**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

****

**BÁO CÁO MÔN TRUY XUẤT THÔNG TIN**

**ĐỒ ÁN: THIẾT KẾ HỆ THỐNG TRUY VẤN THÔNG TIN TRÊN VĂN BẢN**

**Lớp: CS231.I21**

**Sinh viên thực hiện: Mai Quốc Kiệt – 15520400**

**Lê Thiện Duy – 15520158**

# Lời cảm ơn

Chúng em xin cảm ơn giảng viên hướng dẫn là thầy Nguyễn Trọng Chỉnh đã tận tình hướng dẫn và giúp đỡ chúng em trong suốt quá trình học tập và thực hiện đồ án. Trong quá trình làm bài tập và đồ án, do thời gian và khả năng của bản thân còn hạn chế, nên chúng em không tránh khỏi sai sót. Vì vậy chúng em mong được sự bổ sung góp ý của thầy để hoàn thiện tốt hơn.

**Chúng em xin chân thành cảm ơn thầy và chúc thầy gặp nhiều thành công trong cuộc sống**

**CÀI ĐẶT VÀ KẾT QUẢ THỰC HIỆN**

1. **Cài đặt:**

Clone project tại: <https://github.com/QuocKiet/Information-retrieval>

Cài đặt python: https://www.python.org/downloads/windows/

Cài các thư viện cần thiết bằng cmd: pip install -r requirements.txt

Mở file search.py bằng sublime hoặc visual studio code, sau đó tiến hành build bằng ctrl + B cho sublime hoặc sử dụng cmd với lệnh python search.py cho visual studio code hoặc command prompt tại ngay thư mục chứa file search.py

Để chuyển đổi câu query cần thực hiện, thay thế câu query đó vào tham số đầu tiên của hàm queryprocess() tại hàm main() của chương trình:



1. **Kết quả thực hiện:**

Dictionary chứa inverted index: indexing

Dictionary chứa precision: precision và file precision.txt

Dictionary chứa recall: recall và file recall.txt

Thư mục chứa kết quả các truy vấn: “result/”

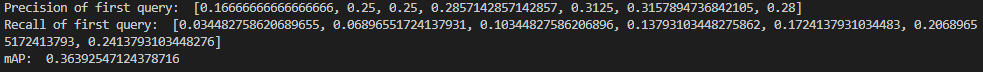
Số lương file liên quan trả về rất nhiều, nên nhóm chỉ lấy số lượng file liên quan bằng với số lượng file trong file kết quả của thầy (ví dụ file kết quả 1.txt của thầy có 25 kết quả thì file 1.txt trong thư mục result của nhóm sẽ có 25 kết quả trả về, sau đó dùng kết quả này để tính precision, recall, map)

Nhóm đã áp dụng stemming và lemmatizing trong xen kẽ nhau trong quá trình thực hiện và nhận thấy lemmatizing cho kết quả tốt hơn

Kết quả 3 văn bản có độ tương đồng cao nhất của query thứ nhất:



Kết quả precision và recall của query đầu tiên và giá trị mAP của nhóm:



Kết quả mAP của nhóm: 0.36

1. **Link github chứa project của nhóm:**

<https://github.com/QuocKiet/Information-retrieval>