Bài tập thực hành:

Indexing and Query

Thời gian làm bài: 1 tuần (Xem dealine trong link nộp bài trên moodle)

Nộp bài:

- Nộp bài lên moodle.
- Đặt tên bài nộp theo định dạng MSSV.rar. Trong đó bao gồm:
  - Tập tin báo cáo.
  - Các tập tin dữ liệu theo yêu cầu của bài tập
- Nếu sử dụng code trên Internet, trước đoạn code sử dụng phải có chú thích đường dẫn đến trang web chứa đoạn code. Trường hợp 2 bài có đoạn code giống nhau, bài không chú thích đường dẫn đến trang web chứa source code sẽ bị tính là gian lận, và bị 0 điểm.

Các hành vi sử dụng toàn bộ/một phần bài làm của người khác sẽ bị 0 điểm cho toàn bộ phần thực hành

1 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:

application: 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 ít nhất 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: 1: 1, 3: 1 // tài liệu số 1: từ xuất hiện ít nhất là web và mining đều là 1 lần, tài liệu số 3: từ xuất hiện ít nhất là mining với 1 lần.

2 Bài tập

Chính sửa đoạn chương trình được giao sao cho chương trình thực hiện đúng các yêu cầu của đề:

- Thêm các thư viện OpenNLP vào project
- Chỉnh sửa/ viết thêm code vào những chỗ //YOUR CODE HERE trong mã nguồn chương trình.

Viết báo cáo:

- Trình bày logic thực thi của chương trình:
  - Từng bước thực hiện từ tập cơ sở dữ liệu thô đến kết quả.
  - Từng bước được thực hiện bằng những hàm nào?
  - Các dòng code trong các hàm thực hiện công việc gì?
  - …
- Đề xuất cải tiến (nếu có) trong các đoạn code của chương trình.