**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----🙞🙜🕮🙞🙜-----



**MÔN: TRUY TÌM THÔNG TIN**

**ĐỀ TÀI:**

**XÂY DỰNG SEARCH ENGINE TÌM KIẾM BÀI HÁT**

**Giảng viên hướng dẫn: ThS. QUÁCH ĐÌNH HOÀNG**

**Nhóm thực hiện:**

1. **Lương Quang Minh 17110331**
2. **Mai Đăng Khoa 18110306**

**Tp.HCM – Tháng 05 / 2023**

**Đề xuất Project**

1. **Giới thiệu:**

**- Giới thiệu chung:** Hiện nay, khối lượng thông tin khổng lồ khiến con người phải tốn nhiều thời gian cho việc tìm kiếm thông tin thích hợp. Do đó, cần đòi hỏi xây dựng các kỹ thuật truy hồi thông tin nhằm giải quyết vấn đề về quá tải thông tin.

- **Giới thiệu về đề tài:**

Ngày nay, nhu cầu nghe nhạc, tìm kiếm các bài hát của người dùng trở nên quan trọng, thiết yếu. Bên cạnh đó, sự phát triển của internet và công nghệ về các thiết bị smart phone, tablet,.. càng đẩy mạnh nhu cầu nghe nhạc, tìm kiếm bài hát thường xuyên của người dùng. Vì vậy, nhóm chúng tôi lên ý tưởng xây dựng một search engine để đáp ứng nhu cầu tìm kiếm đa dạng, thường xuyên, chỉ dựa trên lượng thông tin nhỏ và rời rạc. Từ đó đáp ứng được nhu cầu giải trí tức thời của người dùng.

Người dùng sẽ nhận được kết quả trả về bao gồm thông tin những bài hát phù hợp với những thông tin mà họ cung cấp. Từ đó, họ có sự lựa chọn bài hát phù hợp.

Thông tin kết quả trả về của một bài hát bao gồm những thông tin sau:

- Tên bài hát

- Tên ca sĩ thể hiện

- Lời bài hát

- Tên nhạc sĩ sáng tác (Không có trả về null)

- Độ dài của bài hát

- Số lượt nghe, lượt thích, lượt bình luận,.. của bài hát

1. **Information retrieval system(IRs):**

**Truy tìm thông tin**: là tìm kiếm thông tin ở một dạng không có cấu trúc, thỏa mãn nhu cầu thông tin từ những nguồn thông tin lớn. Truy tìm thông tin thu thập các nguồn thông tin liên quan đến một thông tin cần tìm kiếm, có thể dưạ trên siêu dữ liệu hoặc trên việc đánh chỉ mục toàn văn (hoặc dựa trên nội dung khác).

**Có nhiều dạng tài liệu trong hệ thống IR:**

-Tài liệu văn bản: Dạng ký tự, có cấu trúc đơn giản

-Tài liệu phi văn bản: Dạng phương tiện như hình ảnh, âm thanh, video,..

Truy vấn : diễn tả nhu cầu thông tin cho những tài liệu liên quan đến một đề tài nào đó, là thông tin được đưa vào hệ thống để thực hiện công việc tìm kiếm.

**Các dạng truy vấn:**

- Từ, cụm từ, câu có thể là đoạn văn hoặc một dạng phi văn bản(hình ảnh…)

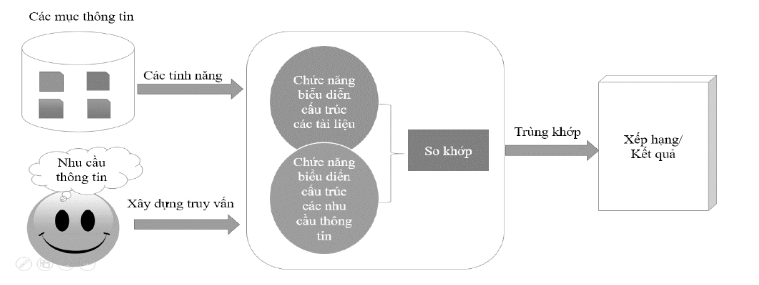
Từ chỉ mục (k): một đơn vị ngữ nghĩa, một từ, một cụm từ ngắn, hay là dạng gốc của một từ. Một tài liệu d sẽ được biểu diễn bởi một tập các từ chỉ mục ki. (i=0, 1, 2,2,…)

Cơ sở dữ liệu(DB): Tập n tài liệu, dj ∊ DB. (j=1,2,…, n)

Bảng từ vựng: Tập t từ chỉ mục, k I ∊ T, i=(1,2,…, t).

**Cấu trúc của một hệ thống IR:**

cấu trúc của một hệ thống IR được mô tả như sơ đồ sau:



Các mục thông tin: Chứa cơ sở dữ liệu lớn các thông tin thu thập được, được lưu trữ ở các hệ thống máy tính lớn.

