**Trường Đại học Khoa học Tự nhiên – ĐHQG TPHCM**

**Khoa Công nghệ Thông tin**



**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

**SEARCH ENGINE**

**Giảng viên hướng dẫn:** TS. Võ Hoài Việt

**Nhóm sinh viên thực hiện:**

Hoàng Trọng Vũ – MSSV: 20120025

Nhan Gia Khâm – MSSV: 20120504

**Lớp:** CNTN2020

Thành phố Hồ Chí Minh – Tháng 7 năm 2021

**MỤC LỤC**

[I. Giới thiệu đồ án 3](#_Toc76986512)

[I.1. Nội dung đồ án 3](#_Toc76986513)

[I.2. Yêu cầu đồ án 3](#_Toc76986514)

[II. Cài đặt chương trình 4](#_Toc76986515)

[III. Các chức năng của chương trình 5](#_Toc76986516)

[III.1. Tìm kiếm 6](#_Toc76986517)

[III.2. Lịch sử tìm kiếm 9](#_Toc76986518)

[III.3. Chỉnh sửa dữ liệu 10](#_Toc76986519)

[III.3.1. Thêm, xóa file 10](#_Toc76986520)

[III.3.2. Cập nhật dữ liệu 12](#_Toc76986521)

[IV. Tổng kết 14](#_Toc76986522)

[V. Tài liệu tham khảo 15](#_Toc76986523)

**I. Giới thiệu đồ án**

**I.1. Nội dung đồ án**

Rút trích đơn giản nội dung chính của văn bản tiếng Việt. Tìm kiếm những văn bản có nội dung tương ứng với từ khóa do người dùng nhập vào, xếp hạng kết quả tìm kiếm theo mức độ liên quan đến từ khóa từ cao đến thấp.

**I.2. Yêu cầu đồ án**

Các sinh viên trong lớp cùng tập hợp dữ liệu thống nhất bao gồm các văn bản tiếng Việt. Các tập tin văn bản được đặt trong cùng một thư mục với tên tập tin tương ứng với tựa đề và phần mở rộng .txt. Dựa trên các văn bản nguồn, sinh viên tự tạo ra tập tin siêu dữ liệu (metadata) ở dạng nhị phân hoặc văn bản: thông tin về các văn bản đã tiền xử lý và nội dung chính của từng văn bản theo cấu trúc tự định nghĩa. Khi thêm/xóa tập tin văn bản, chương trình tự động cập nhật dữ liệu của tập tin siêu dữ liệu.

**Yêu cầu lập trình:**

* Hiển thị menu cho phép người dùng chọn các chức năng.
* Tận dụng các cấu trúc dữ liệu đơn giản đã học.
* Phải truy vấn dựa trên tập tin siêu văn bản, không truy vấn trực tiếp trên các văn bản.
* Tự cài đặt thuật toán sắp xếp, tìm kiếm (không sử dụng thư viện có sẵn) với mảng hoặc danh sách liên kết. Tổng quát hóa bằng cách sử dụng con trỏ hàm (tùy chọn).
* Chỉ sử dụng thư viện string.h của C để xử lý chuỗi.

# II. Cài đặt chương trình

Những công nghệ sử dụng trong quá trình cài đặt:

* Môi trường lập trình: Microsoft Visual Studio 2019.
* Framework sử dụng: Qt5.
* Ngôn ngữ lập trình: C/C++.

Cấu trúc dữ liệu sử dụng:

* Singly Linked List
* Skip List
* Mảng một chiều

Thuật toán sử dụng:

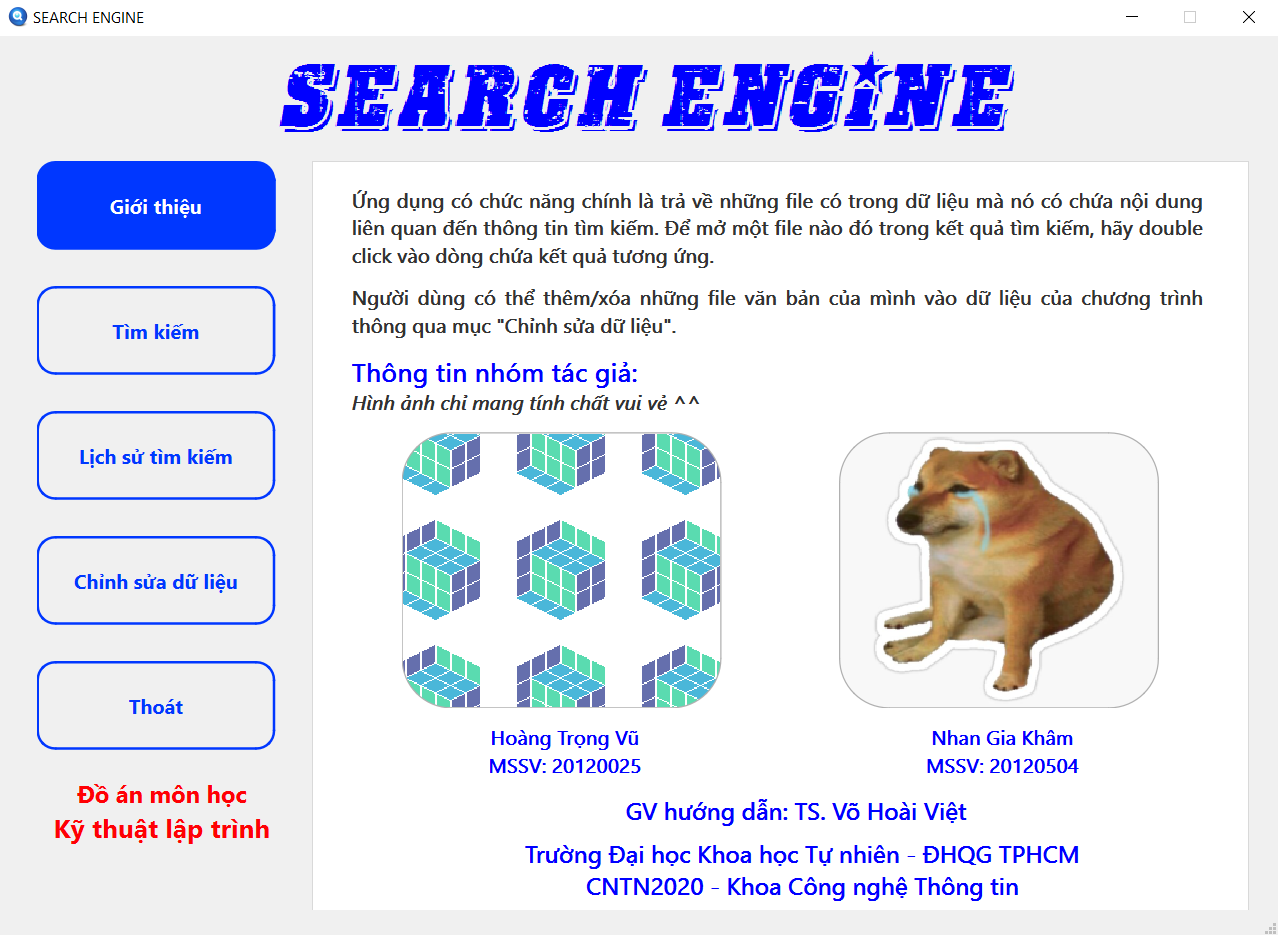
* Thuật toán tìm kiếm nhị phân
* Thuật toán sắp xếp Radix Sort

Tập dữ liệu gốc (database) được lấy từ file *Train.rar* trong đường dẫn sau:  
<https://drive.google.com/drive/folders/1yJ_cM5g69S7GpfxX6ULTgjEn52fqiKFF>

Chi tiết về các cấu trúc dữ liệu, thuật toán, các thao tác và lý do lựa chọn những hướng đi đó được trình bày kĩ trong file *Nội dung thực hiện.pptx* ở cùng thư mục Documents với báo cáo này.

