KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện Thoại: (08) 38.354.266 - Fax:(08) 38.350.096

CTT12 – KỸ THUẬT LẬP TRÌNH

ĐỒ ÁN MÔN HỌC

DAMH: ĐỒ ÁN

I. Thông tin chung

Mã số bài tập:	DAMH
Thời lượng dự kiến:	100 – 120 tiếng
Deadline nộp bài:	
Hình thức:	Nhóm 2 SV
Hình thức nộp bài:	Nộp qua Moodle môn học
GV phụ trách:	Võ Hoài Việt
Thông tin liên lạc với GV:	vhviet@fit.hcmus.edu.vn

II. Chuẩn đầu ra cần đạt

Bài tập này nhằm mục tiêu đạt được các chuẩn đầu ra sau:

- Cấu trúc trong C/C++
- Con trỏ, cấp phát động và quản lý vùng nhớ trong C/C++
- Xử lý tập tin văn bản, tập tin nhị phân bằng C/C++
- Sử dụng cấu trúc danh sách liên kết
- Cài đặt thuật toán sắp xếp, tìm kiếm

III. Mô tả bài tập

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.

Yêu cầu:

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.

IV. Các yêu cầu & quy định chi tiết cho bài nộp

- Bài nộp được nén .RAR hoặc .ZIP và được nộp trên moodle. Với cấu trúc tên tập tin theo tứ thự mã số sinh viên SV1_SV2.RAR hoặc SV1_SV2.ZIP (Ví dụ: 0912496_0912407.RAR)
- Cấu trúc thư mục nộp bài gồm như sau:
  - Documents: chứa báo cáo đồ án, slide trình bày nội dung, bảng phân công thành viên và hướng dẫn sử dụng
  - Release: chứa tập tin exe và các dữ liệu cần thiết để chạy chương trình.
  - Program: file thực thi và mã nguồn chương trình
  - Demo: chứa video demo các chức năng của chương trình.

V. Gợi ý làm bài

Danh sách văn bản:

- Tập tin index.txt cùng thư mục với các tập tin văn bản, chứa danh sách các tập tin tin văn bản có trong thư mục, mỗi dòng tương ứng với một tập tin.
- Chương trình nhận được đường dẫn đến thư mục chứa chỉ mục và các văn bản bằng nhập chuẩn hoặc tham số hàm main.

Rút trích nội dung chính:

- Nếu tìm kiếm những văn bản bằng cách tìm trong các tập tin văn bản một cách tìm trong từng văn bản thì phải tốn rất nhiều thời gian do đó các văn bản được rút trích nội dung chính để lưu lại vào tập tin siêu dữ liệu.
- Nội dung chính được xác định dựa trên các từ quan trọng xuất hiện trong văn bản. Độ quan trọng của một từ được tính bằng tần suất xuất hiện của từ đó trong văn bản chia cho tổng số từ có trong văn bản. Tỷ lệ này nằm trong khoảng (0, 1). Sinh viên tự chọn giá trị a và b sao cho nếu một từ có độ quan trọng nằm trong đoạn [a, b] thì từ đó là một từ quan trọng. Giải thích lý do chọn giá trị a, b trong báo cáo.
- Một số từ có tần suất xuất hiện cao như: a, à... (sinh viên tự xác định hoặc sử dụng link phần tham khảo) gọi là stopword sẽ bị loại bỏ khỏi danh sách từ quan trọng của văn bản do không có ý nghĩa trong
việc tìm kiếm.

Tổ chức siêu dữ liệu:

- Phát sinh tập tin siêu dữ liệu dựa vào các văn bản có trong index.txt
- Khi thêm/xóa văn bản trong index.txt, chương trình tự động cập nhật tập tin siêu văn bản, tính thời gian cập nhật siêu dữ liệu. Sinh viên tự xây dựng cấu trúc siêu dữ liệu sao cho các thao tác thêm/xóa các văn bản, siêu dữ liệu được cập nhật nhanh chóng.

Truy vấn thông tin:

- Cho phép người dùng nhập vào một chuỗi từ khóa. Kết quả trả về các văn bản liên quan xếp hạng giảm dần theo mức độ liên quan.
- Để tìm được các văn bản liên quan, người ta dựa trên độ tương tự của văn bản với các từ khóa nhập vào. Độ tương tự của văn bản với từ khóa được tính bằng cách:
  - Tìm tập giao tập từ quan trọng của văn bản với tập từ khóa, nếu có thì văn bản được xem là có liên quan.
  - Số lượng từ giao càng nhiều thì độ liên quan càng cao, văn bản trong kết quả được xếp hạng càng cao
- Kết quả trả về bao gồm: thứ hạng, tựa đề văn bản, tên tập tin văn bản, mức độ liên quan.
- Số lượng văn bản có thể rất nhiều do đó các thao tác tìm kiếm, sắp xếp cần được tối ưu để tăng tốc việc truy vấn. Các nhóm sinh viên so sánh với nhau thời gian truy vấn, các nhóm đạt có kết quả truy vấn càng tốt và thời gian truy vấn càng ngắn sẽ được cộng điểm

VI. Cách đánh giá

STT	Tên kết quả	Tỉ lệ điểm	Ghi chú
1	Mã nguồn	40%	Cung cấp các thư viện và mã nguồn đầy đủ để biên dịch.
2	Phong cách lập trình	20%	Cấu trúc chương trình rõ ràng, hàm/ biến đặt tên dễ hiểu và gợi nhớ và tuân thủ các qui tắc lập trình.
3	Báo cáo, demo và hướng dẫn sử dụng	20%	Báo cáo các chức năng của chương trình, hướng dẫn sử dụng cho người dùng, video minh họa từng bước và dữ liệu test
4	Slide trình bày	20%	Trình bày nội dung thực hiện, thuật toán, cấu trúc dữ liệu liên quan, kết quả thực nghiệm

VII. Tài liệu tham khảo

Slide bài giảng lý thuyết
Sinh viên tự nghiên cứu.

https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords.txt (Stopwords)

http://infolab.stanford.edu/~backrub/google.html

https://drive.google.com/drive/folders/1yJ_cM5g69S7GpfxX6ULTgjEn52fqiKFF (Tập dữ liệu)

VIII. Các quy định khác

- Chương trình phải có hướng dẫn sử dụng (Không có hướng dẫn sử dụng sẽ bị trừ 50% số điểm của phần phần chương trình).
- Tất cả các bài làm sai quy định đều bị 0 điểm cho mỗi bài.
- Hai bài giống nhau từ 80% trở lên sẽ bị 0 điểm cho cả hai bất kể ai là tác giả.

- Các trường hợp sử dụng mã nguồn không ghi rõ nguồn tham khảo sẽ bị điểm 0 cho tất các các bài và các tác giả).
- Các bài làm xuất sắc sẽ được điểm cộng.
- Không nhận bài nộp trễ qua mail.