**Đề Cương Đồ Án Cuối Kỳ**

**Đề tài: Xây dựng Máy Tìm Kiếm Chuyên Sâu (Vertical Search Engine)**

1. **Máy Tìm Kiếm Chuyên Sâu Là Gì?**

Google có thể tìm mọi thứ, nhưng nếu bạn chỉ muốn tìm:

* 🎬 Phim ảnh
* 👩‍🍳 Công thức nấu ăn
* 📚 Tài liệu học tập
* 📱 Sản phẩm công nghệ

… thì kết quả từ Google thường quá rộng, không chính xác. Vì vậy, bạn sẽ xây dựng một máy tìm kiếm chỉ tập trung vào một lĩnh vực cụ thể.

Ví dụ:

* Nhóm A: Máy tìm kiếm món ăn
* Nhóm B: Máy tìm kiếm phim
* Nhóm C: Máy tìm kiếm tài liệu học tập
* Nhóm D: Máy tìm kiếm sản phẩm công nghệ

Bạn được tự chọn lĩnh vực, miễn là có dữ liệu trên web để thu thập.

1. **Mục Tiêu Của Đồ Án**

Kiến thức bạn sẽ học:

* Hiểu cách máy tìm kiếm hoạt động từ A → Z
* Áp dụng các kỹ thuật: thu thập dữ liệu, xử lý văn bản, xây dựng chỉ mục, tìm kiếm, đánh giá

Kỹ năng bạn sẽ rèn luyện:

* Viết code backend để xử lý dữ liệu
* Làm giao diện web đơn giản
* Làm việc nhóm, chia việc hợp lý

Ứng dụng thực tế:

* Có thể demo cho giảng viên và người dùng thử
* Nếu làm tốt, có thể dùng thật trong một phạm vi nhỏ

1. **Các Bước Thực Hiện (5 Modules)**

**🔹Module 1: Thu Thập Dữ Liệu (Crawling)**

📌 Mục tiêu: Viết chương trình tự động lấy dữ liệu từ 1–2 website

📋 Việc cần làm:

* Chọn website phù hợp với lĩnh vực
* Viết crawler bằng Python (Scrapy hoặc BeautifulSoup)
* Tuân thủ robots.txt, không spam website
* Lưu dữ liệu dạng JSON, CSV hoặc database nhỏ (SQLite, MongoDB)

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

**🔹 Module 2: Xử Lý Văn Bản & Xây Dựng Chỉ Mục**

📌 Mục tiêu: Làm sạch dữ liệu và xây dựng hệ thống tìm kiếm nhanh

📋 Việc cần làm:

* Làm sạch văn bản:
  + Tách từ (tokenization)
  + Bỏ từ dừng (ví dụ: “và”, “của”, “là”…)
  + Chuyển về chữ thường
  + (Nâng cao) Dùng thư viện tiếng Việt như underthesea để tách gốc từ
* Xây dựng chỉ mục ngược (Inverted Index):
* Mỗi từ → danh sách các tài liệu chứa từ đó
* Mỗi tài liệu lưu: docID, số lần xuất hiện, vị trí từ

📌 Kết quả: Truy vấn nhanh hơn, không cần quét toàn bộ dữ liệu

🔹 **Module 3: Truy Vấn & Xếp Hạng Kết Quả**

📌 Mục tiêu: Nhận truy vấn từ người dùng và trả về kết quả phù hợp

📋 Việc cần làm:

* Xử lý truy vấn giống như xử lý dữ liệu
* Tính độ liên quan giữa truy vấn và tài liệu bằng: TF-IDF
* Có thể thêm trọng số cho:
* Tiêu đề
* Các trường quan trọng (tên món, tên phim…)

📌 Kết quả: Trả về danh sách kết quả xếp hạng theo độ phù hợp

🔹 **Module 4: Giao Diện Web**

📌 Mục tiêu: Xây dựng website đơn giản để demo

📋 Việc cần làm:

* Dùng Flask hoặc Django để làm web
* Trang chủ: có ô nhập tìm kiếm
* Trang kết quả:
* Hiển thị tiêu đề (link về trang gốc)
* Đoạn tóm tắt có highlight từ khóa
* Phân trang nếu có nhiều kết quả

📌 Yêu cầu: Giao diện rõ ràng, dễ dùng, không cần quá đẹp

**🔹 Module 5: Đánh Giá Hệ Thống**

📌 Mục tiêu: Kiểm tra xem hệ thống hoạt động tốt không

📋 Việc cần làm:

* Tạo 15–20 truy vấn mẫu
* Với mỗi truy vấn, đánh dấu 5–10 kết quả đúng (ground truth)
* Viết script tính các chỉ số:
* Precision@10: độ chính xác trong 10 kết quả đầu
* MAP (Mean Average Precision): chất lượng trung bình

📌 Kết quả: Có số liệu chứng minh hệ thống hoạt động hiệu quả

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**5. Kết Quả Cần Nộp**

📦 Nộp những gì?

* Source code đầy đủ, có comment rõ ràng
* Báo cáo PDF gồm:
  + Kiến trúc hệ thống
  + Giải thích thiết kế & thuật toán
  + Kết quả đánh giá (Precision, MAP)
  + Phân công công việc từng thành viên
* Video demo ngắn
* Thuyết trình & Q&A trước giảng viên