# III. Các chức năng của chương trình

Giao diện chính của chương trình như sau:



Các chức năng chính của chương trình bao gồm:

* Tìm kiếm
* Xem lịch sử tìm kiếm
* Chỉnh sửa dữ liệu: thêm file *.txt* của người dùng vào dữ liệu của chương trình, xóa những file *.txt* mà người dùng đã thêm ra khỏi dữ liệu

## III.1. Tìm kiếm

Chương trình sẽ trả về những file văn bản có nội dung liên quan đến thông tin tìm kiếm mà người dùng nhập vào theo thứ tự giảm dần về mức độ liên quan, đồng thời cho biết số kết quả tìm thấy và thời gian tìm kiếm.

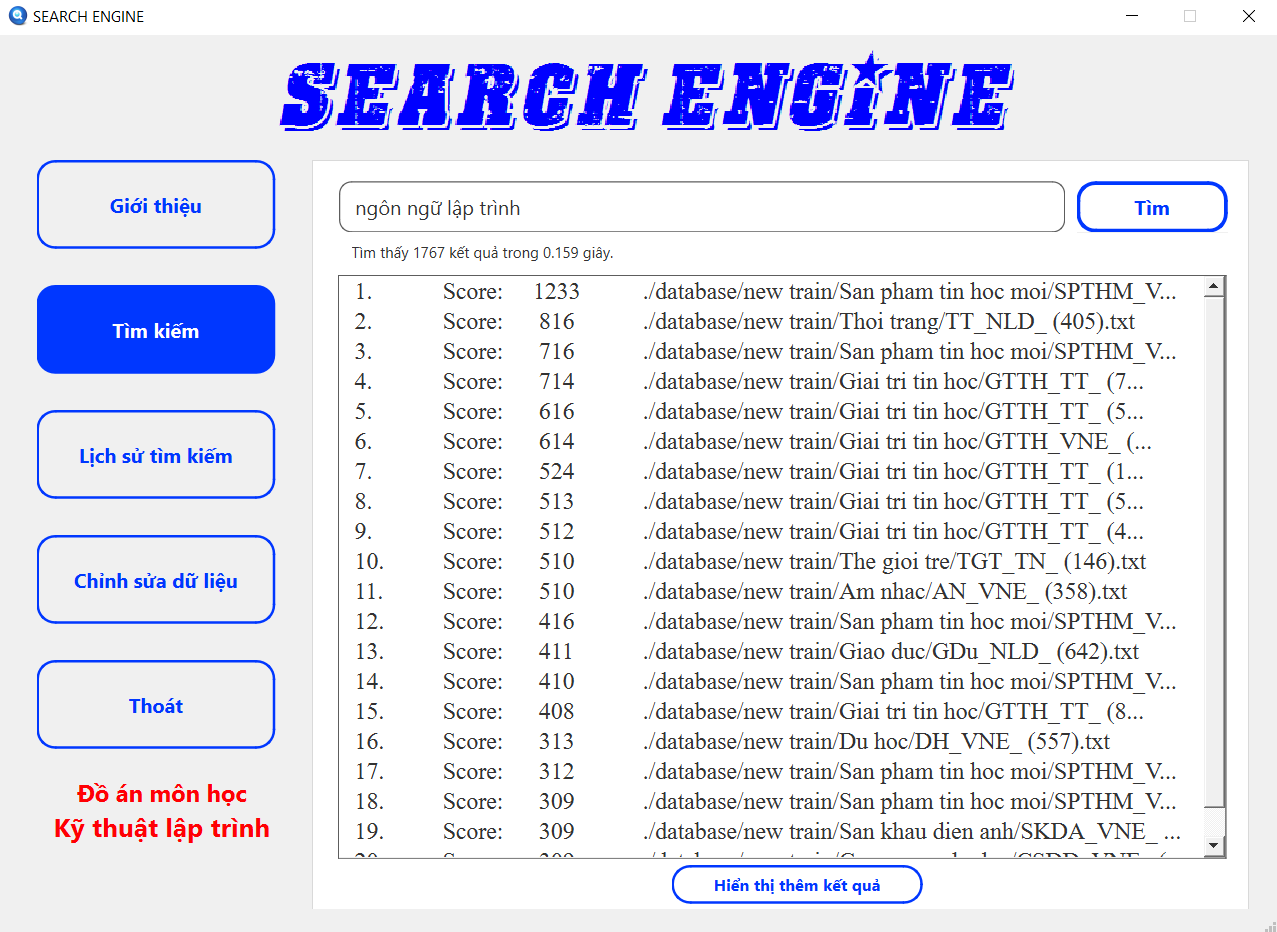
Mặc định, 20 kết quả tìm kiếm có độ liên quan cao nhất đến thông tin tìm kiếm sẽ được hiển thị. Người dùng có thể xem thêm những kết quả khác thông qua nút “**Hiển thị thêm kết quả**”. Mỗi lần nhấn sẽ có thêm nhiều nhất 20 kết quả nữa được hiển thị ra (khi ta đã hiển thị ra gần đủ số kết quả tìm thấy thì con số này có thể sẽ thay đổi).

Người dùng có thể xem nội dung của các file xuất hiện trong kết quả tìm kiếm bằng cách *double-click* vào các dòng kết quả tương ứng.

**Lưu ý.** Khi người dùng nhập vào một chuỗi chứa toàn những từ, cụm từ được xếp vào ***stopword*** của chương trình thì kết quả tìm kiếm là “*Không tìm thấy dữ liệu liên quan*”. Ví dụ như những chuỗi “ạ”, “dạ”, “đột nhiên”, “bỗng nhiên”,…

