

TÌM DỮ LIỆU văn bản cục bộ sEARCH ENIGNE local text data

Thực hành Kỹ thuật lập trình – Đồ án môn học cuối kì – Giai đoạn I

GV phụ trách: Võ Hoài Việt



20120090 – Nguyễn Thế Hoàng

20120090 – Nguyễn Văn Hưng

June 23, 2021

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CNTT – LỚP 20CTT1TN

Mục lục

[I. TỔNG QUAN 1](#_Toc75359399)

[1.1. Nội dung đồ án 1](#_Toc75359400)

[1.2. Yêu cầu đồ án 1](#_Toc75359401)

[1.3. Thông tin nhóm 2](#_Toc75359402)

[II. PHƯƠNG PHÁP 2](#_Toc75359403)

[2.1 Xử lý danh sách văn bản truy vấn 2](#_Toc75359404)

[2.2 Rút trích nội dung chính 2](#_Toc75359405)

[III. CÀI ĐẶT MÃ NGUỒN 3](#_Toc75359406)

[3.1. Môi trường lập trình 3](#_Toc75359407)

[3.2. Xử lý danh sách văn bản truy vấn 3](#_Toc75359408)

[3.2.1 Tạo danh mục văn bản truy vấn "index.txt" 3](#_Toc75359409)

[3.3 Rút trích nội dung chính 3](#_Toc75359410)

[3.3.1 Chuẩn bị Stopword 3](#_Toc75359411)

[3.4 Giao diện người dùng 3](#_Toc75359412)

[IV. HƯỚNG DẪN SỬ DỤNG 3](#_Toc75359413)

[4.1 Chương trình xử lý chính 3](#_Toc75359414)

[4.2 Giao diện người dùng 3](#_Toc75359415)

[V. PHÂN CÔNG CÔNG VIỆC 3](#_Toc75359416)

[VI. Chú giải, viết tắt 3](#_Toc75359417)

# I. TỔNG QUAN

## 1.1. Nội dung đồ án

Đồ án này thuộc danh mục sau:

* Môn học: Thực hành Kỹ thuật lập trình
* GV phụ trách: Võ Hoài Việt
* Mã số đồ án: DAMH
* Lớp: 20CTT1TN
* Loại: Đồ án cuối học kỳ

Đồ án yêu cầu thiết kế phương pháp, thuật toán; Tổ chức, xây dựng mã nguồn thực thi công việc tìm kiếm các văn bản có liên qua tới từ khoá tìm kiếm mà người dùng cung cấp đồng thời xếp hạng mức độ liên quan của văn bản đối với từ khoá đó.

## 1.2. Yêu cầu đồ án

Chiếu theo yêu cầu từ đề bài đồ án do giáo viên cung cấp, chương trình phải thoả mãn các tiêu chí sau:

* Có giao diện trực quan giúp người dùng thao tác với chương trình một cách thuận lợi, rõ ràng.
* Xử lý được các công việc liên quan đến tổ chức thư mục và tệp văn bản, cụ thể:
  + - Truy cập được tới thư mục do người dùng chỉ định
    - Truy cập được các thư mục/tệp văn bản cần tìm kiếm, truy vấn (sau đây gọi là "Thư mục/văn bản truy vấn") nằm bên trong thư mục người dùng chỉ định.
    - Hỗ trợ việc đọc, thêm, xoá thư mục/văn bản truy vấn).
* Trích xuất từ các văn bản truy vấn nội dung chính và thông tin của văn bản đó thành các tập tin siêu dữ liệu (sau đây gọi là "Metadata").
* Truy vấn trong Metadata theo từ khoá người dùng yêu cầu để trả về danh sách các văn bản truy vấn có nội dung tương thích với từ khoá đó và xếp hạng dựa theo mức độ tương thích.

Chương trình được xây dựng bằng cấu trúc dữ liệu, ngôn ngữ lập trình nằm trong phạm vi môn học cùng một số công cụ bổ trợ tự nghiên cứu. Các yêu cầu trên cùng giải pháp thực hiện sẽ được trình bày, diễn giải chi tiết trong các phần tiếp theo của tài liệu.

## 1.3. Thông tin nhóm

|  |  |  |
| --- | --- | --- |
| Họ và tên | Nguyễn Thế Hoàng | Nguyễn Văn Hưng |
| MSSV | 20120090 | 20120009 |
| Email | 20120090@student.hcmus.edu.vn | 20120009@student.hcmus.edu.vn |
| SĐT | 0901264425 | 0937142920 |

Bảng 1: Thông tin thành viên trong nhóm đồ án

# II. PHƯƠNG PHÁP

Để chương trình có thể thực thi các yêu cầu trên, mã nguồn và kiểu dữ liệu được thiết kế, xây dựng với tư tưởng sau:

## 2.1 Xử lý danh sách văn bản truy vấn

Đầu vào của chương trình là đường dẫn tới thư mục chính do người dùng chỉ định (gọi là thư mục "Main"). Thư mục *Main* chứa toàn bộ các văn bản truy vấn, các văn bản này có thể lại nằm trong các thư mục khác là thư mục con của *Main*. Bằng các lệnh hệ thống do ngôn ngữ lập trình cùng hệ điều hành cung cấp, chương trình tiến hành truy xuất tất cả các thư mục bên trong *Main* (nếu có), liệt kê các văn bản truy vấn hiện có (phải ở định dạng .txt).

