NHẬP MÔN HỌC MÁY

Giảng viên hướng dẫn
TS.Bùi Tiến Lên
GV. Võ Nhật Tân
GV. Ngô Đức Bảo

THÀNH VIÊN

Hoàng Lê Minh Đăng	22127051
Lại Ngọc Phương Nam	22127283
Phạm Thành Nhân	22127307
Nguyễn Quang Sáng	22127364
Nguyễn Văn Tú	22127434

MỤC LỤC

1 GIỚI THIỆU

2 BÀI TOÁN

3 GIAO DIỆN

4 KẾT LUẬN

GIỚI THIỆU

VẤN ĐỀ THỰC TẾ

Tài chính - Ngân hàng - Bảo hiểm	7 ngành
Kinh tế - Quản trị kinh doanh - Thương Mại	18 ngành
Công nghệ thông tin - Tin học	23 ngành
Báo chí - Marketing - Quảng cáo - PR	11 ngành
Sư phạm - Giáo dục	35 ngành
Y - Dược	20 ngành
Bác sĩ thú y	3 ngành
Công an - Quân đội	30 ngành
Thiết kế đồ họa - Game - Đa phương tiện	5 ngành
Xây dựng - Kiến trúc - Giao thông	22 ngành

100+

HỆ QUẢ

Quyết định sai lệch

Lãng phí thời gian, tiền bạc, công sức

Nguy cơ thất nghiệp

Ý TƯỞNG

- Xây dựng hệ thống giúp học sinh hiểu rõ bản thân, tìm ngành học phù hợp.
- Xây dựng nền tảng tư vấn tuyển sinh thông minh, tích hợp AI và dữ liệu chính thống, minh bạch.

MỤC TIÊU

300+

Cung cấp thông tin về ngành học

150+

Cung cấp thông tin về trường đại học

Xây dựng lộ trình học tập

BÀI TOÁN

THU THẬP DỮ LIỆU

1

Thu thập URL website của các trường tại Việt Nam

2

Lấy thông tin từ website của mỗi trường

THU THẬP DỮ LIỆU

```
https://hust.edu.vn
https://ts.hust.edu.vn
https://www.ctu.edu.vn
https://tuyensinh.ctu.edu.vn
https://vsat.ctu.edu.vn
https://duytan.edu.vn
https://duytan.edu.vn/tuyen-sinh
https://hueuni.edu.vn
https://tuyensinh.hueuni.edu.vn
https://tnu.edu.vn
https://tuyensinh.tnus.edu.vn
https://vinhuni.edu.vn
https://tuyensinh.vinhuni.edu.vn
https://www.vnam.edu.vn
https://www.vnam.edu.vn/NewsDetail.aspx?lang=\u0026ItemID=1887
https://hvannd.edu.vn
https://hvannd.edu.vn/bv/ct/16032/thong-bao-chi-tieu-tuyen-sinh-tuyen-moi-dao-tao-trinh-do-dai-hoc-nam-2025-tai-hoc-vien-an-ninh-nhan-dan
https://daotaoajc.edu.vn
https://hvcsnd.edu.vn
```

