BÁO CÁO GIỮA KÌ TIỂU LUẬN CHUYÊN NGÀNH

ĐỀ TÀI: XÂY DỰNG HỆ THỐNG QUẢN LÝ TÀI LIỆU VÀ KIỂM TRA TÀI LIỆU TRÙNG LẮP TRÊN KHO DỮ LIỆU CHUNG

----------------

**GVHD:** Nguyễn Đức Khoan

**SVTH:** Nguyễn Văn Hà 17110130

Nguyễn Văn Thắng 17110230

1. **PHẠM VI NGHIÊN CỨU**
2. **MỤC TIÊU**

Xây dựng hệ thông quản lý tài liệu và kiểm tra tài liệu trùng lặp trên kho dữ liệu chung:

* Tìm hiểu,cài đặt và cấu hình nodejs,reactjs,mongodb.
* Tìm hiểu thuật toán vector hóa văn bản.
* Tìm hiểu, cài đặt và cấu hình công cụ elasticsearch.

1. **NỘI DUNG**

Tổng quan về thuật toán vector hóa văn bản.

* Giới thiệu về các phương pháp vector hóa văn bản và thuật toán.
* Giới thiệu về phương pháp Bag of Words (BoW).

Tổng quan về elasticsearch.

* Giới thiệu về elaticsearch.
* Giới thiệu về full-text search.
* Cài đặt và cấu hình elasticsearch.
* Nghiên cứu sử dụng full-text search.

Tìm hiểu nghiệp vụ webisite quản lý tài liệu và áp dụng vào đề tài từ đó ta có thể:

* Vẽ usecase
* Thiết kê architecture
* Thiết kế database

Xây dụng web application sử dụng expressjs framework để quản lý kho dữ liệu và kiểm tra trùng lắp:

* Xây dựng các modules như:
* Bảng điều khiển
* Dữ liệu
* Nhân viên
* Phòng ban
* Tài khoản
* Tìm kiếm,…
* Các chức năng cơ bản cho người dùng
* Login/logout(Theo role)
* Upload dữ liệu
* Các chức năng cơ bản của admin
* login/logout
* quản lý thông tin nhân viên
* quản lý tài khoản người dùng

Xây dựng web API cho việc trao đổi dữ liệu từ reactJs sang database:

* Viết API Sử dụng nodeJs

1. **KHỐI LƯỢNG CÔNG VIỆC ĐÃ THỰC HIỆN**

* **Chuẩn bị**
* Nghiên cứu tài liệu về vector hóa văn bản.
* Nghiên cứu tài liệu về elasticsearch.
* Cài đặt và cấu hình nodejs, reactjs, mongodb, elasticsearch.
* Hiểu được cách thức hoạt động của elasticsearch(full-text search).
* Hiểu được thuật toán vector hóa văn bản(Bag of Words).
* Yêu cầu phần mềm
* Thiết kế usecase
* Thiết kế architecture
* Thiết kế database
* Back-end
* Xây dựng và đưa database lên mongodb
* Viết các api cần thiết để sử dụng(login,upload dữ liệu,…)
* Front-end
* Dashboard UI
* Chức năng login
* Quản lý tài khoản người dùng
* Upload dữ liệu

So với những nội dung mục tiêu đã đặt ra thì nhóm xin tự đánh giá hoàn thành 50% tiến độ.

1. **KHÓ KHĂN**

* Chưa thể thực hiện việc đưa 1 văn bản đã vector hóa nhiều chiều về số chiều tối thiểu.

*TP.Hồ Chí Minh, Ngày 25 tháng 11 năm 2020*

**GIÁO VIÊN HƯỚNG DẪN SINH VIÊN THỰC HIỆN**

Thầy Nguyễn Đức Khoan Nguyễn Văn Thắng Nguyễn Văn Hà