Nhu cầu thông tin: Là những truy vấn người dùng muốn thực hiện

Chức năng biểu diễn cấu trúc các tài liệu: Các mục thông tin ban đầu không có cấu trúc, hệ thống IR cần phải biểu diễn cấu trúc các tài liệu hay các mục thông tin thành dạng cấu trúc nào đó bằng các mô hình truy tìm thông tin.

Chức năng biểu diễn cấu trúc các nhu cầu thông tin: Mặc định các nhu cầu thông tin của người dùng là ngôn ngữ tự nhiên, không có cấu trúc. Đòi hỏi hệ thống truy tìm thông tin cần phải phân tách yêu cầu người dùng thành các thông tin theo một cấu trúc nào đó để thực hiện so khớp với các tài liệu.

So khớp: Thực hiện so khớp nhu cầu thông tin của người dùng các mục thông tin bằng các thuật toán và các phương pháp của hệ thống IR.

Xếp hạng/kết quả: Đưa ra kết quả truy tìm và mức độ liên quan giảm dần nhờ chức năng xếp hạng của hệ thống.

**Các phương pháp truy tìm thông tin:**

**Mô hình boolean:** là mô hình truy vấn đơn giản nhất, bao gồm mối quan hệ giữa các thuật ngữ của yêu cầu thông tin và các tài liệu. Các câu truy vấn được biểu diễn bởi tập các khóa kết nối với tập phép toán Bool. Ba loại toán tử thường được sử dụng: OR, AND, NOT.

Toán tử OR: Xem xét hai thuật ngữ đồng nghĩa.

Toán tử AND: Tổ hợp các thuật ngữ(hay từ khóa) vào một câu thuật ngữ.

Toán tử NOT: là hạn chế hay thuật ngữ hẹp, thông thường được sử dụng cùng với toán tử AND.

1. **Elasticsearch:**

**- Elasticsearch:** là một công cụ tìm kiếm dựa trên nền tảng Apache Lucene. Nó cung cấp một bộ máy tìm kiếm dạng phân tán, có đầy đủ công cụ với một giao diện web HTTP có hỗ trợ dữ liệu JSON. Elasticsearch được phát triển bằng Java và được phát hành dạng nguồn mở theo giấy phép Apache.

Elasticsearch (ES) là một search engine và là công cụ hỗ trợ tìm kiếm văn bản hoặc tài liệu dạng văn bản text, string, file (PDF, word, excel),..được kế thừa từ Lucene Apache Elasticsearch nên ES hoạt động như 1 web server, có khả năng tìm kiếm nhanh chóng và khả năng phân tích và thống kê dữ liệu. Elasticsearch chạy trên server riêng và đồng thời giao tiếp thông qua RESTful do vậy nên nó không phụ thuộc vào client viết bằng gì hay hệ thống hiện tại được viết bằng gì. Nên việc tích hợp nó vào hệ thống trở nên dễ dàng, chỉ cần gửi request http để nó trả về kết quả.

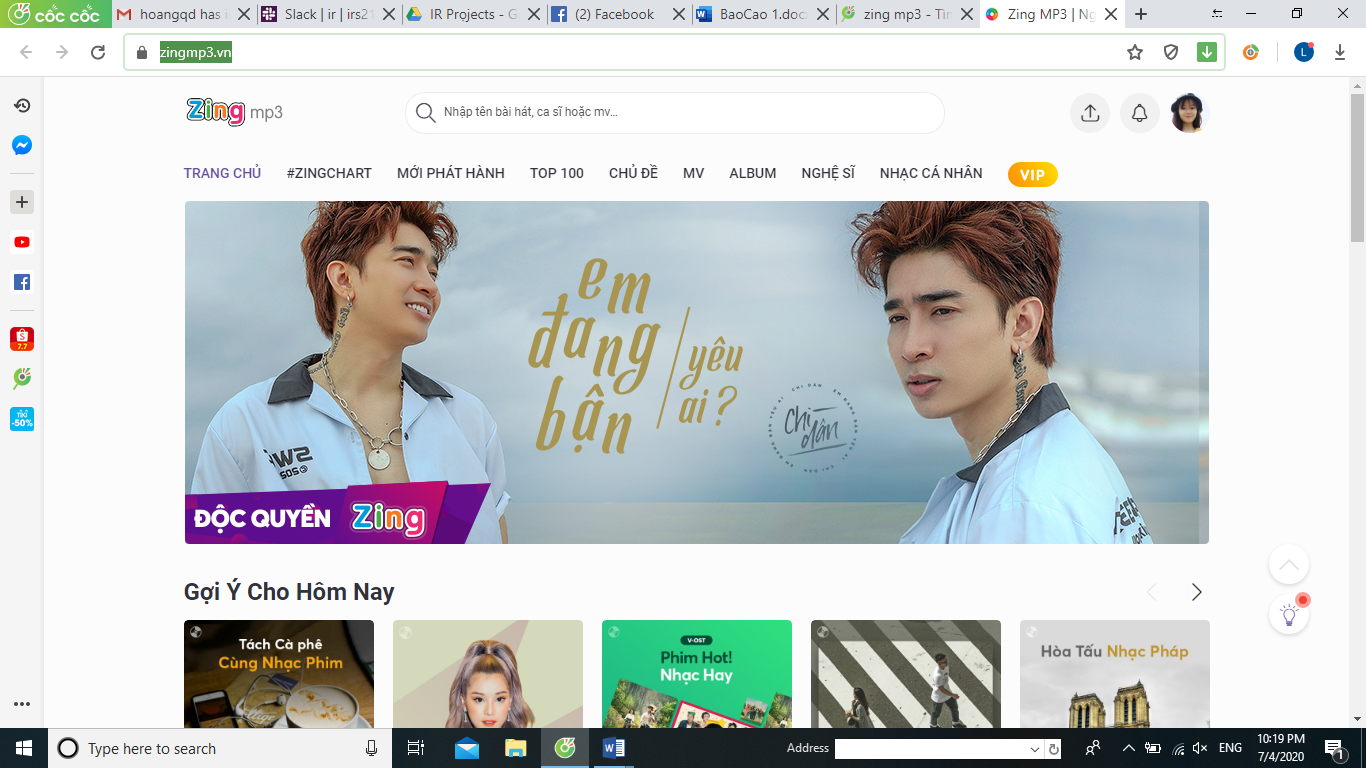
Elasticsearch là một open source được phát triển bằng Java nên có hệ thống phân tán và có khả năng mở rộng tốt. Ngoài ra, Elasticsearch là một server riêng biệt để phục vụ việc tìm kiếm dữ liệu. ES có thể được sử dụng để làm DB chính tuy nhiên ES không mạnh trong các thao tác CRUD nên thường sẽ dùng song song với một DB chính( SQL,MySQL,..)