THU THẬP DỮ LIỆU

Sau đó, sử dụng Crawl4AI để crawl dữ liệu về từ các url vừa chiết xuất được.

```
Đăng nhập
[![Đại học Bách khoa Hà Nội](https://hust.edu.vn/uploads/sys/logo-dhbk-1-02_130_191.png)](https://hust.edu.vn/vi/ "Đại học Bách khoa Hà Nội
  * [Đại học bách khoa Hà Nội](https://hust.edu.vn)
  * [HaNoi University of science and technology](https://hust.edu.vn)


  * [Đại học](https://hust.edu.vn)
  * [Bách khoa Hà Nội](https://hust.edu.vn)
  * [HaNoi University](https://hust.edu.vn)
  * [of science and technology](https://hust.edu.vn)


[![Đại học Bách khoa Hà Nội](https://hust.edu.vn/uploads/sys/logo-dhbk-1-02_130_191.png)](https://hust.edu.vn/vi/ "Đại học Bách khoa Hà Nội
# Đại học Bách khoa Hà Nội
## Đại học Bách khoa Hà Nội - HaNoi University of science and technology
  * [Tuyển dụng](https://tuyendung.hust.edu.vn/ "Tuyển dụng")
  * [Tài nguyên số](https://hust.edu.vn/vi/tai-nguyen-so/tai-nguyen-so/tai-nguyen-so-1.html "Tài nguyên số")
  * [Email](https://mail.hust.edu.vn/ "Email")
  * [Lịch công tác](https://hust.edu.vn/vi/lich-lam-viec/Truong-dai-hoc-BKHN/ "Lịch công tác")
  * [Liên hệ](https://hust.edu.vn/vi/contact/ "Liên hệ")


](https://hust.edu.vn/#flisearchform)
[ ](https://hust.edu.vn/vi/ "Tiếng Việt") [ ](https://hust.edu.vn/en/ "English")
  * [Giới thiệu](https://hust.edu.vn/vi/about/ "Giới thiệu")
    * [Thông điệp của Giám đốc Đại học](https://www.hust.edu.vn/vi/about/thong-diep-cua-giam-doc-dai-hoc.html "Thông điệp của Giám đốc Đại
    học")
    * [Tổng quan](https://hust.edu.vn/vi/about/tong-quan.html "Tổng quan")
      * [Lịch sử](https://hust.edu.vn/vi/about/cac-giai-doan-lich-su-191736.html "Lịch sử")
      * [Sứ mạng, Tầm nhìn, Giá trị cốt lõi và Chính sách chất lượng](https://hust.edu.vn/vi/about/
      su-mang-tam-nhin-gia-tri-cot-loi-va-chinh-sach-chat-luong.html "Sứ mạng, Tầm nhìn, Giá trị cốt lõi và Chính sách chất lượng")
      * [Chiến lược phát triển](https://hust.edu.vn/vi/about/chien-luoc-phat-trien.html "Chiến lược phát triển")
      * [Danh hiệu - Khen thưởng](https://hust.edu.vn/vi/about/danh-hieu-khen-thuong.html "Danh hiệu - Khen thưởng")
      * [Báo cáo thường niên](https://hust.edu.vn/vi/about/bao-cao-thuong-nien.html "Báo cáo thường niên")
    * [Cơ cấu tổ chức](https://hust.edu.vn/vi/co-cau-to-chuc-bai-viet/ "Cơ cấu tổ chức")
      * [BCH Đảng bộ Đại học](https://www.hust.edu.vn/vi/co-cau-to-chuc-bai-viet/bch-dang-bo-dai-hoc.html "BCH Đảng bộ Đại học")
```

