Thực hành môn Truy hồi thông tin và tìm kiếm web

Buổi 2. Xây dựng chỉ mục ngược

Phần I. Cài đặt thư viện tìm kiếm Whoosh

- Gõ lệnh "`pip install whoosh`" ở dấu nhắc Anaconda.
- Tài liệu hướng dẫn sử dụng thư viện: https://whoosh.readthedocs.io/en/latest

Phần II. Đọc hướng dẫn sử dụng thư viện Whoosh

1. Đối tượng chỉ mục ngược `Index`

- Trước khi tạo chỉ mục ngược, cần có đối tượng `Schema` dùng để biểu diễn lược đồ của chỉ mục ngược, chỉ rõ mỗi tài liệu gồm những trường nào, như tiêu đề, tóm tắt, nội dung… Ví dụ, tạo lược đồ có hai trường `title` và `content`:

```
from whoosh.fields import Schema, TEXT
schema = Schema(title=TEXT, content=TEXT)
```

  Ở đây, `title` và `content` đều có kiểu là `TEXT`, tức là văn bản thông thường.

- Tạo chỉ mục ngược trong thư mục `index` và có lược đồ `schema`:

```
from whoosh.index import create_in
ix = create_in("index", schema) # Trả về đối tượng Index
```

2. Tạo và mở chỉ mục ngược

- Tạo chỉ mục mới:

```
import os
from whoosh.index import create_in
if not os.path.exists("index"):
    os.mkdir("index") # Tạo thư mục nếu nó chưa tồn tại
ix = create_in("index", schema)
```

- Mở chỉ mục đã có:

```
from whoosh.index import open_dir
ix = open_dir("index")
```

- Một số lệnh hữu ích:
  - Xem thư mục hiện hành là thư mục nào: `os.getcwd()`
  - Xem nội dung thư mục hiện hành: `os.listdir()`
  - Thay đổi thư mục hiện hành: `os.chdir(<thư mục mới>)`

3. Xây dựng chỉ mục ngược dùng đối tượng `IndexWriter`

```
writer = ix.writer() # Trả về đối tượng IndexWriter
writer.add_document(title="My document",
                    content="This is my document!")
writer.add_document(title="Second try",
                    content="This is the second example.")
writer.add_document(title="Third time's the charm",
                    content="Examples are many.")
writer.commit() # Phải gọi hàm commit để kết thúc
```

4. Tìm kiếm dùng đối tượng Searcher

- Lấy về đối tượng `Searcher`:

```
searcher = ix.searcher()
```

- Dùng cú pháp `with` nếu muốn các file đang mở được đóng tự động sau khi tìm kiếm xong:

```
with ix.searcher() as searcher:
    <viết các câu lệnh tìm kiếm ở đây>
```

- Gọi hàm `search` để xử lý câu truy vấn và nhận về kết quả:

```
results = searcher.search(<đối tượng truy vấn>)
```

5. Tạo đối tượng truy vấn `Query`

- Dùng đối tượng `QueryParser` để phân tích một xâu truy vấn thành một đối tượng truy vấn `Query`:

```
from whoosh.qparser import QueryParser
# "content" là trường tìm kiếm ngầm định
parser = QueryParser("content", ix.schema)
myquery = parser.parse(<xâu truy vấn>)
```

Phần III. Bài tập lập trình

Bài 1. Đọc hiểu rồi chạy thử chương trình sau đây:

```
from whoosh.index import create_in
from whoosh.fields import *
# "stored=True" nghĩa là sẽ lưu trữ trường này trong chỉ mục;
# kiểu trường "ID" nghĩa là không đánh chỉ mục trên trường này.
```

```
schema = Schema(title=TEXT(stored=True), path=ID(stored=True),
                content=TEXT)
# Đảm bảo thư mục "indexdir" tồn tại trong thư mục hiện hành
ix = create_in("indexdir", schema)
writer = ix.writer()
writer.add_document(title="First document", path="/a",
    content="This is the first document we've added!")
writer.add_document(title="Second document", path="/b",
    content="The second one is even more interesting!")
writer.commit()
from whoosh.qparser import QueryParser
with ix.searcher() as searcher:
    query = QueryParser("content", ix.schema).parse("first")
    results = searcher.search(query)
    print("Number of results:", len(results))
    print(results[0]["title"])
```

Bài 2. Thực hiện lần lượt các yêu cầu sau:

1. Gõ trực tiếp một tập tài liệu D gồm 10 tài liệu vào trong code. Mỗi tài liệu gồm tiêu đề và nội dung. Có thể tạo hai danh sách: một danh sách lưu các tiêu đề và danh sách kia lưu các nội dung tương ứng.
2. Xây dựng chỉ mục ngược trên tập tài liệu D, trong đó có lưu mã tài liệu (chỉ số của phần tử tương ứng trong danh sách ở bước 1) và có lưu phần tiêu đề vào chỉ mục, nhưng không lưu phần nội dung (vì kích thước lớn).
3. Nhập vào (từ bàn phím) một câu truy vấn gồm nhiều từ.
4. Xử lý câu truy vấn và hiển thị kết quả theo quy cách: dòng đầu tiên là số kết quả, còn mỗi dòng tiếp theo gồm số thứ tự (bắt đầu từ 1) và phần tiêu đề của mỗi kết quả.
5. Nhập vào (từ bàn phím) số thứ tự của một kết quả.
6. Hiển thị phần nội dung của kết quả đã chọn (dùng mã tài liệu có lưu trong kết quả để truy nhập danh sách ở bước 1).