Quá trình hoạt động của ES được mô tả : tạo Mapping Reindex Data & Push Data lên Server ES Get Data từ server ES về.

**-Logtash:** Với những hệ thống lớn việc quản lý log và phân loại log bằng việc xem file log của server để xác định thông tin của log, phân loại log là khá khó khăn. Cần thiết phải có một công cụ quản lý log một cách tốt hơn, sớm phát hiện những lỗi phát sinh của server hoặc kiểm tra các thông tin về log. Hiện nay cũng có khá nhiều công cụ để quản lý log khác nhau. Bộ công cụ Logstash có nhiều ưu điểm như phần mềm mã nguồn mở hoàn toàn miễn phí, cung cấp dịch vụ quản lý log rất tốt và dễ sử dụng.

Logtash là một công cụ sử dụng để thu thập, xử lý log được viết bằng java. Nhiệm vụ chính của logstash là thu thập log sau đó chuyển vào Elastichsearch. Mỗi dòng log của logstash được lưu trữ đưới dạng json.

1. **Xử lý:**
2. **Dữ liệu**
   1. **Lấy dữ liệu từ Website**

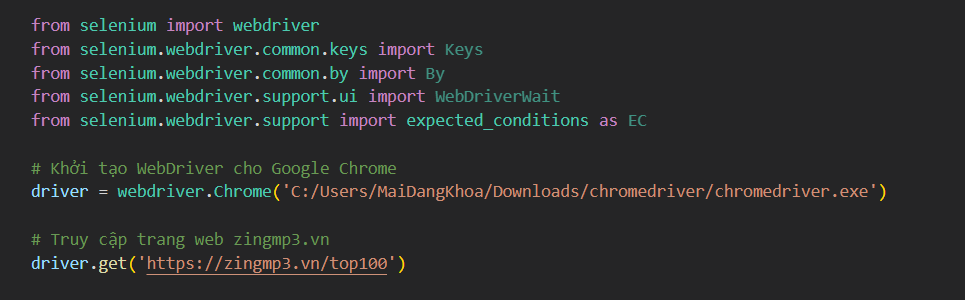
Dữ liệu được lấy từ trang Web: <https://zingmp3.vn/>

* 1. **Xử lý dữ liệu**

Sử dụng phương pháp Scrapy, ngôn ngữ Python để lấy dữ liệu từ trang web Zingmp3.

