Bộ Giáo Dục & Đào Tạo
Trường Đại Học Công Nghệ TP.HCM-HUTECH

Môn học: **KHAI THÁC THÔNG TIN**
Mã đề tài: **02**
Tên đề tài: **Xây dựng hệ truy hồi thông tin tiếng Việt dựa trên tiếp cận không gian vector (vector space)**

1. Nội dung thực hiện

Nội dung của đề tài gồm các yêu cầu sau:

- Xây dựng một hệ thống truy hồi thông tin/tìm kiếm web dựa trên nội dung theo hướng tiếp cận không gian vector.
- Áp dụng TF-IDF cho việc học mô hình biểu diễn/chuyển đổi các văn bản/tài liệu về dạng các vectors.
- Có giao diện (có thể xây dựng ứng dụng dạng web) tìm kiếm cho người dùng nhập vào truy vấn (query).
- Có xử lý stopwords cho tiếng Việt [1].
- Có sử dụng cơ chế tách từ tối ưu cho tiếng Việt, có thể sử dụng các thư viện NLP có sẵn như: JvnTextPro, VnTokenizer, v.v.
- Có cơ chế chỉ mục các tài liệu/văn bản và lưu trữ chỉ mục – học viên có thể **tùy chọn** tự xây dựng hoặc sử dụng bất cứ nền tảng cơ sở dữ liệu lưu trữ dữ liệu chỉ mục.

2. Sản phẩm yêu cầu

Đề tài yêu cầu các sản phẩm sau:

- **Thuyết minh đề tài:** nộp dạng bài tiểu luận **từ 20-25 trang**. Học viên nộp bài dưới cả 02 hình thức, gồm [**Bản mềm**]: Nộp trên hệ thống hỗ trợ học tập E-learning HUTECH (http://e-graduate.hutech.edu.vn/portal) và [**Bản cứng** – 2 bản]: Nộp cho Giảng viên phụ trách học phần. Nội dung trình bày trong bài thuyết minh:
  - Lý thuyết nền tảng.
  - Cách tiếp cận và các bước xây dựng ứng dụng.
  - Minh họa & demo về sản phẩm.

[1] https://github.com/phamtheanhphu/IRS_Course/tree/master/data/stopwords

- **Thuyết trình trước lớp**: có slides – thời gian thuyết trình từ 30-45 phút (mọi thành viên trong nhóm đều phải thuyết trình).
- Ứng dụng được xây dựng – có demo.

3. Hướng dẫn

- Thu thập dữ liệu các tài liệu/văn bản tiếng Việt – tầm > 1000 tài liệu (có thể thu thập từ các website tin tức, mạng xã hội, v.v.).
- Học viên có thể dùng bất cứ ngôn ngữ lập trình nào để xây dựng ứng dụng.

4. Tài liệu tham khảo

Học viên có thể tham khảo bất cứ nguồn tài liệu và mã nguồn phần mềm nào có thể lấy được từ Internet. Ngoài ra học viên có thể tham khảo thêm một số dự án mẫu dưới đây:

- https://www.cs.utexas.edu/~mooney/ir-course/proj2/
- https://bart.degoe.de/building-a-full-text-search-engine-150-lines-of-code/
- https://www.cs.toronto.edu/~muuo/blog/build-yourself-a-mini-search-engine/