***Một số hình ảnh minh họa:***

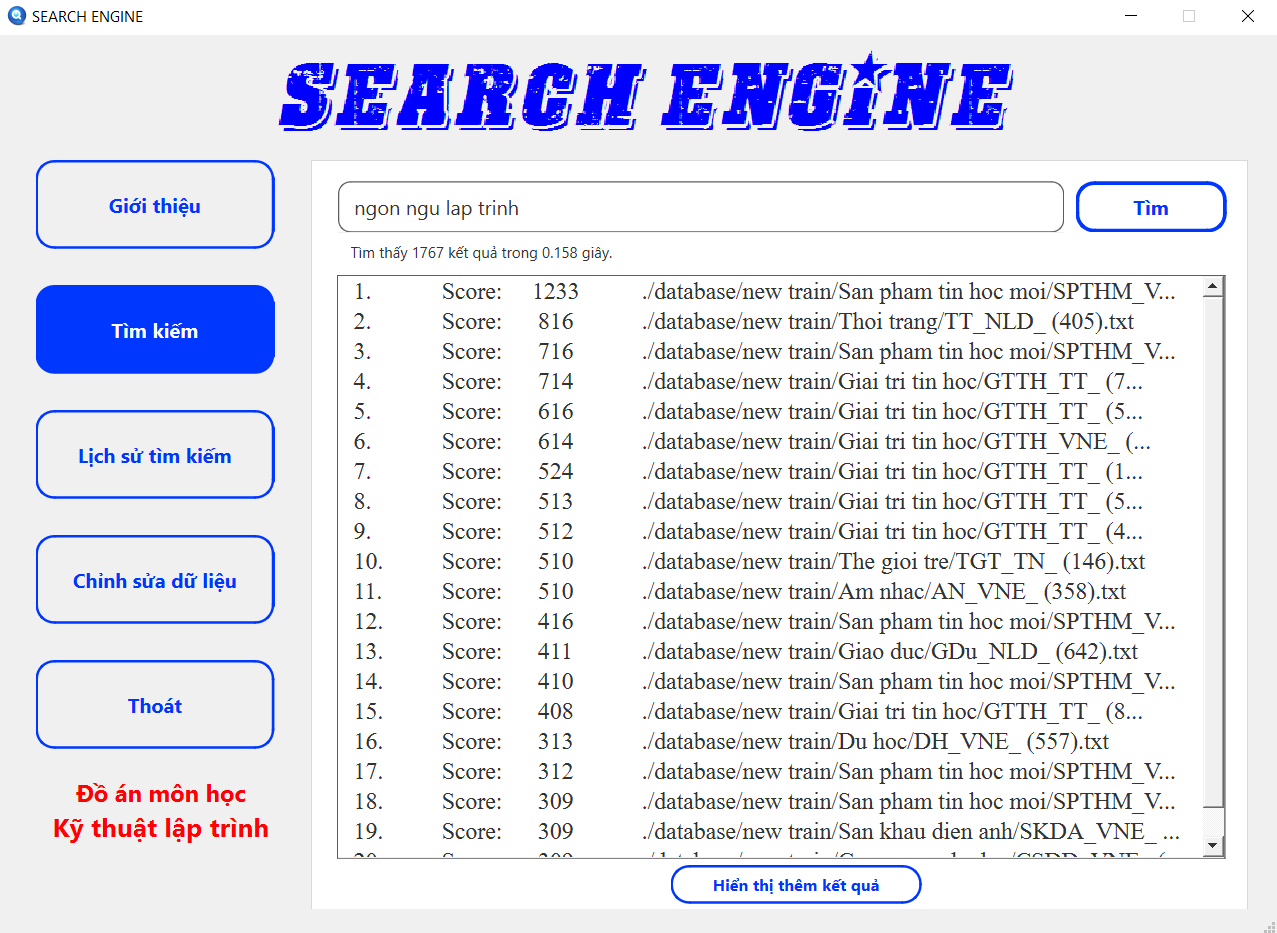
Tìm kiếm thông tin về “*ngôn ngữ lập trình*”, ta có kết quả như sau:



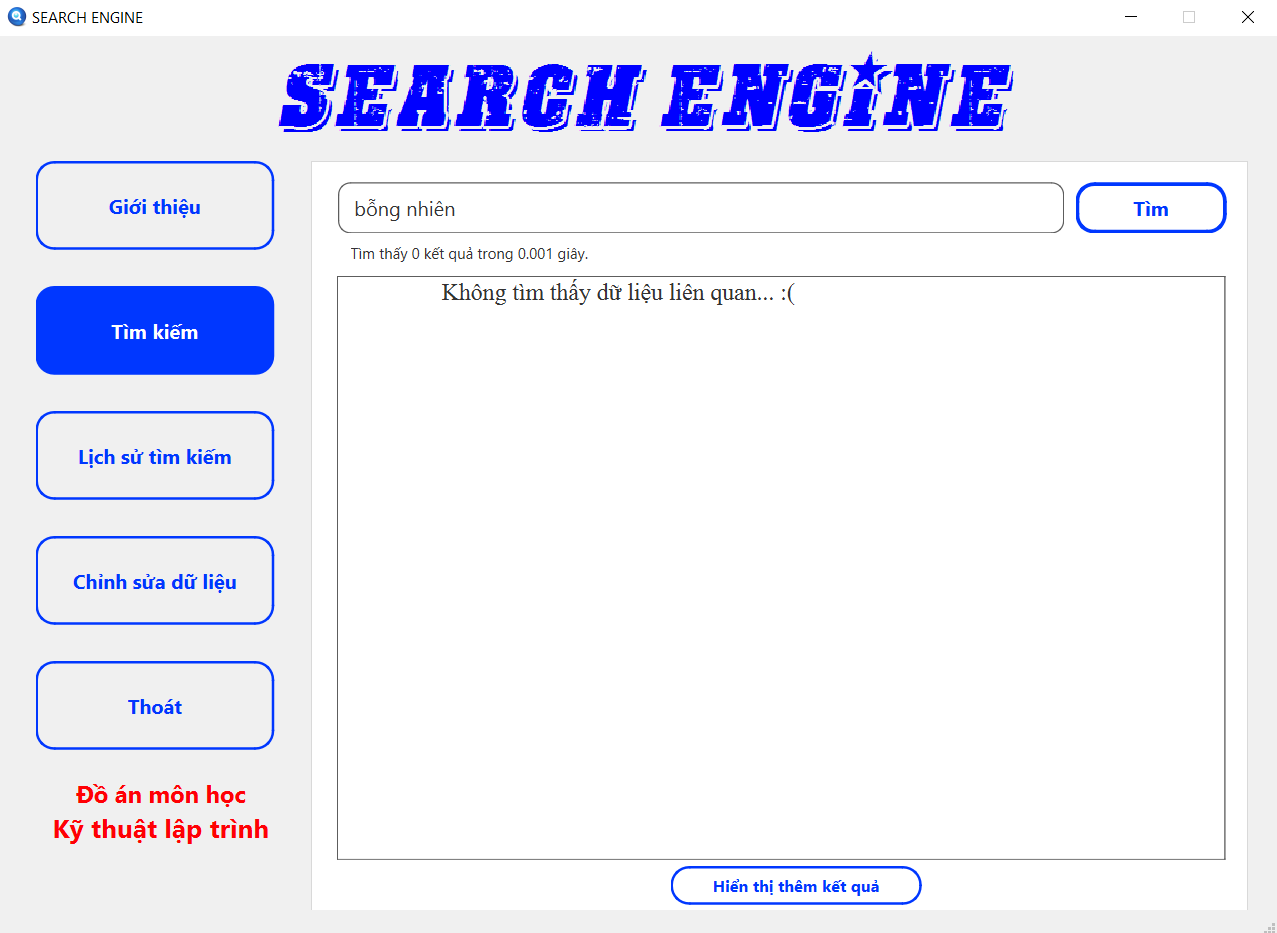
Mở file có độ liên quan cao nhất lên để xem nội dung của nó:



