Bài Tập Index And Query

1. Mô tả

- Để làm được bài thực hành này, bạn cần hoàn thành thao tác xử lý tách từ ở các bài tập trước.
- Các yêu cầu trong bài này sẽ có mức độ khó tăng dần, câu sau có sử dụng kết quả của câu trước, vì vậy bạn cần hoàn thành các câu theo thứ tự.

2. Yêu cầu

Cho danh sách các tập tài liệu như sau:

Doc 1: Web mining is useful.

Doc 2: Usage mining applications.

Doc 3: Web structure mining studies the Web hyperlink structure.

a) Viết chương trình đọc tất cả các tập tài liệu đã có, thực hiện tách từ, loại bỏ stopword để tạo ra tập ngữ vựng V. Sắp xếp tập ngữ vựng V theo thứ tự alphabet và ghi vào file kết quả tapNguVungV.txt.
   Ví dụ: Với các tập tài liệu như trên, tập ngữ vựng V sẽ là {Web, mining, useful, applications, usage, structure, studies, hyperlink}
   Tập ngữ vựng V sau khi sắp xếp sẽ là:
   V = {applications, hyperlink, mining, structure, studies, usage, useful, web}
b) Viết chương trình xử lý sử dụng tập ngữ vựng V và tập các tài liệu ban đầu, tạo ra chỉ mục đảo như mô tả sau:

Từ trong tập V	Tập tài liệu có xuất hiện từ	Số lần từ xuất hiện trong tập tài liệu

Ví dụ:

applications: 2: 1 // tài liệu số 2, xuất hiện 1 lần

hyperlink: 3: 1 // tài liệu số 3, xuất hiện 1 lần

mining: 1: 1, 2: 1, 3: 1 // tài liệu số 1: 1 lần, tài liệu số 2: 1 lần, tài liệu số 3: 1 lần

structure: 3: 2 // tài liệu số 3: 2 lần

Ghi kết quả chỉ mục đảo vào file kết quả.

c) Với chỉ mục đảo đã có, ta tiến hành truy vấn thông tin. Yêu cầu kết quả trả về phải sắp xếp các tài liệu theo thứ tự tần số xuất hiện của từ trong tài liệu giảm dần. Kết quả trả về sẽ có dạng như sau:

Tập tài liệu có xuất hiện từ	Số lần từ xuất hiện trong tập tài liệu

Ta sẽ có các loại truy vấn khác nhau:

- Nhập vào 1 từ
- Nhập vào 2 từ
- Nhập vào 3 từ

Ví dụ:

- Nhập vào 1 từ: web
  - Kết quả trả về sẽ là: 3: 2, 1: 1 // tài liệu số 3: 2 lần, tài liệu số 1: 1 lần

- Nhập vào 2 từ web mining, kết quả trả về sẽ là: 3: 3, 1: 2 // tài liệu số 3: tổng xuất hiện của 2 từ là 3 lần, tài liệu số 1: tổng xuất hiện của 2 từ là 1 lần.

3. Lưu ý:

Các bài được nộp lên moodle phải do chính bạn làm, nếu phát hiện gian lận, copy bài của nhau sẽ bị 0 điểm thực hành.