Các thông tin được lấy:

* id: id của bài hát trên web
* name: tên bài hát
* author: tên ca sĩ thể hiện
* album: tên album bài hát
* producer: tên nhà sản xuất
* Genre: thể loại bài hát
* Lyric: Lời bài hát
* Time: thời lượng bài hát
* Duration: chiều dài bài hát (ms)
* countLike: Lượt like bài hát
* countListen: lượt nghe bài hát
  1. **Tiến hành lấy dữ liệu**
* Dữ liệu nhóm sử dụng danh sách các bài hát thu thập từ trang web mp3.zing sử dụng Selenium để Crawl dữ liệu



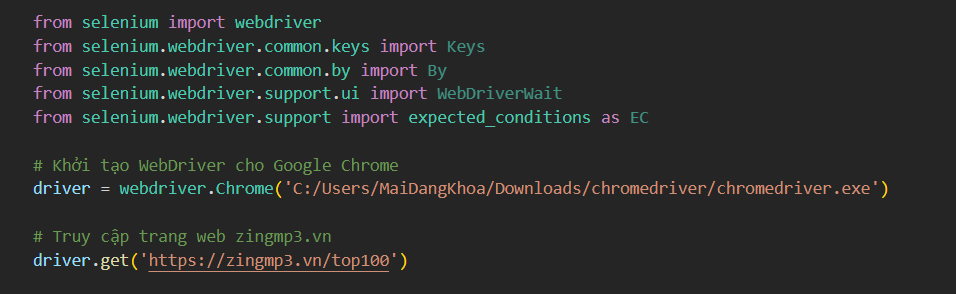
* Dữ liệu gồm 1200 document thu thập từ 12 danh sách top 100 bài hát các thể loại khác nhau.

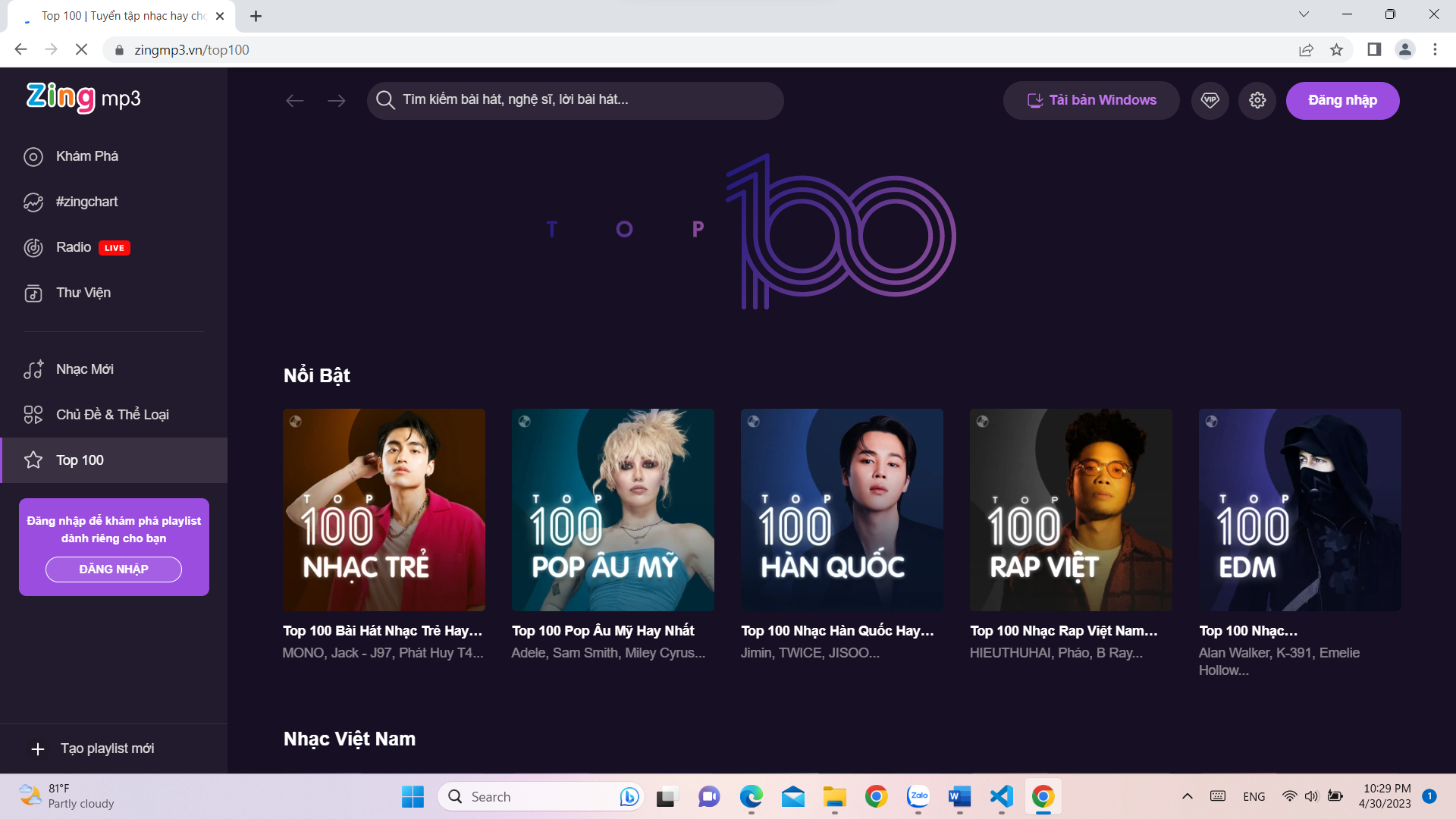
Dữ liệu bao gồm các fields: id, title, author, album, producer, lyrics, time, countLike, countListen

