**TRUY XUẤT THÔNG TIN:**

**VECTOR SPACE MODLE VÀ LANGUAGE MODEL**

1. **Giới thiệu**

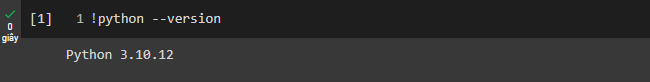
Trong báo cáo này, chúng tôi đã xây dựng các mô hình truy xuất thông tin sử dụng Vector Space Model và Language Model. Chúng tôi đã thực hiện các thí nghiệm này trên môi trường Google Colab và đã đảm bảo yêu cầu và thư viện cần thiết đã được cài đặt trước.

Để giúp người đọc có thể tái tạo kết quả của chúng tôi, chúng tôi đã tạo một tệp readme đi kèm với mã nguồn của báo cáo này. Tệp readme này sẽ hướng dẫn người đọc về cách thực hiện các thí nghiệm và chạy mã nguồn của chúng tôi một cách đầy đủ.

1. **Cài đặt chương trình.**

**Bước 1: Cài đặt môi trường và các thư viện không được hỗ trợ sẵn***.*

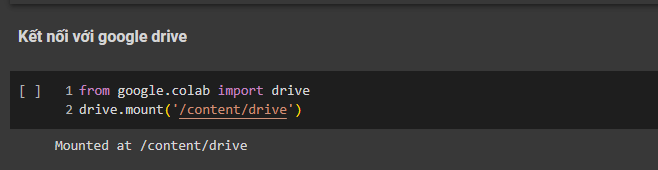
Chương trình của chúng em có thể được triển khai trên nhiều phiên bản python khác nhau, với nhiều môi trường chạy. Cụ thể nhóm chúng em đã sử dụng python phiên bản python 3.10.12



Một thư viện sử dụng nhiều và xuyên suốt, cung cấp các ‘stopwords’, ‘punkt’, ‘wordnet’ … là thư viện nltk. Do đó bạn phải chắc chắn rằng môi trường chạy của bạn đảm bảo được điều này. Khi đó, tiến hành download các công cụ hỗ trợ cho đồ án của mình.



Nếu bạn sử dụng google colab để tiến hành chạy code và triển khai chương trình, thay vì lưu dữ liệu trong phần ‘Tệp’ , sẽ bị mất dữ liệu khi bạn kết thúc thời gian chạy, gây tốn thời gian cho mỗi lần sử dụng. Thì chúng ta nên dùng google drive để lưu trữ dữ liệu và kết nối với Google Colab.



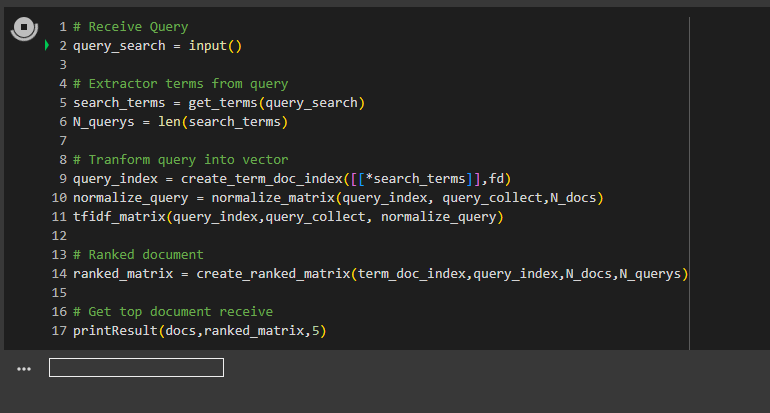
Cài đặt thêm thư viện Whoosh (nếu sử dụng vs code thì sẽ cài đặt ở terminal hoặc Command Prompt)



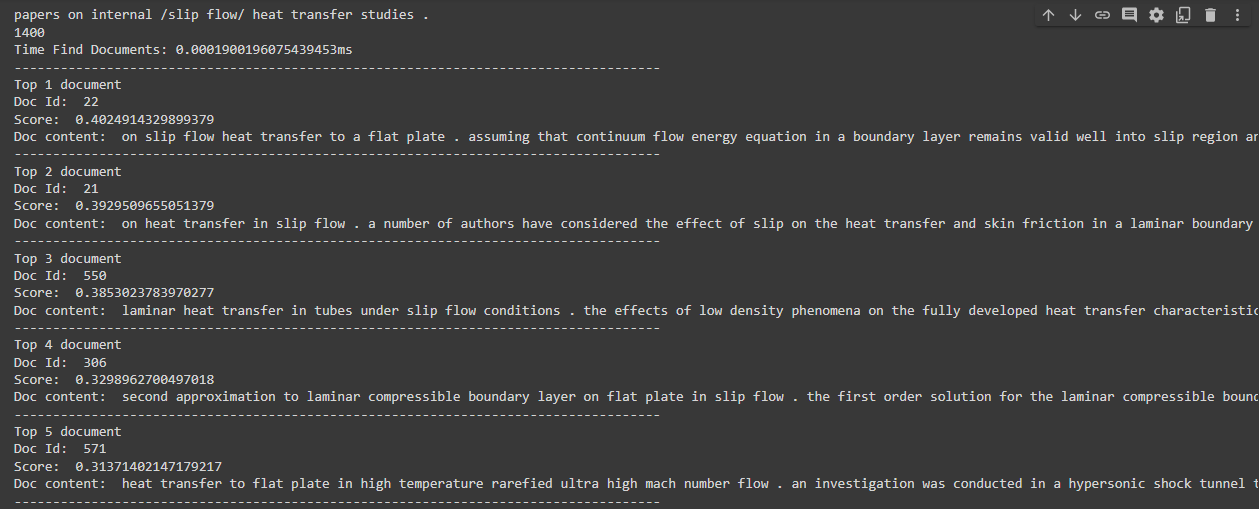
**Bước 2: Nhập câu truy vấn. Hoàn tất chương trình**

Sau khi hoàn tất và đảm bảo được các thư viện không được cài đặt sẵn, chúng ta chỉ cần chạy từng cell code theo thứ tự được trình bày sẵn trong file ipynb bằng tổ hợp phím: **ctrl + Enter.**

Sau khi hoàn tất các bước cài đặt và huấn luyện mô hình, khi testing kết quả, chương trình sẽ ngừng lại và yêu cầu bạn nhập câu truy vấn như hình bên dưới:



Chọn câu truy vấn với nội dung người dùng yêu cầu, nhấn **Enter** để hệ thống truy xuất thông tin tìm kiếm các document phù hợp và có liên quan.



Kết quả trả về top5 document có độ liên quan, tương đồng cao nhất với câu truy vấn, điểm tương đồng và thời gian truy xuất thông tin.

1. **Kết luận**

Nhóm đã thực hiện thử nghiệm một số mô hình, thư viện và công cụ hỗ trợ truy xuất thông tin, bao gồm thư viện Whoosh và các mô hình truy xuất thông tin khác như Vector Space Model (VSM) và Language Model (LM) để hiểu rõ hơn về hệ thống truy xuất. Các công cụ này có những ưu điểm và hạn chế riêng, và người dùng cần phải lựa chọn công cụ phù hợp cho nhu cầu của mình.

Các bước chạy, hoàn tất hệ thống truy xuất thông tin cũng được trình bày cụ thể như trên. Nếu quá trình chạy gặp trục trặc, bạn có thể liên hệ đến các thành viên trong nhóm.