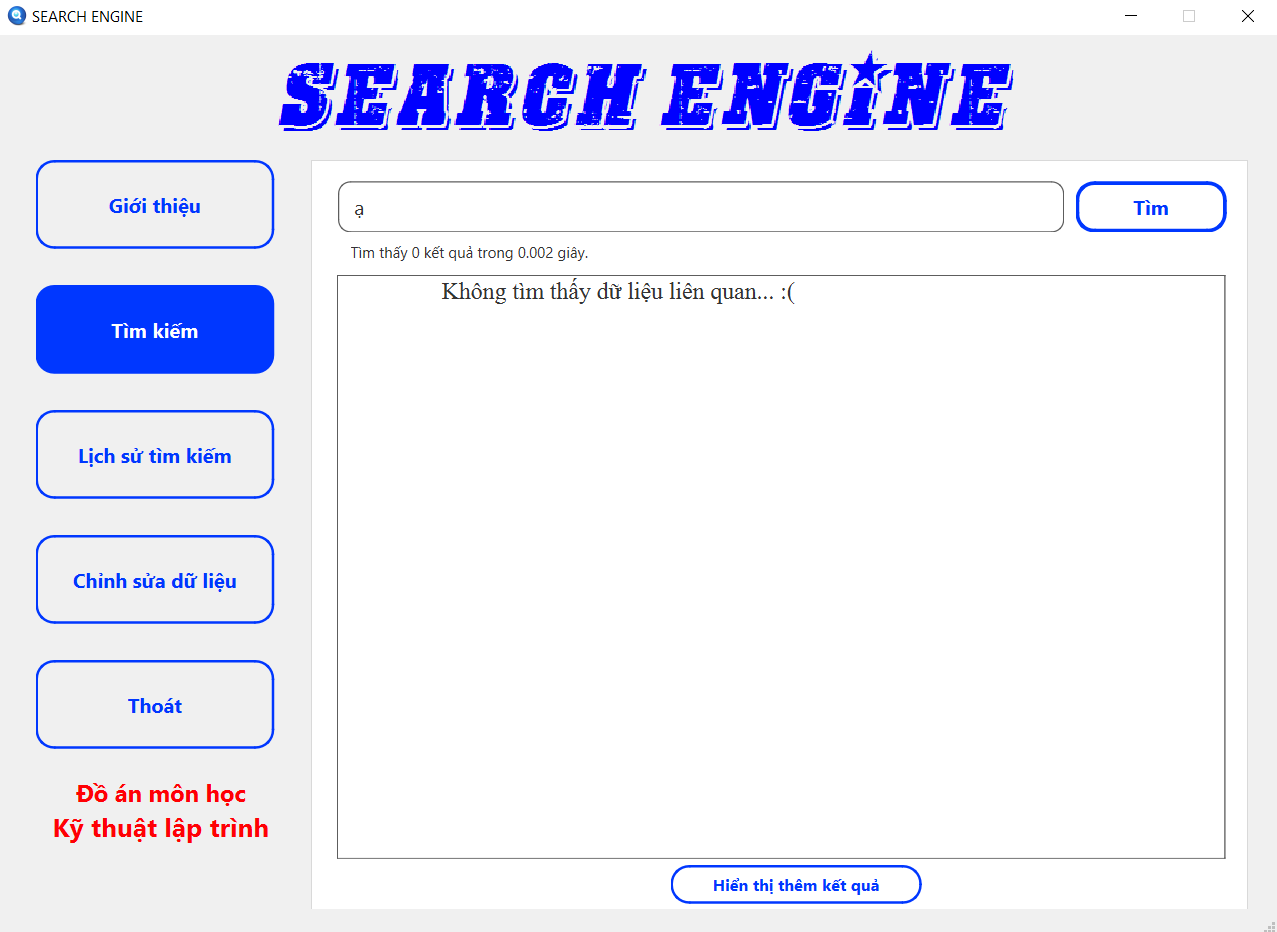
Khi nhập vào chuỗi không dấu “*ngon ngu lap trinh*”, ta cũng thu được kết quả tương tự như trên:



Khi nhập những chuỗi tìm kiếm chứa toàn những cụm từ được xếp vào ***stopword***:



*Nhập vào chuỗi “bỗng nhiên”*



*Nhập vào chuỗi “ạ”*

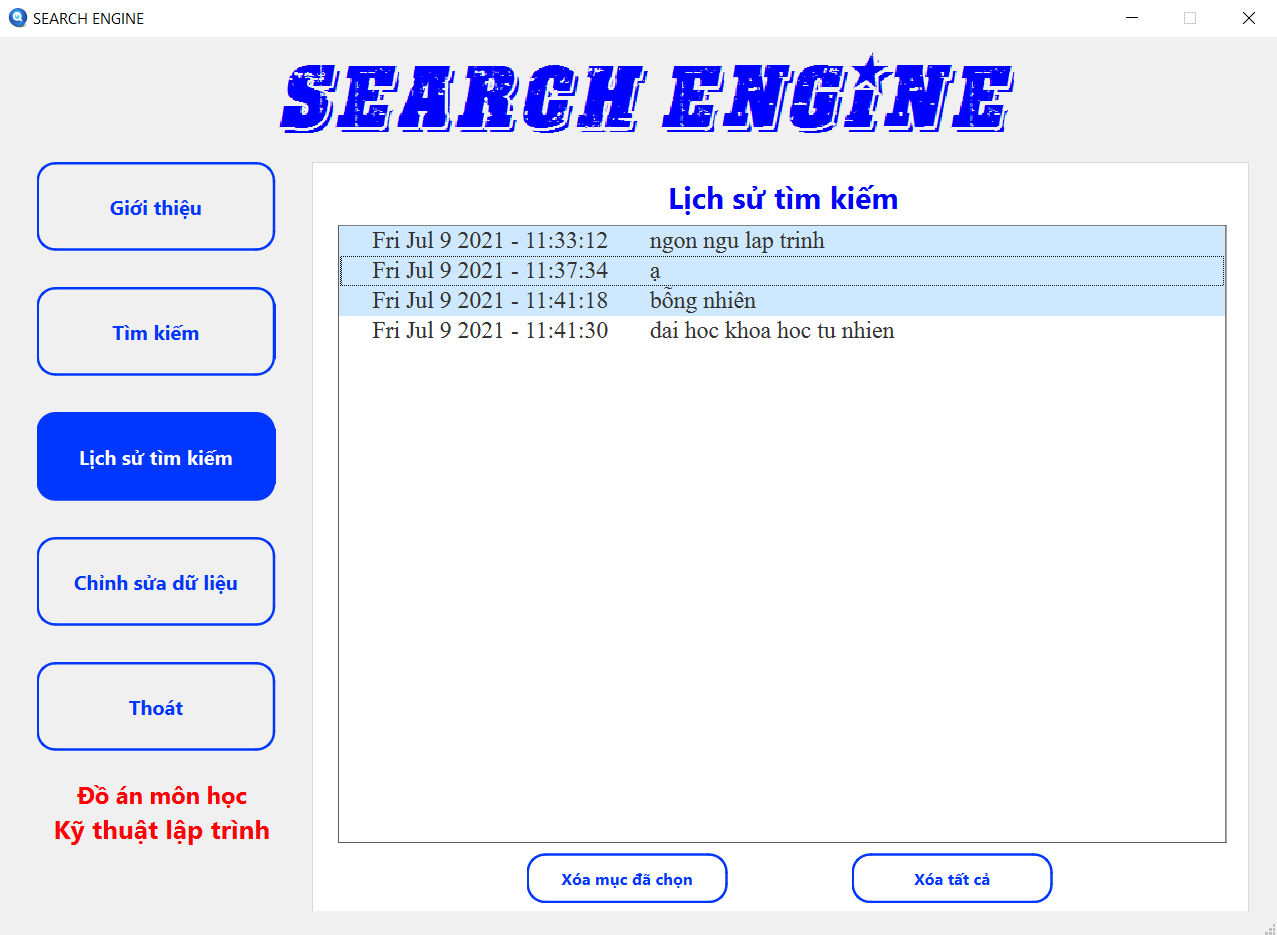
## III.2. Lịch sử tìm kiếm

Người dùng có thể xem lại lịch sử tìm kiếm của mình ***tính từ thời điểm khởi động chương trình*** với các thông tin cần thiết như ngày tháng tìm kiếm và nội dung tìm kiếm.