Trong đó:

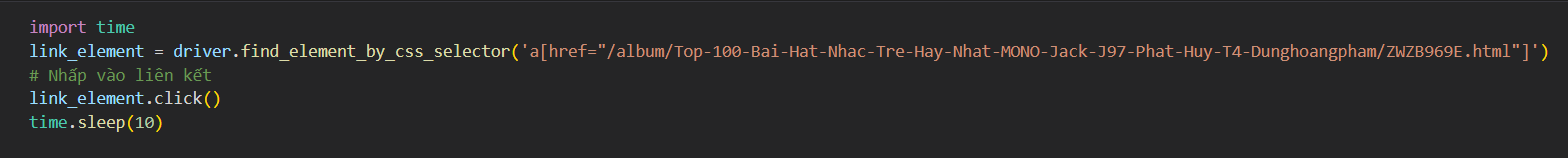
* Id: Id của bài hát
* Title: tên bài hát
* Author: Tác giả bài hát
* Album: Album bài hát
* Producer: Nhà sản xuất
* Lyrics: Lời bài hát
* Time: Thời gian bài hát
* countLike: Số lượt yêu thích
* countListen: Số lượt xem
* Cách lấy dữ liệu

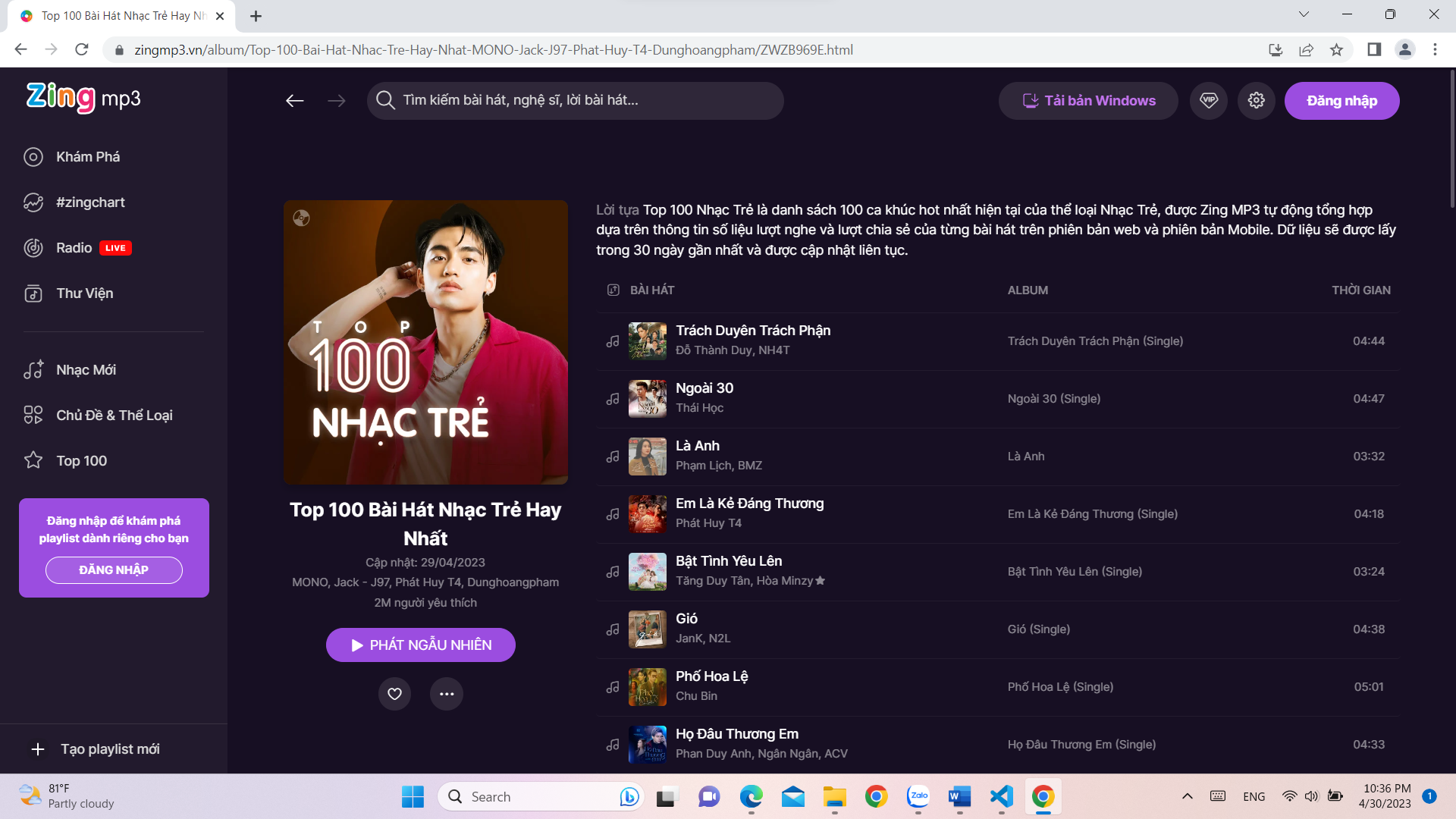
Truy cập zingmp3.vn danh sách top 100



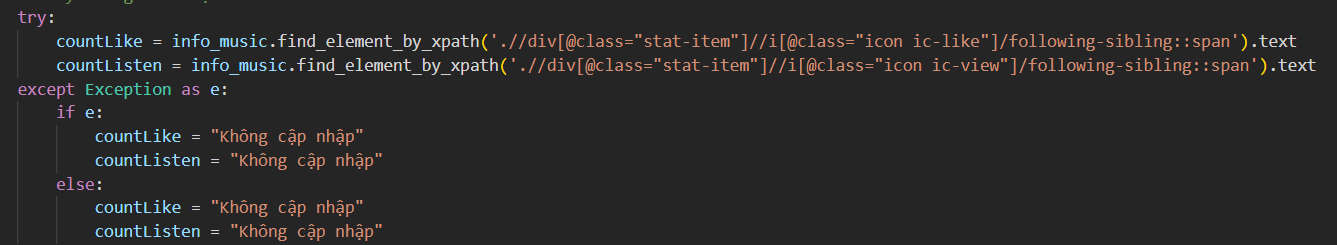