Để thuận tiện cho các giai đoạn sau của chương trình, danh sách đường dẫn tuyệt đối tới các tệp văn bản truy vấn đã truy xuất như trên sẽ được lưu lại dưới dạng văn bản trong file "index.txt".

Chương trình cũng xây dựng các tính năng thêm/bớt thư mục/văn bản truy vấn theo lệnh người dùng. Sử dụng Qt xây dựng bộ khung giao diện người dùng có thể dễ dàng giải quyết yêu cầu này một cách đơn giản.

## 2.2 Rút trích nội dung chính

Tìm kiếm trong một tập hợp khổng lồ các văn bản truy vấn là cực kỳ chậm chạp và không thực tế ngay cả thuật toán so khớp chuỗi thời gian tuyến tính. Để xử lý vấn đề này, các văn vản truy vấn được rút trích chỉ giữ lại những từ ngữ đóng vai trò truyền đạt nội dung chính của văn bản. Vì vậy, kích thước không gian tìm kiếm giảm đi rất nhiều và tăng hiệu quả việc truy vấn thông tin quan trọng.

Việc xác định đâu là các từ quan trọng dựa vào hai yếu tố. Một là đó có phải là từ *Stopword* không? Hai là tần xuất xuất hiện của từ đó trong văn bản có ý nghĩa cao không? Khi loại bỏ được các từ *Stopword* và từ ý nghĩa thấp ra khỏi văn bản truy vấn sẽ giữ lại được văn bản ngắn hơn nhưng vẫn truyền đạt đủ ý chính, từ đó làm cơ sở để truy vấn.

# III. CÀI ĐẶT MÃ NGUỒN

## 3.1. Môi trường lập trình

* Trình biên dịch sử dụng: Microsoft Visual C++ (MSVC) v- 14.29
* IDE: Visual Studio 2019 Version 16.9 + Tích hợp Qt v- 6.1.1

## 3.2. Xử lý danh sách văn bản truy vấn

### 3.2.1 Tạo danh mục văn bản truy vấn "index.txt"

Hàm removeStopWord() [tạm thời] nhận tham số là chuỗi biểu diễn đường dẫn tới một thư mục. Hiện tại hàm có thể trích xuất tên/đường dẫn các văn bản truy vấn nằm bên trong thư mục đó và lưu vào file "index.txt" (vị trí "index.txt" được hiển thị trong console, nhưng sẽ luôn nằm chung thư mục với các văn bản truy vấn).

## 3.3 Rút trích nội dung chính

### 3.3.1 Chuẩn bị Stopword

Các từ là *Stopword* được tổng hợp từ nhiều nguồn và liệt kê trong file "stopwords.txt" (mỗi dòng là một từ *Stopword*). Hàm *prepare()* tiến hành mở file "stopwords.txt", dùng fgetws đọc từng dòng và đánh dấu chuỗi đọc được là *Stopword* [tạm thời]. Có thể kiểm tra việc đọc dữ liệu này bằng cách danh sách các *Stopword* đọc được vào một file văn bản khác.

## 3.4 Giao diện người dùng

Giao diện người dùng xây dựng bằng công cụ Qt. Giai đoạn hiện tại có thể hiện ra cây thư mục và danh sách các văn bản truy vấn, cũng như đọc nội dung của văn bản đó, cùng một số tính năng chưa hoàn thiện phác thảo chương trình cuối.

# IV. HƯỚNG DẪN SỬ DỤNG

Do giai đoạn hiện tại chương trình chưa hoàn thiện, mã nguồn và file thực thi được tách ra làm hai phần độc lập như sau:

## 4.1 Chương trình xử lý chính

Nằm trong thư mục /Release/Release\_Source, gọi chương trình bằng dòng lệnh không tham số (chỉ dùng tên chương trình là "main.exe"). Cần một thư mục chứa các văn bản truy vấn.

## 4.2 Giao diện người dùng

Nằm trong thư mục /Release/Release\_GUI, chạy file "Search\_Engine\_GUI.exe" để mở giao diện đồ hoạ và thực hiện thao tác cơ bản.

# V. PHÂN CÔNG CÔNG VIỆC

|  |  |  |
| --- | --- | --- |
| Họ và tên | Công việc | Mô tả |
| Nguyễn Văn Hưng |  |  |
| Nguyễn Thế Hoàng |  |  |

Bảng : Bảng phân công công việc giữa các thành viên

# VI. Chú giải, viết tắt