Dữ liệu thu thập được vẫn chưa hoàn toàn sạch

TIỀN XỬ LÝ DỮ LIỆU

GPT-4o mini

Sử dụng API của OpenAI (GPT-4o mini) để tổng hợp và dọn dẹp dữ liệu từ các file markdown

```
# Đại học An ninh nhân dân (ANS)
## Giới thiệu
Trường Đại học An ninh nhân dân (ANS) là một trong những cơ sở đào tạo hàng đầu của ngành Công an, nổi tiếng về đào tạo các
chuyên ngành về an ninh và điều tra. Các ngành mạnh và uy tín nhất của trường bao gồm: An ninh mạng và phòng chống tội phạm
sử dụng công nghệ cao, Trinh sát an ninh, và Điều tra hình sự. Trường không được xếp hạng chính thức trong các bảng xếp
hạng đại học do tính chất đặc thù của ngành đào tạo. Tuy nhiên, ANS được đánh giá cao về chất lượng đào tạo, với đội ngũ
giảng viên có trình độ cao và cơ sở vật chất hiện đại. Nhiều sinh viên tốt nghiệp từ trường đã trở thành những cán bộ công
an xuất sắc, đảm nhiệm các vị trí quan trọng trong lực lượng Công an nhân dân.
## Tin Tức Mới Nhất
  * Website chính thức của trường: <https://dhannd.edu.vn>
  * Trang thông tin tuyển sinh: <https://dhannd.edu.vn/Thong-tin-tuyen-sinh>


## 1. Thay đổi chính trong năm 2025
### Chỉ tiêu tuyển sinh
  * Tổng chỉ tiêu tuyển sinh đại học chính quy: 260 chỉ tiêu (234 nam, 26 nữ)
  * Khu vực tuyển sinh: Phía Nam (từ Đà Nẵng đến Cà Mau)
  * Phân bổ theo vùng tuyển sinh:
    * Vùng 4 (các tỉnh miền Trung): 50 chỉ tiêu
    * Vùng 5 (các tỉnh Tây Nguyên): 40 chỉ tiêu
    * Vùng 6 (các tỉnh Đông Nam Bộ): 80 chỉ tiêu
    * Vùng 7 (các tỉnh Tây Nam Bộ): 75 chỉ tiêu
    * Vùng 8 (chiến sĩ nghĩa vụ): 15 chỉ tiêu


### Phương thức tuyển sinh
  * Bài thi đánh giá của Bộ Công an gồm 3 phần:
    * Tự luận bắt buộc: 1 câu nghị luận xã hội (25 điểm)
    * Trắc nghiệm bắt buộc:
      * Toán: 35 câu (35 điểm)
      * Lịch sử: 10 câu (10 điểm)
      * Ngoại ngữ (Tiếng Anh): 20 câu (15 điểm)
    * Trắc nghiệm tự chọn: 15 câu (15 điểm), chọn 1 trong 4 môn:
      * Vật lý (CA1)
      * Hóa học (CA2)
      * Sinh học (CA3)
      * Địa lý (CA4)
```

Theo format file markdown của nhóm gồm 5 mục chính: Giới thiệu, tin tức mới nhất, thông tin tuyển sinh, chương trình đào tạo, ngành học và lĩnh vực cho tất cả các trường đã tổng hợp được.

MÔ HÌNH SỬ DỤNG

Thông tin mô hình

- **Mô hình sử dụng**: Qwen2.5-7B-Instruct
- **Loại**: Tiền huấn luyện
- **Tham số**: 7.61B (6.53B không tính embedding)
- **Kiến trúc**: Transformer
- **Ngữ cảnh tối đa**: 131,072 tokens
- **Đầu ra tối đa**: 8,000 tokens

Ưu điểm nổi bật

- Hiểu tốt dữ liệu cấu trúc (bảng, JSON) dài
- Cải thiện khả năng tuân theo chỉ dẫn và tạo văn bản dài
- Hỗ trợ hơn 29 ngôn ngữ, bao gồm tiếng Việt
- Tăng khả năng xử lý prompt đa dạng

HUẤN LUYỆN MÔ HÌNH

Mục tiêu: Tinh chỉnh mô hình Qwen2.5-7B-Instruct để hiểu và phản hồi hội thoại tiếng Việt trong ngữ cảnh tư vấn đại học

Cấu hình mô hình

- **Mô hình gốc**: Qwen/Qwen2.5-7B-Instruct
- **Kỹ thuật tinh chỉnh**: PEFT + LoRA (Low-Rank Adaptation)
  - r=16, lora_alpha=32, dropout: 0.05
  - Mục tiêu vào các module: q_proj, v_proj
- **Cơ chế tối ưu**: FlashAttention 2 + Gradient Checkpointing
- **Kiểu dữ liệu**: float16

Tham số huấn luyện

- **Batch size**: 6 (với gradient accumulation = 2)
- **Epochs**: 7
- **Learning rate**: 1e-5
- **Data collator**: DataCollatorForCompletionOnlyLM