Chọn link nhạc trẻ

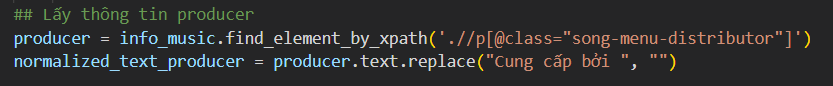




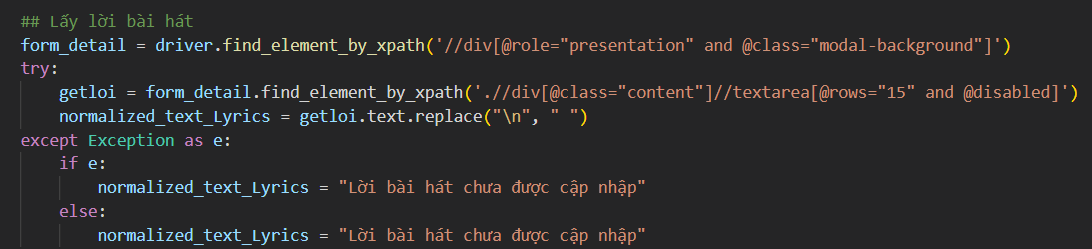
Lấy thông tin countLike, countListen



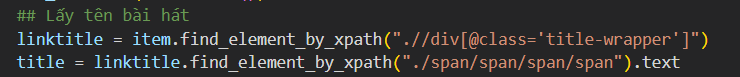
Lấy thông tin nhà sản xuất



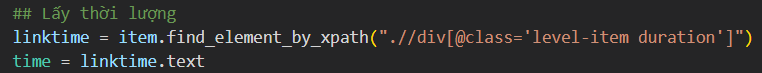
Lấy thông tin lời bài hát



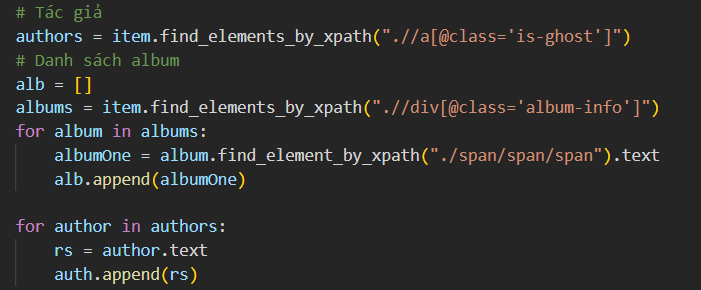
Lấy thông tin tiêu đề bài hát



Lấy thông tin thời lượng bài hát



