TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

Search Engine for Truyện Kiều

Môn: Truy vấn thông tin Thị giác

Giáo viên hướng dẫn: Võ Hoài Việt
Phạm Minh Hoàng
Phạm Thanh Tùng

Lớp: 20TGMT

Sinh viên thực hiện: Giang Gia Bảo – 20127446

Tp Hồ Chí Minh, 22 tháng 06 năm 2023

MỤC LỤC

I. THÔNG TIN SINH VIÊN

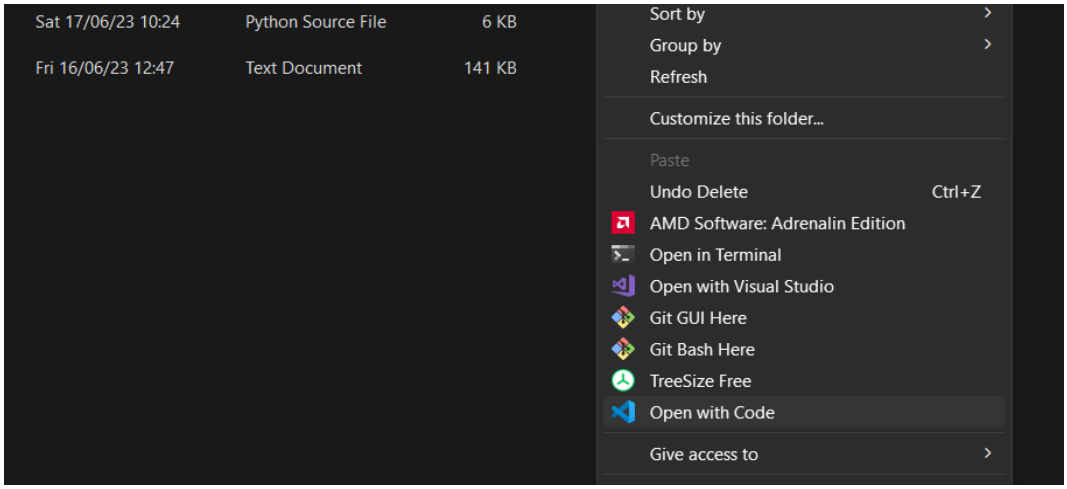
MSSV	Họ và tên	Lớp	Note
20127446	Giang Gia Bảo	20TGMT	

II. ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH

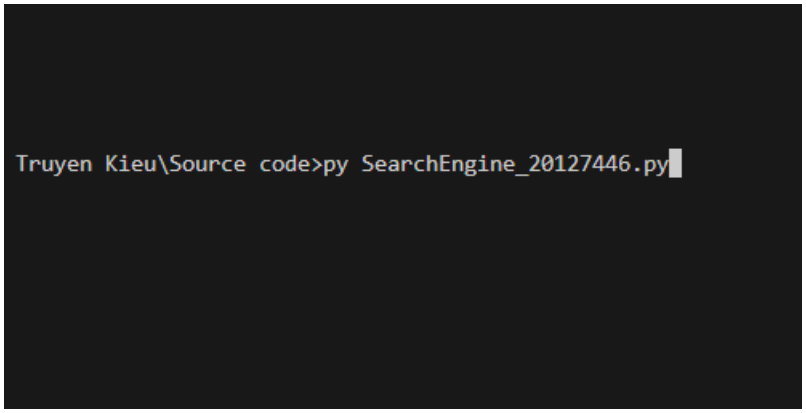
STT	Yêu cầu	Hoàn thành
1	Build system	100%
2	Evaluate system: precision & recall metrics	100%
3	Analyze the result and direction for improvement.	100%

