KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện Thoại: (08) 38.354.266 - Fax:(08) 38.350.096

CTT009 – NHẬP MÔN CÔNG NGHỆ THÔNG TIN 2

ĐỒ ÁN MÔN HỌC

DAMH-02: XỬ LÍ NGÔN NGỮ TỰ NHIÊN

I. Mô tả đồ án

1. Nội dung chính

Xây dựng một hệ thống tìm kiếm văn bản (Search Engine).

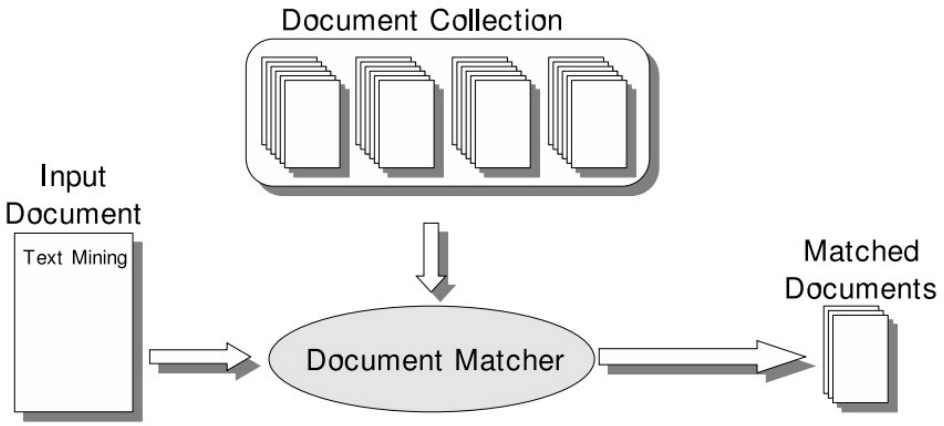
2. Mục tiêu đồ án

Sinh viên thực hiện đồ án sẽ nắm được các kiến thức

- Kiến thức xác định đặc trưng và vector hóa trong xử lý văn bản
- Kiến thức về tính độ tương tự giữa các vector văn bản
- Kiến thức về xếp hạng trong tìm kiếm văn bản
- Kiến thức về lập trình python

3. Nội dung chi tiết của đồ án

Xây dựng một hệ thống tìm kiếm (Search Engine) như hình bên dưới:



Cụ thể, cho trước một tập văn bản D, người dùng sẽ nhập vào một văn bản đầu vào (Input Document), nhiệm vụ của một hệ thống tìm kiếm văn bản là trả về các văn bản có độ tương đồng giảm dần so với “Input Document”.

II. Hướng dẫn

Các bước thức hiện như sau:

Bước 1: Vector hóa các văn bản trong D dưới một tập đặc trưng (Mô hình cơ bản: Bag-of-words và TF-IDF)

https://en.wikipedia.org/wiki/Bag-of-words_model

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Bước 2: Vector hóa "Input Document" bằng mô hình đã sử dụng ở bước 1

Bước 3: Tính độ tương đồng (similarity measure) giữa vector đại diện cho "Input Document" và vector đại diện cho từng văn bản trong D (các phương pháp tính độ tương đồng giữa hai vector văn bản: Euclidean, Cosine …)

https://en.wikipedia.org/wiki/Similarity_measure

Bước 4: Chọn ra top k văn bản trong D tương đồng giảm dần với "Input Document".

Tập dữ liệu: http://www.mediafire.com/file/njw6g524o506lav/ChiNhan.rar

Ngôn ngữ lập: python (*Ưu tiên cài giao diện*)

III. Các kết quả cần đạt được

- Báo cáo tìm hiểu.
- Báo cáo tiến độ (tuần sau)
- Sản phẩm