ĐÁNH GIÁ MÔ HÌNH

Chỉ số	Base Model	LoRA Fine-Tuned	Chênh lệch
BLEU	0.1225	**0.1342**	+0.0117
ROUGE-1	0.4488	**0.4625**	+0.0137
ROUGE-2	0.2628	**0.2709**	+0.0081
ROUGE-L	0.3542	**0.3628**	+0.0086

- **BLEU:** Chỉ số BLEU tăng nhẹ (+0.0117), cho thấy mô hình fine-tuned tạo ra các cụm từ gần hơn với đáp án tham chiếu, phản ánh khả năng học được các mẫu biểu đạt chính xác hơn sau tinh chỉnh.
- **ROUGE-1**: Tăng vừa phải (+0.0137), phản ánh độ bao phủ từ đơn (unigram recall) đã được cải thiện. Mô hình học được cách dùng từ sát ngữ cảnh hơn mà vẫn giữ được độ đa dạng.
- **ROUGE-2**: Tăng nhẹ (+0.0081), thể hiện khả năng tái tạo tốt hơn các cụm hai từ liên tiếp, góp phần làm câu mạch lạc và tự nhiên hơn.
- **ROUGE-L**: Tăng ổn định (+0.0086), cho thấy sự cải thiện về mặt cấu trúc ngôn ngữ trong đầu ra của mô hình sau tinh chỉnh

CÔNG NGHỆ SỬ DỤNG

FastAPI

FastAPI framework, high performance, easy to learn, fast to code, ready for production

tiangolo.com

unclecode/crawl4ai: 🚀🤖 Crawl4AI: Open-source LLM Friendly Web Crawler & Scraper. Don't be shy, join here:...

🚀🤖 Crawl4AI: Open-source LLM Friendly Web Crawler & Scraper. Don't be shy, join here: https://discord.gg/jP8KfhDhyN - unclecode/crawl4ai

GitHub

Qwen/Qwen2.5-7B · Hugging Face

We’re on a journey to advance and democratize artificial intelligence through open source and open science.

huggingface

HKUDS/LightRAG: "LightRAG: Simple and Fast Retrieval-Augmented Generation"

"LightRAG: Simple and Fast Retrieval-Augmented Generation" - HKUDS/LightRAG

GitHub

PostgreSQL

The world's most advanced open source database.

PostgreSQL / May 9

CÔNG NGHỆ SỬ DỤNG

React

React is the library for web and native user interfaces. Build user interfaces out of individual pieces called components written in JavaScript. React is designed to let you seamlessly combine component...

reactjs

Chakra UI

Simple, Modular & Accessible UI Components for your React Applications

chakra-ui.com

GIAO DIỆN

UI/UX

/

Trang chủ

Trang Về Chúng tôi

Trang đăng nhập

Trang đăng ký

UI/UX

/

Trang liệt kê các trường đang có

Trang thông tin chi tiết của trường

Trang Chatbox

KẾT LUẬN

TỔNG QUAN

- **Đáp ứng nhu cầu định hướng học vấn**: Hệ thống tư vấn tuyển sinh thông minh tích hợp AI, web crawling, RAG hỗ trợ học sinh THPT trong bối cảnh giáo dục đại học Việt Nam đa dạng.

- **Định hướng học vấn phù hợp**: Cá nhân hóa về ngành, trường, lộ trình học giúp giảm rủi ro quyết định sai lầm.

- **Dữ liệu tin cậy**: Được thu thập từ nguồn chính thống, xử lý nội dung có chọn lọc.

HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI

- Tăng cường độ chính xác của dữ liệu
- Cải tiến độ chính xác và khả năng xử lý Tiếng Việt chatbot AI
- Xây dựng ứng dụng di động
- Đảm bảo bảo mật dữ liệu
- Tăng cường tư vấn cá nhân hóa

THANK YOU!