III. CÁCH CHẠY SOURCE CODE

1. Mở file .py bằng vscode (hoặc các IDE khác)



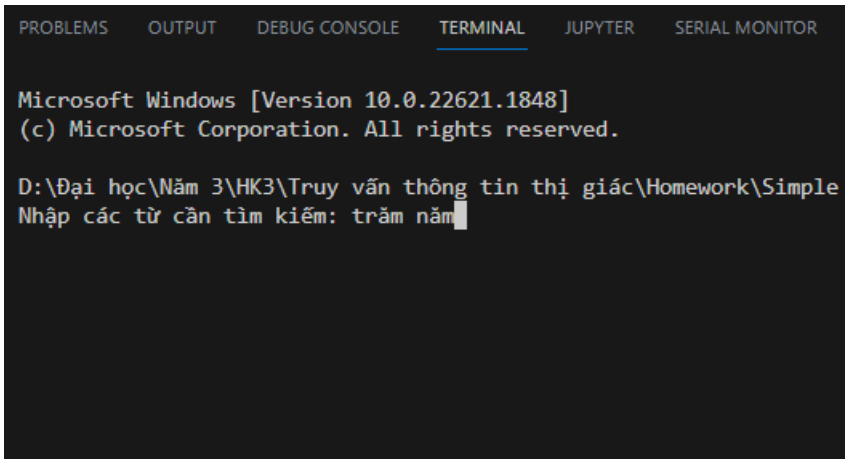
2. Mở terminal, gõ command để chạy chương trình 'py SearchEngine_20127446.py'



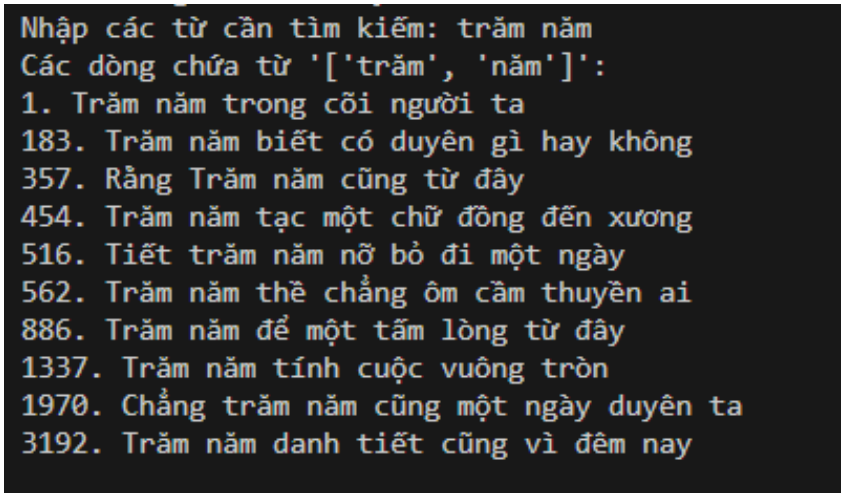
Lưu ý: cần kiểm tra đường dẫn của file .txt và sửa đúng đường dẫn trước khi chạy.

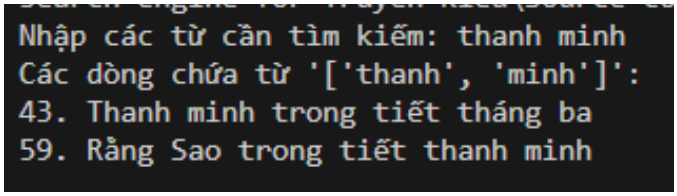
```
input_file_path = "./truyen_kieu_data.txt" # original file
output_file_path = "./truyen_kieu_dict.txt" # file dictionary
file_unidecode = "./truyen_kieu_unidecode.txt" # file after lọc hết dấu
```

3. **Gõ các từ muốn search (phải có dấu như trong bài thơ gốc)**



4. **Kết quả tìm kiếm**

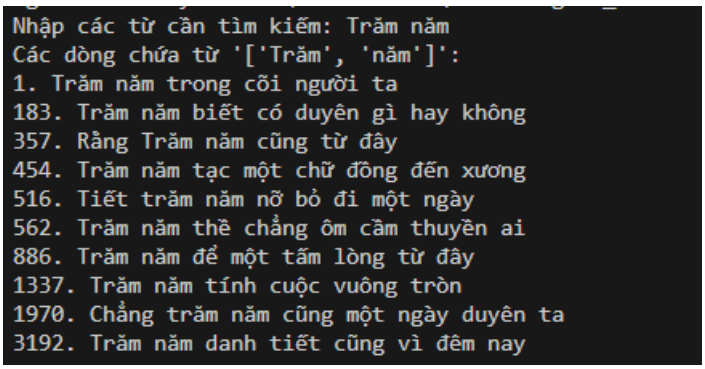




IV. EVALUATE SYSTEM

Trong ví dụ này ta kiểm tra hệ thống với cụm từ “Trăm năm”.