Chương trình cho phép xóa một vài mục lịch sử đã chọn hoặc xóa đi tất cả lịch sử.

***Hình ảnh minh họa:***

Nhấn vào nút “**Xóa mục đã chọn**” để xóa những mục lịch sử đã được chọn, hoặc nhấn vào nút “**Xóa tất cả**” để xóa toàn bộ lịch sử tìm kiếm.



## III.3. Chỉnh sửa dữ liệu

### III.3.1. Thêm, xóa file

Thao tác chỉnh sửa dữ liệu cho phép người dùng thực hiện những công việc sau:

* Thêm những file *.txt* từ máy tính của mình vào dữ liệu của chương trình để phục vụ cho nhu cầu tìm kiếm.
* Xóa đi những file mà người dùng đã thêm trước đó khỏi dữ liệu của chương trình (*thao tác này chỉ xóa file đó ra khỏi dữ liệu của chương trình chứ không ảnh hưởng gì đến file gốc của người dùng*).

**Lưu ý.** Để thao tác thêm file hoạt động được hiệu quả, có một số ràng buộc về dữ liệu như sau:

* Đường dẫn đến file muốn thêm phải là ***đường dẫn không dấu*** (không chứa bất kì từ tiếng Việt có dấu nào). Hơn nữa, đây nên là một ***đường dẫn tuyệt đối***.

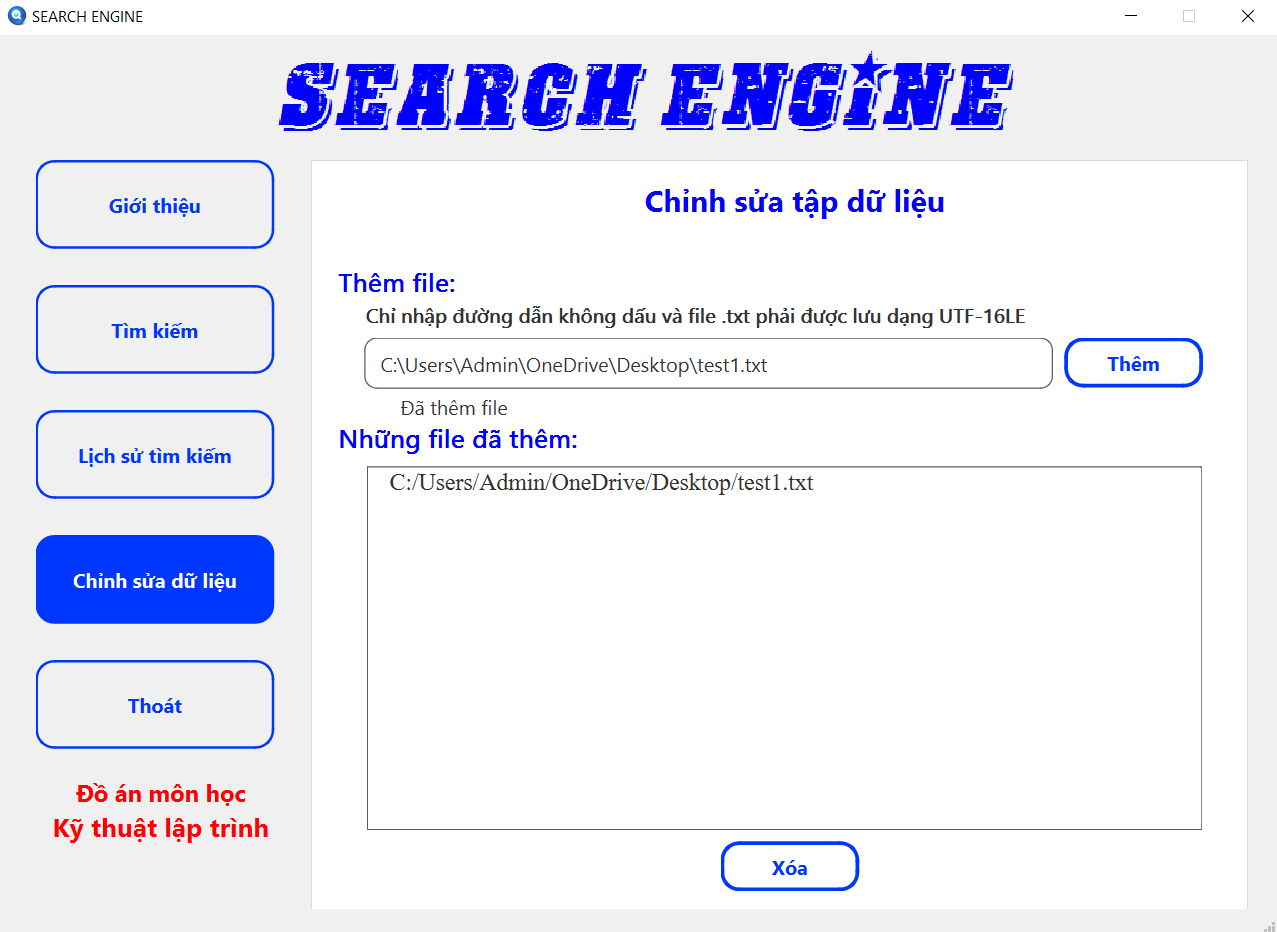
*Ví dụ:* Đường dẫn *“E:/Giải Trí/demo.txt”* là không hợp lệ.

* Các file *.txt* muốn thêm vào phải được lưu dưới dạng **UTF-16LE**.

