MÔ TẢ

1. Giới thiệu:

Bài toán phân loại văn bản pháp luật là một lĩnh vực trong xử lý ngôn ngữ tự nhiên (NLP), nơi mà mục tiêu là tự động phân loại các văn bản pháp luật vào các danh mục hoặc lớp cụ thể dựa trên nội dung của chúng. Đây là một nhiệm vụ quan trọng trong việc quản lý, tìm kiếm và sử dụng các tài liệu pháp luật, và có nhiều ứng dụng thực tế, chẳng hạn như:

* Tìm kiếm thông tin pháp luật: Giúp người dùng nhanh chóng tìm thấy các tài liệu pháp luật liên quan đến một chủ đề cụ thể.
* Hỗ trợ quyết định pháp lý: Hỗ trợ các luật sư và thẩm phán trong việc tìm kiếm các án lệ và quy định pháp luật liên quan.
* Quản lý tài liệu pháp luật: Giúp các cơ quan pháp luật và công ty luật tổ chức và quản lý hiệu quả các tài liệu pháp luật.\

1. Các bước hoạt động:

B1) **Thu thập dữ liệu**:

Thu thập một tập hợp các văn bản pháp luật, bao gồm các văn bản luật, nghị định, thông tư, quyết định, và các tài liệu pháp lý khác.

B2) **Tiền xử lý dữ liệu**: Bao gồm các bước như:

* Đầu tiên sẽ là tách bộ luật theo các điều khoản
* Với các điều khoản sẽ được xử lí text theo thư pyvi để biến text thành các từ có liên quan sẽ thành 1 cụm để xử lí
* Sau đó sử dụng 1 số phương pháp tách text thành số:
  + Count Vectors
  + TF-IDF vectors:
    - Word Level
    - Ngram\_level
* Sử dụng thuật toán SVD import từ thư viện để giám số lượng chiều nhưng mà vẫn đảm bảo được các yếu tố gốc

B4) **Xây dựng mô hình phân loại**: Sử dụng các thuật toán máy học để huấn luyện mô hình phân loại. Một số thuật toán phổ biến bao gồm:

* Naive Bayes
* Logistic Regression
* Deep Learning models (như LSTM, BRNN)

B4) **Đánh giá mô hình**: Sử dụng các tập dữ liệu kiểm tra để đánh giá hiệu quả của mô hình.

B5) **Triển khai mô hình**: Sau khi mô hình đạt hiệu quả cao, nó có thể được triển khai vào hệ thống thực tế để phân loại các văn bản pháp luật mới.

1. Công nghệ sử dụng:
2. Thư viện sử dụng:

* Pyvi: Thư viện NLP tiếng việt
* Tqdm:
* Numpy
* Gensim: Thư viện NLP
* Os
* Pathlib
* Pickle: lưu file dạng byte
* Sklearn
* Tensorflow

1. Các mô hình đang sử dụng:

* Logistic Regression:
* Long short term memory (LSTM):
* Bidirectional Recurrent Neural Network (BRNN):

1. Tổ chức file/folder:

* 2 folder cha:
  + TEST\_FULL
  + TRAIN\_FULL
* Trong mỗi folder cha sẽ có folder con tương đương với là nhãn của văn bản mà các folder con chứa ví dụ:
  + TEST\_FULL:
    - Dan\_su
    - Lao\_dong
    - Kinh\_te
    - Dat\_dai
    - Dau\_thau
    - Hinh\_su
  + TRAIN\_FULL:
    - Dan\_su
    - Lao\_dong
    - Kinh\_te
    - Dat\_dai
    - Dau\_thau
    - Hinh\_su

1. Training Model:

* Logistic Regression:
  + Dự đoán đúng 86%
  + Thời gian chạy: nhanh
  + Trong trường hợp đata to hơn chưa chắc tốt
* Long short term memory (LSTM):
  + Dự đoán đúng 84%
  + Thời gian chạy: nhanh
  + Cần điều chỉnh lại tập train để cải thiện
* Bidirectional Recurrent Neural Network (BRNN):
  + Dự đoán đúng 85%
  + Thời gian chạy: lâu hơn so với LSTM
  + Cần điều chỉnh lại tập train để cải thiện