Thực hành môn Truy hồi thông tin và tìm kiếm web

Buổi 1. Xử lý văn bản

Phần I. Cài đặt phần mềm

1. Tìm bộ cài Anaconda trên web rồi cài đặt lên máy tính để chuẩn bị môi trường lập trình Python.
2. Cài đặt thư viện xử lý ngôn ngữ NLTK: Gõ lệnh "`pip install nltk`" ở dấu nhắc Anaconda để cài thư viện, sau đó gõ lệnh "`import nltk`" và lệnh "`nltk.download('popular')`" ở dấu nhắc Python để cài dữ liệu ngôn ngữ đi kèm.

Phần II. Bài tập lập trình

Bài 1. [Tách từ tiếng Anh] Dùng thư viện NLTK để thực hành tách các từ trong một tài liệu tiếng Anh (dùng hàm `nltk.word_tokenize`(<xâu tài liệu>)) và tách gốc của các từ tiếng Anh bằng thuật toán Porter (dùng hàm `nltk.PorterStemmer().stem`(<xâu từ>)).

Bài 2. [Tách từ tiếng Việt] Nhập một tài liệu d gồm một hoặc nhiều câu tiếng Việt. Nhập tiếp một từ t rồi kiểm tra xem từ t có mặt trong văn bản d hay không. Chú ý: Sinh viên tự tìm một thư viện tách từ tiếng Việt trên web (ví dụ: https://github.com/trungtv/pyvi) để dùng trong bài tập này và các bài tập khác. Một số hàm Python hữu ích: Hàm `split` tách ra và trả về một danh sách các xâu con dựa trên dấu cách, hàm `lower` trả về xâu chữ thường, toán tử `in` kiểm tra một phần tử có xuất hiện trong một danh sách hay không.

Bài 3. [Tần số từ] Nhập một tài liệu d. Tách ra các từ riêng biệt trong tài liệu d rồi đếm xem mỗi từ đó xuất hiện bao nhiêu lần. Gợi ý: Dùng kiểu từ điển dict trong Python.

Bài 4. [Từ vựng và tần số tài liệu] Cho một tập tài liệu D gồm 3-4 tài liệu hoặc nhiều hơn, mỗi tài liệu chỉ gồm một câu (Hãy gõ trực tiếp tập D vào trong code, tức là không cần nhập từ bàn phím khi chạy chương trình). Tách ra các từ riêng biệt trong tập tài liệu D và đếm xem mỗi từ đó xuất hiện trong bao nhiêu tài liệu khác nhau (Ví dụ, nếu tập D gồm 10 tài liệu thì kết quả đếm sẽ nằm trong khoảng từ 1 đến 10).

Bài 5. [Danh sách thẻ định vị] Nhập hai danh sách số nguyên đã sắp xếp tăng dần. Tìm giao và hợp của hai danh sách đó.

Bài 6. [Truy vấn] Thực hiện lần lượt các yêu cầu sau:

1. Tạo 10 file văn bản (mỗi file là một tài liệu) và đặt vào cùng một thư mục. Nội dung mỗi file tự nghĩ ra hoặc tìm trên web. Đặt tên file phản ánh nội dung file.

2. Mở từng file (`f=open(<tên file>,'r'`), đọc nội dung file (`f.read()`) rồi lưu vào bộ nhớ.
3. Tách ra các từ riêng biệt từ tập tài liệu.
4. Lưu các từ đã tách vào một biến kiểu từ điển. Mỗi phần tử trong từ điển là một cặp khóa-giá trị, trong đó khóa là từ, còn giá trị là danh sách số thứ tự của các tài liệu có chứa từ đó.
5. Người dùng gõ vào một từ truy vấn thì tìm (số thứ tự của) các tài liệu chứa từ đó trong từ điển rồi trả về.
6. Người dùng chọn (số thứ tự của) tài liệu nào thì hiện toàn văn tài liệu đó lên màn hình.