**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

A picture containing logo

Description automatically generated**🙡🕮🙣**

**BÁO CÁO ĐỒ ÁN 2**

MÔN: KHAI THÁC DỮ LIỆU VĂN BẢN VÀ ỨNG DỤNG

Lớp: 21CNTThuc

NHÓM 10

GVLT: Lê Thanh Tùng

Nguyễn Trần Duy Minh

Thành phố Hồ Chí Minh, tháng 4 năm 2024

Mục lục

[**I.** **THÔNG TIN NHÓM 10** 3](#_Toc164361760)

[**II.** **BẢNG PHÂN CÔNG** 3](#_Toc164361761)

[**III.** **THÔNG TIN VỀ ĐỒ ÁN** 3](#_Toc164361762)

[**IV.** **CÁC NHIỆM VỤ CHÍNH** 4](#_Toc164361763)

[**1)** **EDA** 4](#_Toc164361764)

[**2)** **phoBERT** 6](#_Toc164361765)

[**3)** **Electra** 7](#_Toc164361766)

[**4)** **So Sánh** 8](#_Toc164361767)

[**V.** **TÀI LIỆU THAM KHẢO** 9](#_Toc164361768)

# **THÔNG TIN NHÓM 10**

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ Và Tên** | **Email** |
| 21127326 | Cao Trung Kiên | ctkien21@clc.fitus.edu.vn |
| 21127257 | Phạm Trần Minh Duy | ptmduy21@clc.fitus.edu.vn |
| 20127568 | Nguyễn Hữu Nam | nhnam20@clc.fitus.edu.vn |
| 20127385 | Huỳnh Hoàng Gia Uy | hhguy20@clc.fitus.edu.vn |
| 20127318 | Phan Trí Tài | 20127318@student.hcmus.edu.vn |

1. **BẢNG PHÂN CÔNG**

|  |  |  |
| --- | --- | --- |
| **Công việc** | **Người thực hiện** | **Đánh giá** |
| EDA | Huỳnh Hoàng Gia Uy | 80% |
| phoBERT | Phạm Trần Minh Duy  Nguyễn Hữu Nam | 90% |
| Alexa | Cao Trung Kiên | 90% |
| Báo cáo, thuyết trình | Phan Trí Tài | 100% |

# **THÔNG TIN VỀ ĐỒ ÁN**

* 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 thứ 2: trích xuất câu trả lời từ ngữ cảnh liên quan. Cụ thể, các bạn cần thực hiện việc xây dựng mô hình trích xuất câu trả lời từ câu truy vấn (câu hỏi) và văn bản ngữ cảnh (cho trước).

# **CÁC NHIỆM VỤ CHÍNH**

## **EDA**

* Mục Đích:
  + Hiểu rõ hơn về dữ liệu: EDA giúp chúng ta hiểu rõ hơn về cấu trúc, mô hình và mối quan hệ giữa các biến trong dữ liệu. Điều này rất quan trọng để xác định cách tiếp cận vấn đề trích xuất văn bản.
  + Phát hiện lỗi và ngoại lệ: EDA giúp phát hiện các giá trị ngoại lệ hoặc lỗi trong dữ liệu, điều này rất quan trọng vì nó có thể ảnh hưởng đến chất lượng của mô hình huấn luyện.
  + Chọn lựa đặc trưng: Qua EDA, chúng ta có thể xác định được những đặc trưng nào có ích cho việc huấn luyện mô hình, và những đặc trưng nào có thể bỏ qua.
  + Định hình mô hình: EDA cũng giúp chúng ta hiểu rõ hơn về phân phối của dữ liệu, điều này có thể hỗ trợ quyết định về việc chọn lựa mô hình phù hợp.
* Phân Tích:
* Thống kê độ dài của câu hỏi và độ dài của đoạn văn.
* Phân bố của nhãn.
* Kết Quả:
* Biểu đồ thể hiện độ dài của câu hỏi và đoạn văn.

A graph with numbers and a line

Description automatically generated with medium confidence

* Bảng thống kê phân bố của nhãn.

A blue square with white text

Description automatically generated

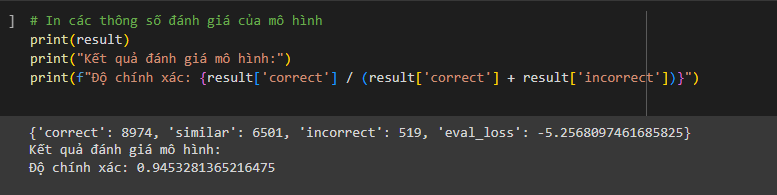
* Nhận Xét:
* Dữ liệu có sự đa dạng về độ dài của câu hỏi và đoạn văn.
* Phân bố của nhãn cần được cân nhắc để xây dựng mô hình hiệu quả.
* Kết luận
* Tất cả các đoạn văn đều có thể trích xuất ra đáp án dựa trên câu hỏi được hỏi (dựa vào bảng phân bố của nhãn)
* Trung bình độ dài câu hỏi khoảng 62 từ, độ dài đoạn văn khoảng 792 từ
* Trung bình số lượng câu hỏi trong đoạn văn khoảng 4 câu hỏi

## **phoBERT**

* Thực Hiện Thử Nghiệm:
* Tiến hành áp dụng mô hình vào bài toán đã cho trên bộ dữ liệu train.json và kiểm tra với bộ dữ liệu test.json
* Kết Quả:

A screenshot of a computer

Description automatically generated



* Nhận Xét:
* Mô hình chạy mượt và nhanh, không cần phải tiền xử lý dữ liệu thêm

## **Electra**

* Thực Hiện Thử Nghiệm:
* Tiến hành áp dụng mô hình vào bài toán đã cho trên bộ dữ liệu train.json và kiểm tra với bộ dữ liệu test.json
* Kết Quả:

A screenshot of a computer

Description automatically generated

A screen shot of a computer code

Description automatically generated

* Nhận Xét:
* Mô hình chạy mượt và nhanh, không cần phải tiền xử lý dữ liệu thêm

1. **So Sánh**

|  |  |
| --- | --- |
| **PhoBERT** | **ELECTRA** |
| * Cho ra độ chính xác cao, nhưng số lần huấn luyện sai cũng cao * eval\_loss và training\_loss thấp và ổn định * Độ chính xác thấp | * Cho ra độ chính xác thấp hơn * eval\_loss và training\_loss cao và đôi khi mất ổn định khi qua các bước kiểm tra * Độ chính xác cao |

1. **TÀI LIỆU THAM KHẢO**

* [**https://simpletransformers.ai/docs/qa-specifics/**](https://simpletransformers.ai/docs/qa-specifics/)
* [**https://arxiv-org.translate.goog/abs/2003.10555?\_x\_tr\_sl=en&\_x\_tr\_tl=vi&\_x\_tr\_hl=vi&\_x\_tr\_pto=sc**](https://arxiv-org.translate.goog/abs/2003.10555?_x_tr_sl=en&_x_tr_tl=vi&_x_tr_hl=vi&_x_tr_pto=sc)
* [**https://huggingface.co/NlpHUST/electra-base-vn**](https://huggingface.co/NlpHUST/electra-base-vn)
* [**https://huggingface.co/vinai/phobert-base**](https://huggingface.co/vinai/phobert-base)