Project 1

1 Description

Đồ án này được xây dựng nhằm giúp các bạn có cái nhìn sơ khởi về hệ thống hỏi đáp tự động. Trong đó, hệ thống hỏi đáp tự động sẽ gồm 2 giai đoạn chính:

- Trích chọn văn bản liên quan
- Trích xuất câu trả lời từ ngữ cảnh liên quan

Trong project này, các bạn sẽ tập trung tìm hiểu giai đoạn đầu tiên: trích chọn văn bản liên quan. Tuy nhiên, nhằm đơn giản hóa bài toán, các bạn cần thực hiện việc xây dựng mô hình dự đoán độ tương đồng/ liên quan giữa câu truy vấn (câu hỏi) và văn bản ngữ cảnh (cho trước) nhằm kết luận rằng ngữ cảnh đã cho có đủ thông tin để trả lời câu hỏi hay không?

Cụ thể, tập ngữ liệu (dataset) sẽ được download tại `https://www.kaggle.com/datasets/duyminhnguyentran/csc15105` và các công việc yêu cầu cụ thể được trình bày dưới đây.

2 Công việc cụ thể

2.1 EDA (Exploratory Data Analysis)

Phân tích Khám phá Dữ liệu giúp chúng ta có cái nhìn đầu tiên về dữ liệu. Trong phần này, các bạn cần thực hiện việc phân tích và thống kê về dữ liệu. Một số phân tích gợi ý: thống kê độ dài của câu hỏi, độ dài của đoạn văn, phân bố của label,....

2.2 Thử nghiệm trên mô hình tự xây dựng

Sinh viên được yêu cầu, sử dụng dataset đã được cung cấp, và tiến hành huấn luyện cùng đánh giá hiệu năng các mô hình cơ bản như sau.

Mô hình bao gồm 2 phần: Word Embedding và classifier. Trong đó:

- Word Embedding:
    0. BERT-multilingual
    1. BARTPho
    2. PhoBERT
    3. Word2Vec/Glove
- Classifier:
    0. Traditional Machine Learning: SVM, CRF, HMM, Naive Bayes, ...
    1. Fully Connected Layer
    2. Convolution Neural Network (sentence embedding) + Fully Connected Layer (classifier)

Cách thức lựa chọn mô hình kết hợp: Mỗi nhóm sẽ lựa chọn tổ hợp mô hình của mình theo quy tắc dưới đây:

- Phần Embedding: Tổng 3 số cuối mã số sinh viên của các thành viên trong nhóm và chia dư cho 4.
- Phần Classifier: Tổng 3 số cuối mã số sinh viên của các thành viên trong nhóm và chia dư cho 3.

Ví dụ: nhóm gồm 4 sinh viên với mã số gồm $\{001, 002, 003, 004\}$ thì sẽ có tổ hợp là:

- Phần Embedding: $(1 + 2 + 3 + 4) = 10 mod 4 = 2 \Rightarrow$ PhoBERT
- Phần Classifier: $10 mod 3 = 1 \Rightarrow$ Fully Connected Layer

2.3 Thử nghiệm trên mô hình đã có sẵn

Nhóm sẽ tự lựa chọn một mô hình đã được giới thiệu gần đây (chỉ được lựa chọn các mô hình được công bố từ năm 2020 đến năm 2024) và áp dụng mô hình vào bài toán đã cho.

3 Submission

1. Source code: gồm các folder con $Q1, Q2, Q3$ chứa code theo từng yêu cầu 2.1, 2.2, 2.3. Lưu ý, code phải chạy được trên nền tảng Google Colab và kèm file readme để cấu hình nếu cần thiết cho từng phần.
2. Báo cáo: yêu cầu cần đầy đủ các thành phần sau đây:
   (a) Bìa
   (b) Mục lục
   (c) Tự đánh giá: Tự đánh giá về mức độ hoàn thành của từng yêu cầu. Bảng phân công công việc của từng thành viên và mức độ hoàn thành của từng thành viên
   (d) Nội dung chính: Chia thành từng mục theo từng yêu cầu của đồ án. Bao gồm:
      i. EDA (Exploratory Data Analysis): thể hiện bằng các biểu đồ, các bảng. Mô tả ý nghĩa và giải thích và nhận xét các số liệu.
      ii. Thử nghiệm trên mô hình tự xây dựng: Mô tả từng thành phần của mô hình và thực nghiệm độ hiệu quả của mô hình trên bộ dataset đã cho và đưa ra nhận xét.
      iii. Thử nghiệm trên mô hình đã có sẵn: Mô tả mô hình. Báo cáo kết quả của thực nghiệm của mô hình trên bộ dataset đã cho và đưa ra nhận xét.
   (e) **Lưu ý 1:** Đồ án sẽ được đánh giá trên tỉ lệ giữa khối lượng công việc được trình bày trong báo cáo và số lượng thành viên. Báo cáo cần được trình bày chỉn chu và chi tiết vì 50% điểm đồ án đến từ báo cáo.
   (f) **Lưu ý 2:** Sinh viên cũng cần đọc hiểu mô hình ở mục **2.3** nhằm thực hiện báo cáo đồ án tại lớp Lý thuyết. Để thực hiện báo cáo đồ án tại lớp Lý thuyết, cần chuẩn bị bài thuyết trình gồm: nội dung mô hình, điểm mạnh mô hình, cách hoạt động, đánh giá kết quả, so sánh các siêu tham số và nhận xét....

4 Notice

- Mỗi nhóm gồm 4 - 5 thành viên. Không chấp nhận các nhóm lẻ. Nhóm có ít hơn / nhiều hơn số thành viên quy định cần gửi email cho GVLT và GVTH cho đến tối đa 2 tuần sau khi công bố đồ án.
- Bất cứ hình thức đạo văn, sao chép có chủ đích dù là vô tình hay cố tình đều sẽ bị 0 điểm toàn môn học. Trong trường hợp cần thiết, GVLT và GVTH có quyền yêu cầu SV/nhóm SV thực hiện vấn đáp để đưa ra quyết định cuối cùng.