Báo cáo bài tập 2 phần 2:

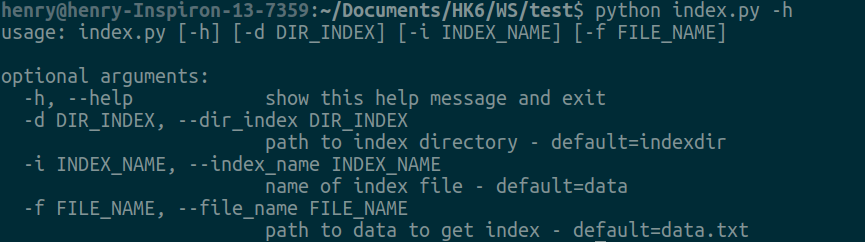
Thành viên:

|  |  |  |
| --- | --- | --- |
| MSSV | Họ và tên | Email |
| 1412161 | Lưu Đức Hiếu | [henryhieu1996@gmail.com](mailto:henryhieu1996@gmail.com) |
| 1412498 | Nguyễn Minh Thành |  |

Ngôn ngữ lập trình:

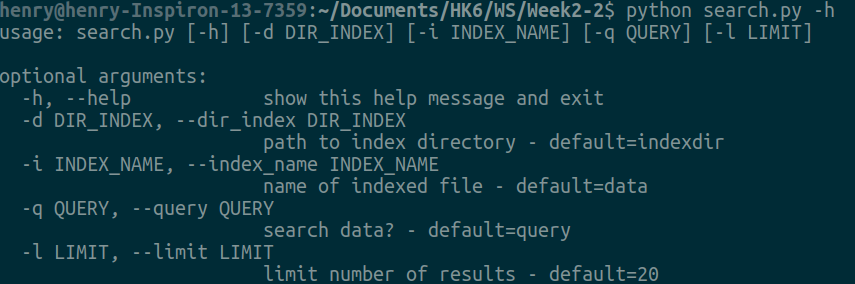
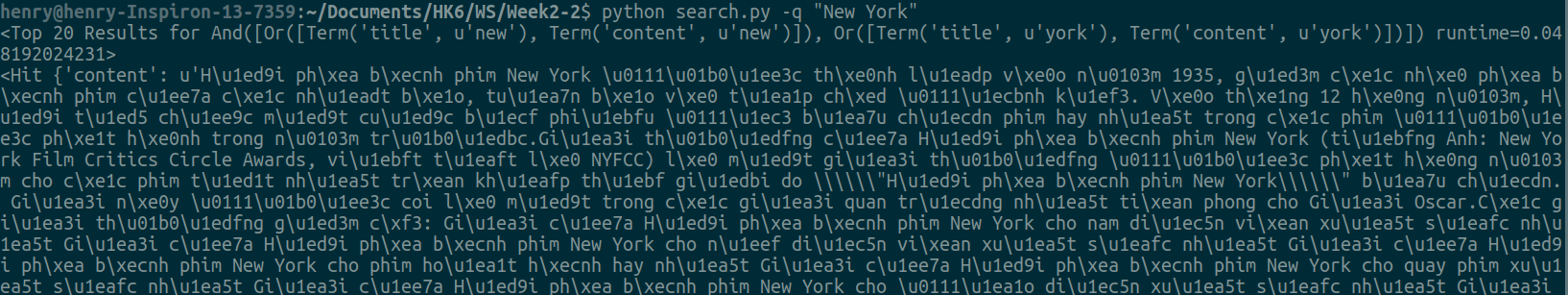
* + Python

Cách sử dụng chương trình: index.py, câu lênh: “python index.py -h” để ra hướng dẫn dùng

* 
  + Để dùng chương trình với các tham số mặc định:
* 

--> Thời gian **index: 1215.978s**

Cách sử dụng chương trình search.py, câu lênh: “python search.py -h” để ra hướng dẫn dùng:

* + 
    - Search với từ khóa “New York”: **thời gian chạy là 0.04s**
  + 
    - Search với từ khóa “Pennsylvania”: **thời gian chạy là 0.019s**
    - 

Cách thực hiện bài tập:

Quá trình index:

* Ở bài tập này, nhóm em dùng thư viện Whoosh, một thư viện hỗ trợ cho ngôn ngữ lập trình python.
* Đầu tiên, nhóm em mở file có kí tự unicode bằng codes, và đọc từng dòng thay vị đọc nguyên file vào bộ nhớ. Đọc từng dòng và tách 1 dòng thành mảng 3 phần tử bởi dấu tab “\t”
* Kế tiếp nhóm tạo 1 kiểu dữ liệu schema: với 3 phần tử đều có thuộc tính stored=True để hiển bị lúc bị truy vấn



* Sau khi thêm toàn bộ file dữ liệu vào bộ index thì ta bắt đầu quá trình index.
* Thời gian index khá nhanh: tầm 1200s ~ 20phút, trong lúc index thì các bộ dữ liệu tạm thời sinh ra chiếm khoảng 2.4gb cho tập tin data đầu vào 320mb.
* Sau khi quá trình hoàn tất. Bộ index được gôm lại còn 900mb.

Quá trình search:

* Sau khi có được bộ index thì ta dùng thư viện Whoosh để mở file và bắt đầu tìm kiếm.
* Nhóm cho tìm kiếm từ khóa trong 2 trường là title và dữ liệu đã được tóm tắt trong tập tin data.txt
* Thời gian search khá là nhanh <0.05s.