Kết quả trả về:



Số dòng chứa đúng cụm từ tìm kiếm là 7 dòng: 1, 183, 454, 562, 886, 1337, 3192.

Số dòng kết quả trả về là 10 dòng.

1. Evaluate Precision

Precision (độ chính xác): Precision đo lường tỷ lệ các kết quả tra cứu đúng so với tổng số kết quả tra cứu.

Precision = Số từ được tra cứu đúng / (Số từ được tra cứu đúng + Số từ được tra cứu sai)

Precision = 7/10 = 70%

2. Evaluate Recall Metric

Recall (độ phủ sóng): Recall đo lường tỷ lệ các từ được tra cứu đúng so với tổng số từ thực sự cần tra cứu.

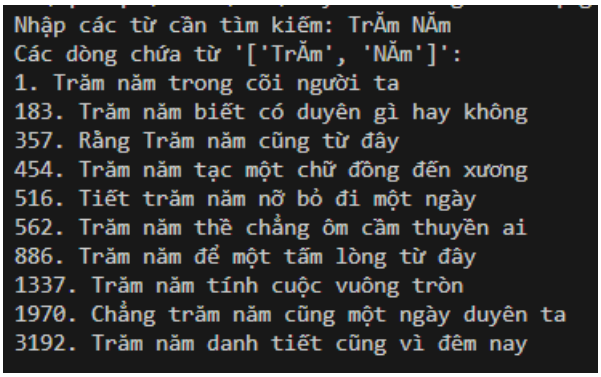
Recall = Số từ được tra cứu đúng / (Số từ được tra cứu đúng + Số từ bị bỏ sót)

Recall = 7/7 = 100%

3. Đánh giá

Do hệ thống chuyển đổi cụm từ tìm kiếm nhập vào thành các từ thường do hàm **lower()** nên kết quả trả về là tất cả các từ không phân biệt viết hoa hay viết thường.

Kết quả cho việc tìm cụm từ tương tự nhưng viết hoa một số ký tự:



Do vậy nên Độ phủ (Recall) luôn đạt 100% nhưng Độ chính xác (Precision) không thể đạt 100% mà phụ thuộc vào số lượng từ đó so với tổng số từ bao gồm cả in hoa và không in hoa.

V. ANALYZE AND IMPROVEMENT

1. Analyze result

Thông qua kết quả tìm kiếm. Hệ thống có thể tìm từ đơn, hoặc các cụm từ, câu thơ một cách nhanh và chính xác. Luôn cho ra tất cả các từ có thể trong văn bản.

Hoạt động bằng cách tìm các dòng chứa từng từ và sau đó lấy các dòng chung chứa tất cả các từ đã tìm kiếm và cho ra kết quả.

Tuy nhiên hệ thống vẫn chưa được phát triển để tìm các từ không dấu, dính nhau, viết hoa, viết thường.

VD: tram nam, tramnam, tRăM NĂm, Trăm năm, …

2. Direction for improvement

Hướng phát triển cho hệ thống:

Có thể tìm các từ không dấu như: tram nam, nhung ngay,…

Có thể phân biệt chữ hoa, chữ thường, chữ viết hoa sai cú pháp…

Phát triển mô hình học máy tự sửa lỗi ngữ pháp như sai chính tả, hoặc gợi ý từ gần giống với những từ tìm kiếm không có kết quả.

VD: người dùng tìm từ “thãnh thơi”, hệ thống sẽ tự sửa lỗi chính tả và cho ra kết quả của “thảnh thơi”, tìm từ “cáNh éN” sẽ sửa lỗi và tìm từ “cánh én”…

VI. REFERENCE

https://github.com/duyet/truyenkieu-word2vec/blob/master/truyen_kieu_data.txt

https://helpex.vn/question/cach-tot-nhat-de-loai-bo-cac-dau-trong-chuoi-unicode-python-la-gi-5cb16654ae03f6169c9dcc85

https://www.w3schools.com/python/python_dictionaries.asp