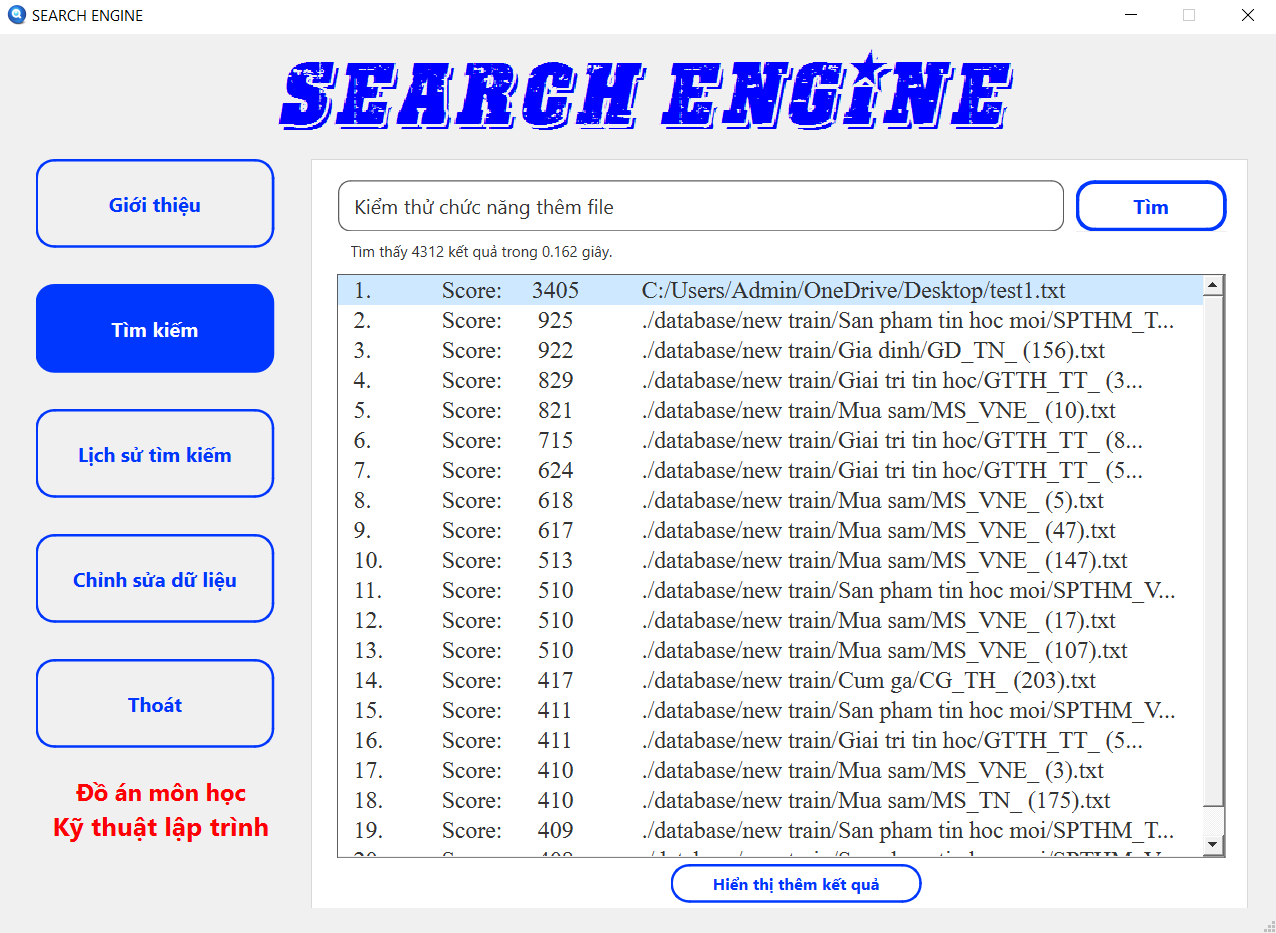
Sau khi người dùng thêm file, đường dẫn đến file đã thêm sẽ xuất hiện trong cửa sổ “**Những file đã thêm**”, giúp người dùng dễ dàng thao tác để xóa file này ra khởi dữ liệu nếu muốn.

***Một số hình ảnh minh họa:***

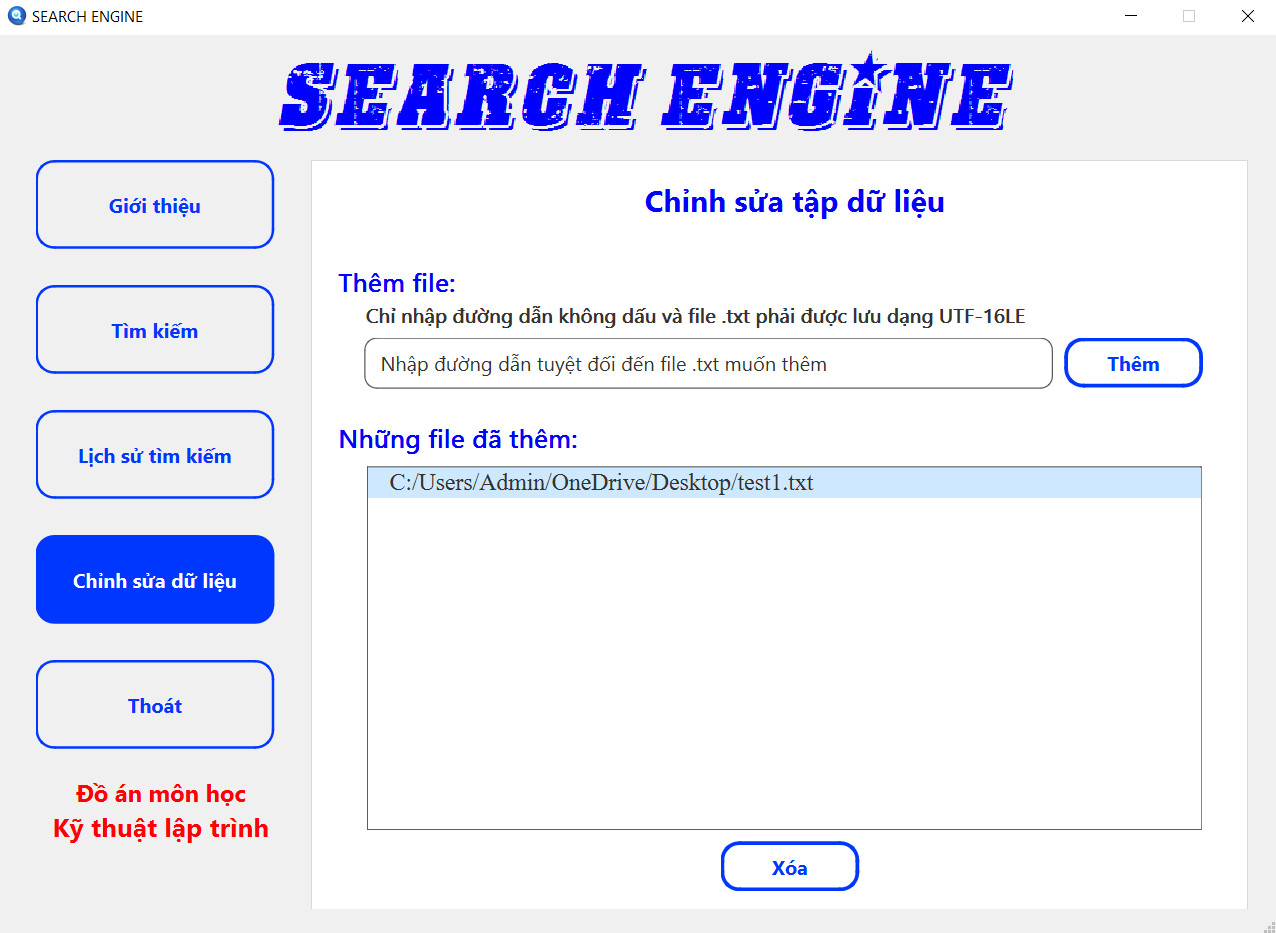
Thêm file có đường dẫn “*C:\Users\Admin\OneDrive\Desktop\test1.txt*” vào dữ liệu:



Thử tìm kiếm nội dung liên quan đến file đã thêm:

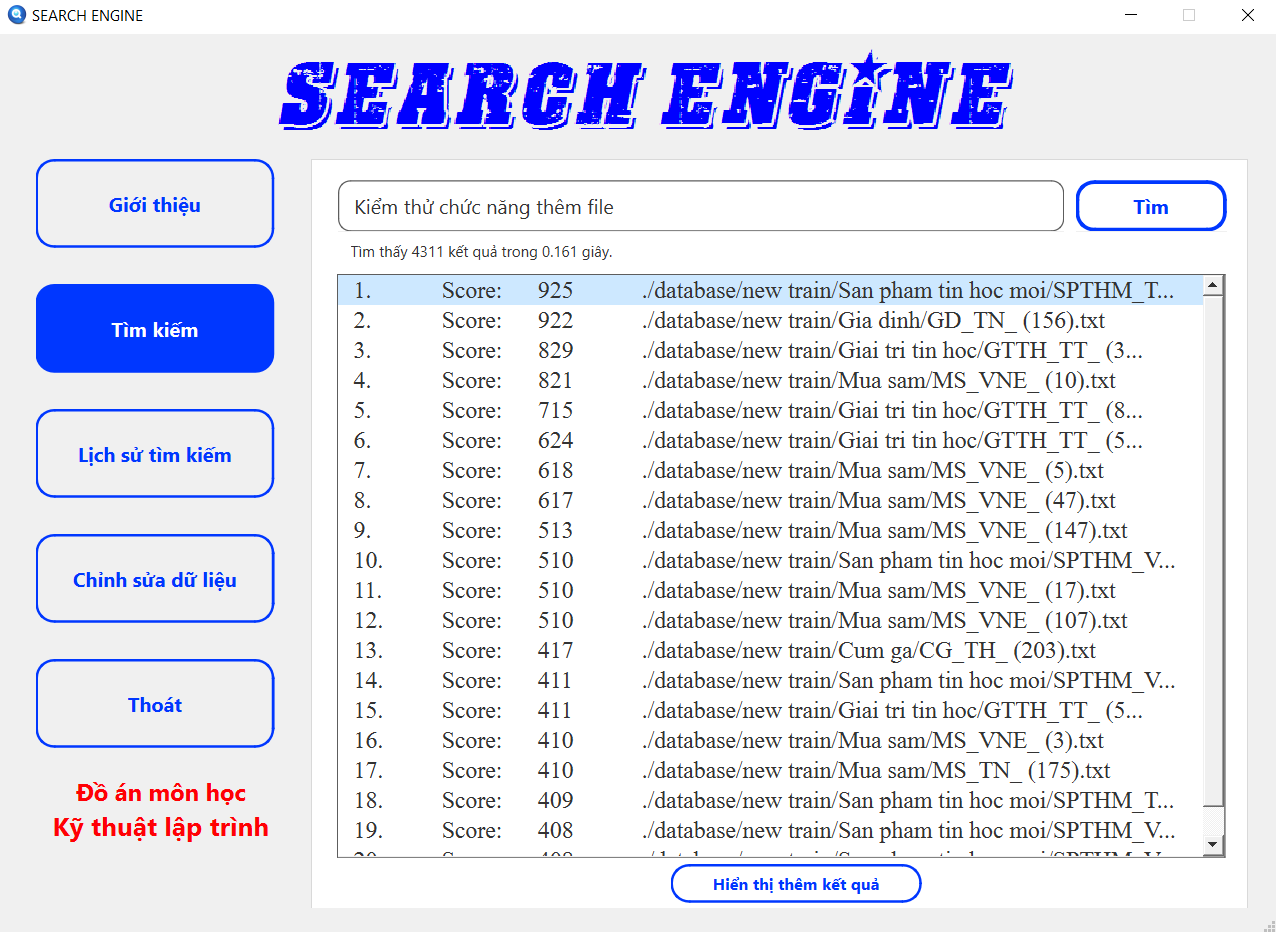


Xóa file vừa thêm khỏi dữ liệu của chương trình:



*Chọn vào đường dẫn và nhấn nút “****Xóa****” để xóa file test1.txt khỏi dữ liệu*

Thử tìm kiếm lại nội dung “*Kiểm thử chức năng thêm file*”:



Tất nhiên rằng, sau khi người dùng thêm file *.txt* của mình vào dữ liệu thì file này có thể sẽ xuất hiện trong kết quả tìm kiếm nếu nó có chứa nội dung liên quan, khi đã xóa file đó ra khỏi dữ liệu thì nó sẽ không thể xuất hiện trong kết quả tìm kiếm nữa.

Chương trình vẫn lưu lại những file người dùng đã thêm vào dữ liệu từ những lần sử dụng trước để tìm kiếm trong các lần sử dụng chương trình sau (nếu nó chưa bị người dùng xóa khỏi dữ liệu hoặc xóa khỏi máy tính của họ).

### III.3.2. Cập nhật dữ liệu

Dữ liệu của chương trình sẽ được cập nhật vào *một trong ba thời điểm sau:*

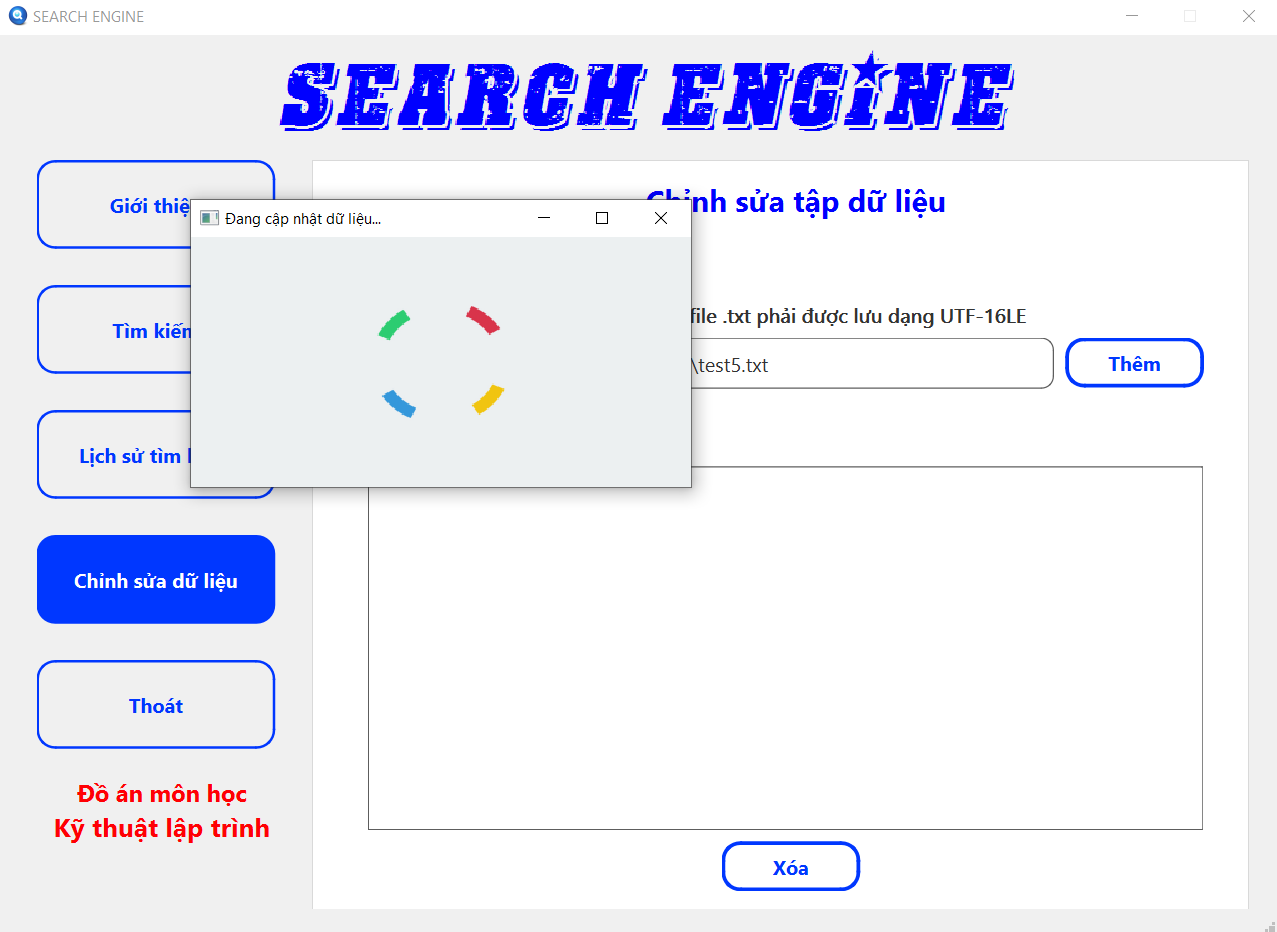
1. Khởi động chương trình.
2. Người dùng đã xóa 4 – 5 file khỏi dữ liệu của chương trình trong quá trình sử dụng.
3. Thoát khỏi chương trình.

