**Bài test AI Engineer**

**Phần 2: Code theo đề bài sau**

Mục tiêu: Đánh giá khả năng xử lý dữ liệu pháp luật, xây dựng hệ thống NLP tiếng Việt, và triển khai mô hình AI cơ bản.

**Bước 1: Tiền Xử Lý Văn bản Pháp luật thô**

Yêu cầu:

* Cho trước 5 file văn bản pháp luật thô trong thư mục .**\data**.
* Nhiệm vụ:
* Phân đoạn văn bản thành các điều luật riêng biệt (ví dụ: Điều 1, Điều 2...).
* Làm sạch dữ liệu:
* Loại bỏ ký tự đặc biệt, chuẩn hóa khoảng trắng.
* Chuyển đổi định dạng số (ví dụ: "1.2m" → "1.200.000 VNĐ").
* Gán nhãn metadata:
* Trích xuất tiêu đề điều luật, số hiệu điều.
* Lưu dưới dạng JSON với cấu trúc:

*{*

*"dieu\_1": {*

*"title": "Quy định về tốc độ",*

*"text": "1. Xe ô tô phải tuân thủ tốc độ tối đa 60 km/h trong khu vực thành phố.\n2. Trường hợp đường phố có biển báo khác, lái xe phải tuân theo biển báo.\n a) Điểm 1: Biển báo giới hạn tốc độ 40 km/h áp dụng cho xe tải.\n b) Điểm 2: Biển báo cấm vượt áp dụng cho tất cả phương tiện." },*

*"dieu\_2": {*

*"title": "Hành vi nghiêm cấm",*

*"text": "1. Lái xe dưới ảnh hưởng của rượu bia hoặc chất kích thích.\n2. Sử dụng điện thoại khi lái xe.\n3. Vượt đèn đỏ.\n4. Chở quá số người quy định.\n a) Điểm 1: Vi phạm chở quá 2 người sẽ bị phạt tiền từ 5-10 triệu đồng."*

*}*

*}*

Công cụ khuyến nghị: Python, **underthesea** (tokenize tiếng Việt), **regex**, **pandas, json**.

**Bước 2: Xây Dựng Hệ Thống Tìm Kiếm Thông Minh**

Yêu cầu:

* Sử dụng dữ liệu đã tiền xử lý từ Phần 1 để xây dựng hệ thống tìm kiếm dựa trên câu hỏi pháp lý.
* Nhiệm vụ:
* Mã hóa văn bản pháp luật:
* Dùng mô hình **PhoBERT** hoặc **Sentence-BERT** (tiếng Việt) để tạo vector embedding cho từng điều luật.
* Tìm kiếm tương đồng:
* Với câu hỏi người dùng (ví dụ: "Xe máy chạy quá tốc độ bị phạt thế nào?"), tính toán độ tương đồng (cosine similarity) giữa câu hỏi và các điều luật.
* Trả về TOP 5 điều luật phù hợp nhất.
* Đánh giá:
* Tính độ chính xác (accuracy) trên tập kiểm tra gồm 10 cặp (câu hỏi, điều luật đúng).
* Thuật toán tìm kiếm có những điểm tiên tiến.

**10 câu hỏi kiểm tra:**

1. NLĐ bị sa thải có được trả lương hay không?
2. Người sử dụng lao động được sa thải người lao động nữ đang mang thai không?
3. Quy định về điều chuyển nhân sự được quy định như thế nào?
4. Người lao động được thuê làm giám đốc doanh nghiệp Nhà nước được hưởng các chế độ về tiền lương, thưởng như thế nào?
5. Làm việc 8h một ngày thì được nghỉ giữa giờ ít nhất bao nhiêu phút?
6. Người sử dụng lao động đào tạo nghề nghiệp và phát triển kỹ năng nghề cho người lao động như thế nào?
7. Nguyên tắc cho thuê lại lao động là gì?
8. Thời hạn của thỏa ước lao động tập thể như thế nào?
9. Hợp đồng lao động được giao kết theo hình thức nào?
10. Nội dung về đào tạo lao động có bắt buộc phải ghi vào hợp đồng lao động?

Công cụ khuyến nghị: **HuggingFace Transformers**, **scikit-learn**, **. . .**

**Bước 3: Triển Khai API Đơn Giản**

Yêu cầu:

* Đóng gói hệ thống tìm kiếm thành API RESTful.
* Nhiệm vụ:
* Xây dựng API:
* Endpoint **/search**: Nhận JSON **{query: "..."}** và trả về danh sách điều luật phù hợp.
* Dùng **FastAPI** hoặc **Flask**.
* Container hóa:
* Viết **Dockerfile** để đóng gói ứng dụng.
* Chạy thử trên máy cá nhân.

**# Thực hiện và đóng gói tất cả thành file zip và gửi tới email *tuyendungai@thuvienphapluat.net***