Lấy danh sách tác giả và album



Danh sách thông tin thu thập

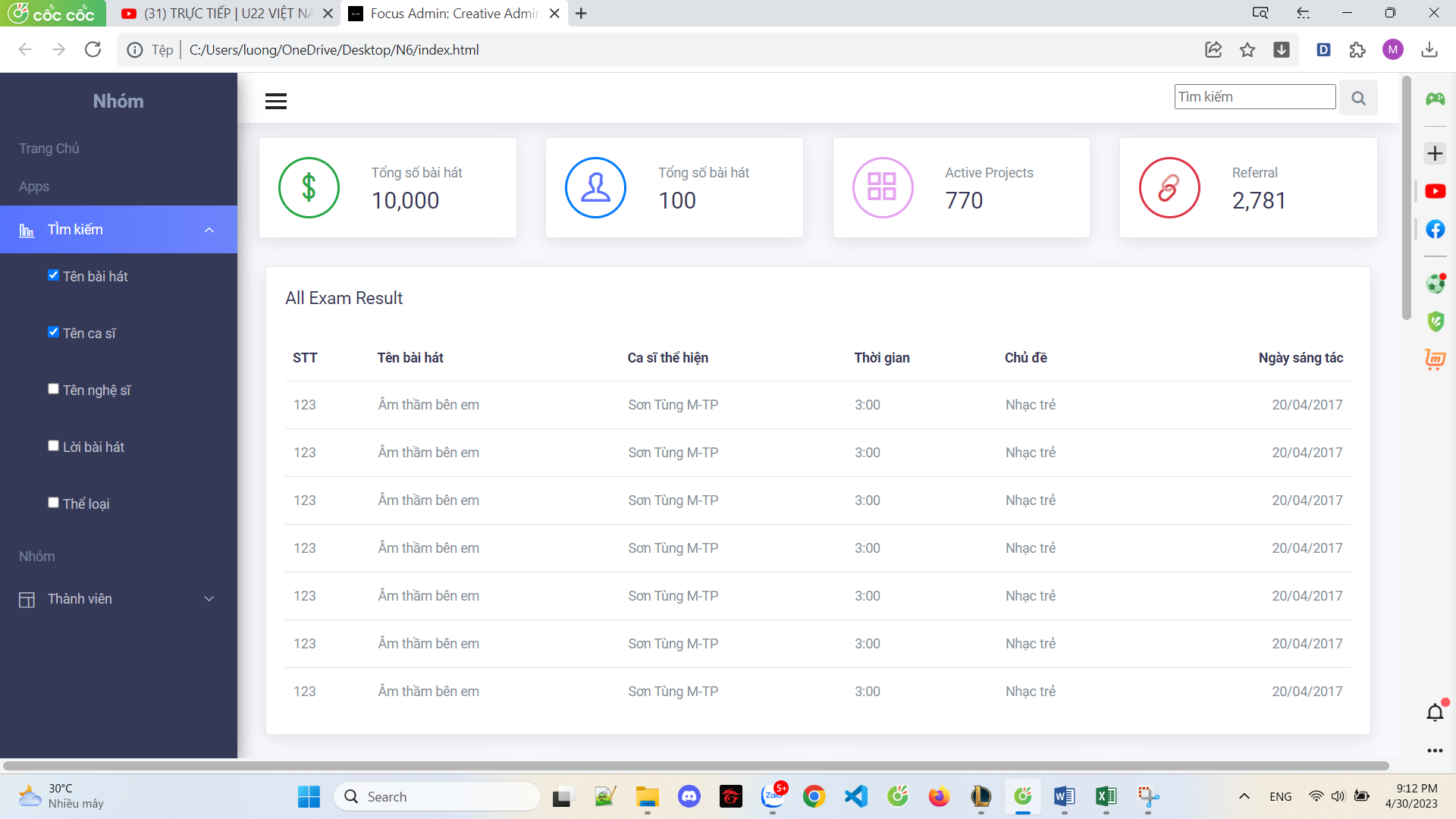


Bước 7: Đưa dữ liệu về file csv.



1. **Xây dựng giao diện web**

Website nhóm xây dựng nhằm mục đích demo cho search engine

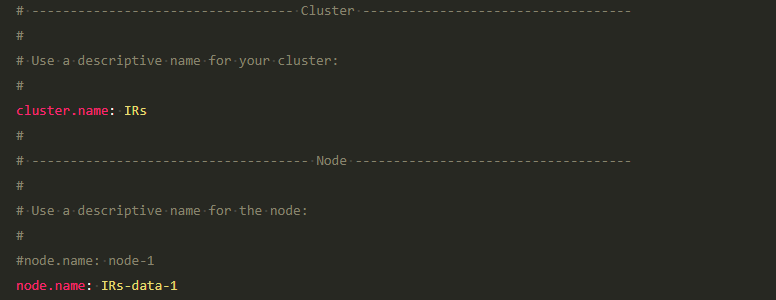


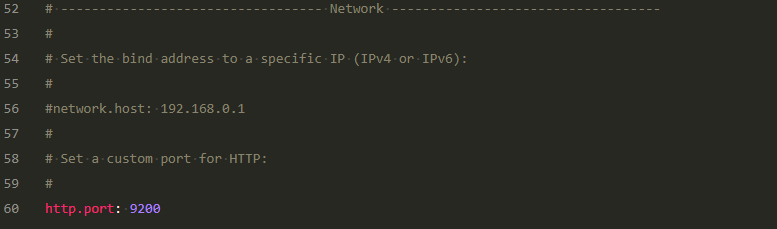
1. **Xây dựng search engine**

Xây dựng search engine bằng Elastic Search để tìm kiếm lời bài hát theo:

