng thuận.

Xác suất đồng thuận của hai người là:

Pr(a) = 0.5953177257525084

Xác xuất độ đồng thuận của 2 người trên từng nhãn:

Autos & Vehicles : 0.0008053601190143287

Comedy : 0.0008053601190143287

Education : 0.0008053601190143287

Entertainment : 0.0008053601190143287

Film and Animation : 0.0008053601190143287

Gaming : 0.0008053601190143287

Howto & Style : 0.0008053601190143287

Music : 0.0008053601190143287

News & Politics : 0.0008053601190143287

Nonprofits & Activism : 0.0008053601190143287

People & Blogs : 0.0008053601190143287

Pets & Animals : 0.0008053601190143287

Science & Technology : 0.0008053601190143287

Sports : 0.0008053601190143287

Travel & Events : 0.0008053601190143287

Xác suất giả định của 2 annotator trên 15 nhãn: Pr(e) = 0.012080401785214934

Độ đồng thuận của 2 người: K = 0.5903692213629828

Theo thang đo độ đồng thuận của Cohen & Kappa, mức đồng thuận này ở mức Moderate agreement cho nên nhóm sẽ tiếp tục sử dụng để thực hiện dán nhãn diện rộng trên toàn bộ dữ liệu.

VI. Mô hình máy học thực hiện

VII. Tổng kết và hướng phát triển tương lai

Sau khi thực hiện, cơ bản nhóm đã xây dựng được một bộ dữ liệu Thông tin các Video trên Youtube từ việc thu thập, tiền xử lý, dán nhãn đến áp dụng vào bài toán thực tế là phân loại video dựa trên tiêu đề. Tuy nhiên vẫn còn nhiều thiếu sót trong bộ dữ liệu mà nhóm đã xây dựng được như số lượng điểm dữ liệu còn quá thấp so với tổng thể video trên Youtube, một số nhãn chưa có sự phân định rõ ràng khi áp dụng từ Youtube Category khi áp dụng cho các nội dung tại Việt Nam(Ví dụ: Entertainment và Music), bộ dư liệu chỉ có ý nghĩa tức thời vì thông tin trên Youtube thay đổi liên tục theo thời gian cho nên dữ liệu sẽ bị lỗi thời sau một thời gian ngắn...

Trong tương lai, nhóm sẽ mở rộng bộ dữ liệu về cả số lượng điểm dữ liệu lẫn số thuộc tính để có thể đa dạng bộ dữ liệu hơn. Các nhãn có thể được thay đổi để có thể sát với tình hình thực tế và các xu hướng của hiện tại.

Mặc dù độ đồng thuận đang ở mức chấp nhận được và việc gán nhãn khá hiệu quả. Tuy nhiên nhóm cần phải cải thiện guideline để có thể đạt được độ đồng thuận tốt hơn và việc gán nhãn được chính xác hơn, tránh các trường hợp gán nhãn nhâm cũng như mô tả chính xác hơn về từng nhãn.

Đồng thời xây dựng một hệ thống tự động thu thập thông tin từ youtube theo pipeline có sẵn để dữ liệu được sát với thực tế nhất bởi thông tin ở Youtube thay đổi liên tục theo thời gian hoặc tự.