Sau khi cập nhật xong tại thời điểm (2), chương trình sẽ thông báo thời gian cập nhật dữ liệu.

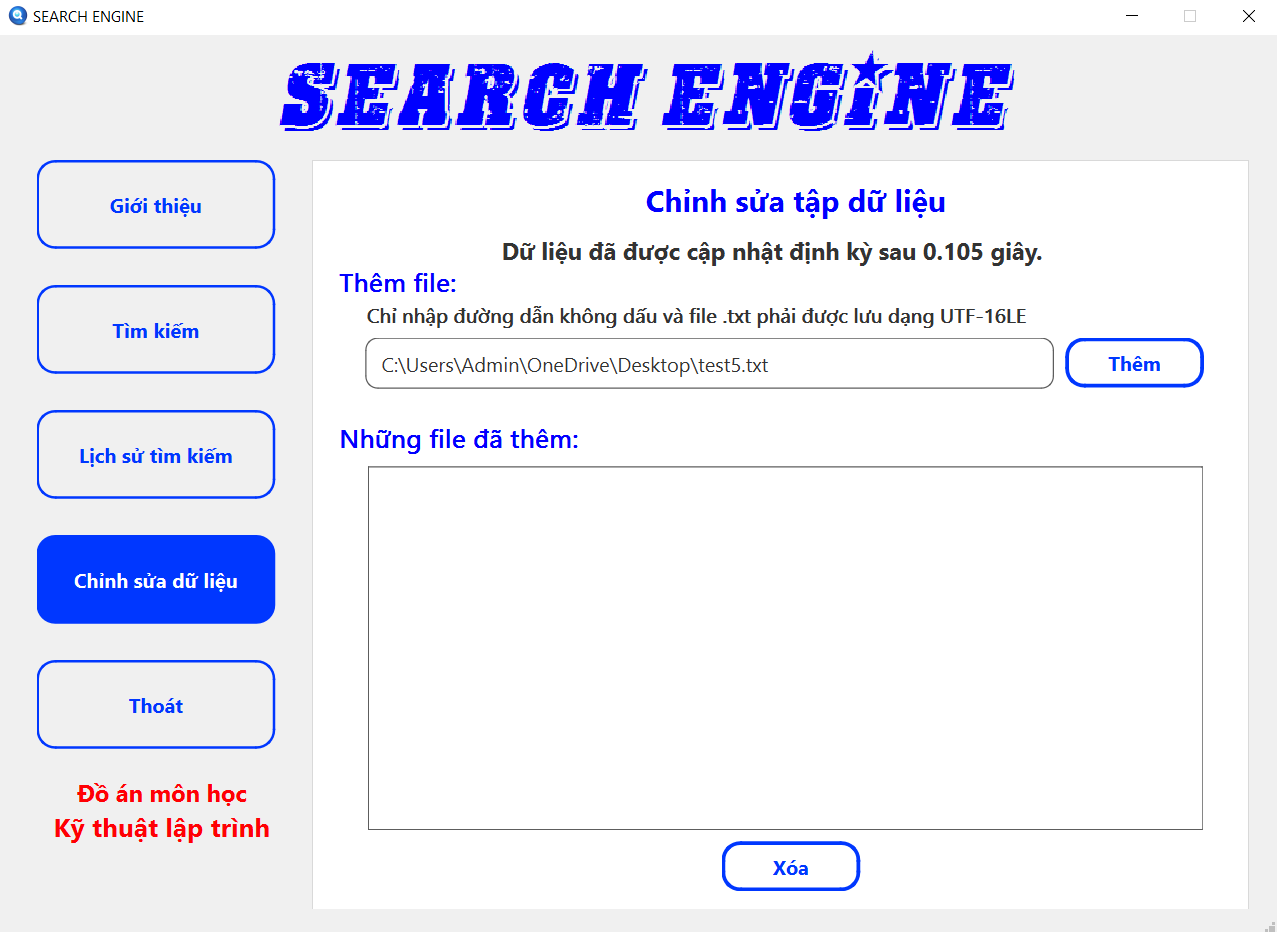
Thao tác cập nhật dữ liệu sẽ đồng bộ dữ liệu trong file *index.txt* và file siêu dữ liệu. Bên cạnh đó, trường hợp “bất thường” như sau sẽ được xử lý:

* Nếu người dùng đã thêm file vào dữ liệu chương trình rồi nhưng lại vô ý xóa nó khỏi máy tính của họ vào một thời điểm nào đó, chương trình cũng sẽ loại những thông tin, nội dung của file đó ra khỏi dữ liệu.

***Một số hình ảnh minh họa:***



*Thông báo dữ liệu đang được cập nhật*



*Sau khi cập nhật dữ liệu thành công*

# IV. Tổng kết

Chương trình **SEARCH ENGINE** của nhóm em cung cấp cho người dùng một số chức năng cơ bản như tìm kiếm, xem lịch sử tìm kiếm và chỉnh sửa dữ liệu.

Bên cạnh đó, chương trình còn có một số hạn chế như những ràng buộc về dữ liệu trong thao tác hỗ trợ người dùng thêm file *.txt* vào dữ liệu của chương trình (đường dẫn không chứa từ Tiếng Việt có dấu nào, file *.txt* phải lưu dưới dạng UTF-16LE), không xử lý được triệt để tình huống nội dung chính của chuỗi tìm kiếm mà người dùng nhập vào chứa toàn những cụm từ thuộc dãy các *stopword*, kết quả tìm kiếm đôi khi không như ý muốn của người dùng.

Một trong những hướng khắc phục trong tương lai là mở rộng thêm các định dạng khác trong tập dữ liệu ban đầu, sử dụng thêm tập từ điển Tiếng Việt cho tình huống chuỗi tìm kiếm có nội dung chính chứa toàn các cụm từ thuộc dãy *stopword*, thay đổi thuật toán rút trích nội dung chính của văn bản và tính toán độ liên quan trong quá trình tìm kiếm để cải thiện kết quả tìm kiếm.

# V. Tài liệu tham khảo

* Qt Document: <https://doc.qt.io/qt-5/index.html>
* Cấu trúc dữ liệu Skip List: <https://en.wikipedia.org/wiki/Skip_list>
* Thuật toán Radix Sort: <https://en.wikipedia.org/wiki/Radix_sort>
* Danh sách stopword Tiếng Việt: <https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords.txt>