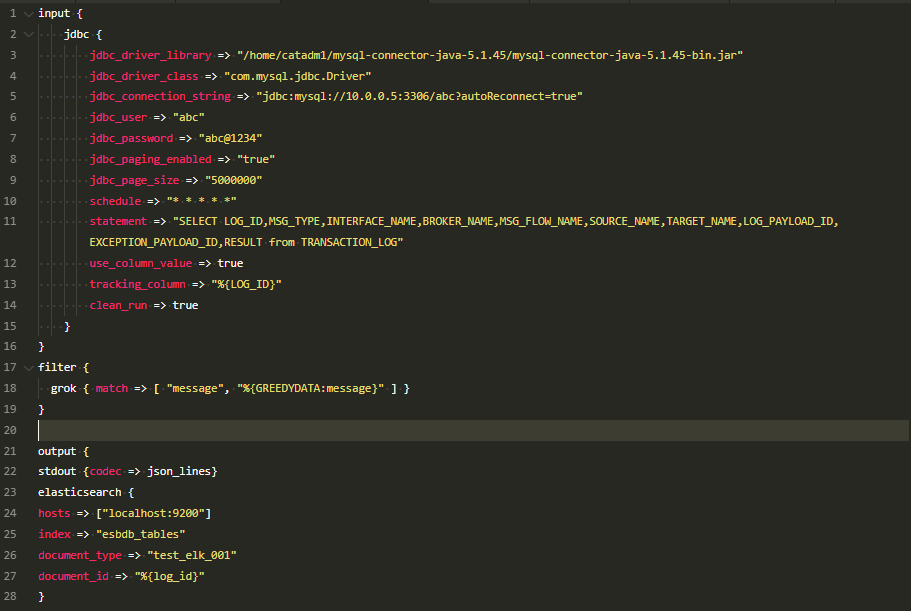
* Tên bài hát
* Ca sĩ thể hiện
* Lời bài hát
* Thời gian phát hành
* Theo thể loại
  1. **Cấu hình elasticsearch:**
  + Sử dụng elasticsearch-7.7.1 được cung cấp sẵn với elasticsearch để cấu hình kết nối tới CSDL, lưu các index,…
  + Trong folder chúng ta cấu hình trong thư mục con config:

+ file *elasticsearch.yml:* Cấu hình cluster, node và networks

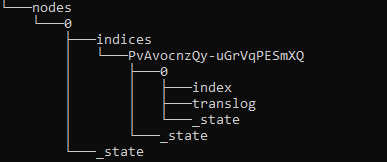




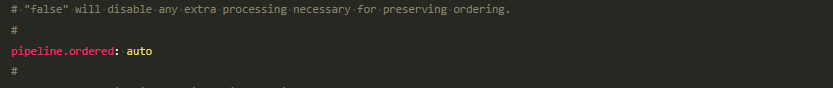
+ Cấu hình trong file *logstash-mySQL.conf:*

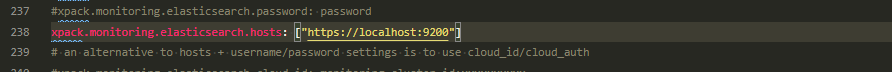


* Khi chạy server elasticsearch thì dũ liệu được index lưu trong folder *data* theo các node:



* 1. **Cấu hình logstash:**
  + Sử dụng logstash-7.7.1 được cung cấp kèm để tự động index, cập nhập dữ liệu thay đổi vào folder data trong folder elasticsearch-7.7.1,…
  + Trong *config/logstash.yml:*





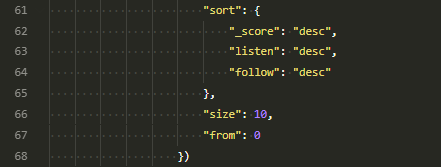
* Trong folder *bin* : tạo một file conf để thực hiện kết nối với elastic server đã cấu hình ở trên và connection string đến CSDL của chúng ta, cùng bộ jar hỗ trợ riêng(ở đây đang kết nối đến MySQL)



* 1. **Xây dựng chức năng:**
* Search thông thường được xây dựng trên function:



* Các bộ filter theo field(s) cũng được tích hợp trong function này, cho phép search lân cận trong từng term.
* Thực hiện sort và quyết định số lượng trích xuất kết quả truy vấn tốt nhất được thực hiện tĩnh trên code (có thể cải thiện hiện thị bằng sử dụng biến thay thế với người dùng)



* Ngoài ra trong project còn có các function render các dự liệu chuẩn hiển thị cho người dùng, xử lý hiển thị số lớn,…
* Tất các các chức năng đều được controls và quản lý trực tiếp trong file *app.py.*