KHOA KỸ THUẬT VÀ CÔNG NGHỆ **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN** **Độc lập – Tự do – Hạnh phúc**

**ĐỀ CƯƠNG CHI TIẾT**

**KHÓA LUẬN TỐT NGHIỆP NGÀNH CÔNG NGHỆ THÔNG TIN**

Họ tên sinh viên: **Phạm Hữu Lộc** MSSV: 110121055

Lớp: DA21TTB Khóa: 2021

Tên đề tài: XÂY DỰNG CƠ CHẾ PHÁT HIỆN AI-GENERATED CODE TRONG BÀI TẬP LẬP TRÌNH CỦA SINH VIÊN

**Mục tiêu của đề tài:** Xây dựng một công cụ hoặc quy trình giúp phát hiện code được sinh ra từ các công cụ AI (như ChatGPT, Copilot) trong các bài tập lập trình của sinh viên năm nhất cụ thể:

* Thu thập và xây dựng tập dữ liệu bao gồm mã nguồn thật từ sinh viên và mã nguồn do AI tạo.
* Phân tích đặc trưng của mã nguồn thông qua các phương pháp như: phân tích AST (Abstract Syntax Tree), độ phức tạp, độ lặp lại, cách đặt tên biến/hàm.
* Xây dựng hệ thống rule-based và/hoặc mô hình học máy nhẹ để đưa ra dự đoán code có phải do AI sinh ra không.
* Đánh giá độ chính xác và khả năng ứng dụng thực tế của công cụ, quy trình.

**Nội dung thực hiện:**

* Tìm hiểu các đặc trưng thường gặp của AI-generated code
* Thu thập và xây dựng tập dữ liệu gồm 2 nhóm:
* Mã nguồn của sinh viên năm nhất (C, Java, Python, …)
* Mã nguồn do các công cụ AI sinh (ChatGPT, Github Copilot, …)
* Trích xuất đặc trưng từ mã nguồn:
* AST (Abstract Syntax Tree)
* Cyclomatic Complexity
* Độ lặp từ khóa / cấu trúc
* Tên biến và hàm
* Xây dựng mô hình phát hiện:
* Phương pháp rule-based: định ngưỡng thủ công
* Phương pháp học máy nhẹ: Decision Tree / Logistic Regression (nếu đủ dữ liệu)
* Đánh giá kết quả
* So sánh giữa rule-based và machine learning
* Kiểm thử hệ thống trên các bài tập thực tế

**Phương pháp thực hiện:**

* Phương pháp phân tích và trích xuất đặc trưng
* Sử dụng AST để phân tích cấu trúc mã nguồn
* Tính toán các đặc trưng như: số lượng node AST, số lần xuất hiện của cấu trúc điều khiển, độ sâu cây AST, độ lặp mã.
* Phương pháp xây dựng mô hình phát hiện

**Bố cục đề tài:**

1. Tài liệu tham khảo:

1. Kế hoạch thực hiện đề tài

|  |  |  |  |
| --- | --- | --- | --- |
| **Tuần** | **Từ ngày - đến ngày** | **Công việc thực hiện** | **Ghi chú** |
| **Tuần 1** | 8/04 – 13/04 | **Phân tích đề tài** - Lựa chọn công nghệ - Lập kế hoạch - Viết đề cương chi tiết khóa luận |  |
| **Tuần 2** | 14/04 – 20/04 | **Nghiên cứu cơ bản** - Khảo sát các công cụ AI phổ biến (ChatGPT, Copilot, Deepseek, Remini, …) - Thu thập mẫu mã nguồn sinh từ AI - Tìm hiểu các phương pháp phân tích đặc trưng code (code fingerprint, syntax tree, cyclomatic complexity, clone detection) - Trao đổi với GVHD |  |
| **Tuần 3** | 21/04 – 27/04 | **Xây dựng tập dữ liệu mẫu** - Thu thập bài tập từ sinh viên - Sinh mẫu code từ AI với cùng đề bài - Phân loại thủ công mã AI vs mã con người - Báo cáo tiến độ tuần 2 với GVHD |  |
| **Tuần 4** | 28/04 – 04/05 | **Phân tích đặc trưng mã nguồn** - Viết tool Python để đo: + Độ phức tạp (radon) + Độ lặp lại (token-level) + Tỉ lệ comment, biến đặt tên chung chung - Định nghĩa bộ đặc trưng đầu tiên - Báo cáo tiến độ tuần 3 với GVHD |  |
| **Tuần 5** | 05/05 – 11/05 | **Phân tích và thống kê dữ liệu** - Chạy phân tích trên tập dữ liệu - Thống kê đặc trưng nào để phân biệt tốt giữa AI và người - Lập bảng, biểu đồ so sánh, nhận xét - Báo cáo tiến độ tuần 4 với GVHD |  |
| **Tuần 6** | 12/05 – 18/05 | **Xây dựng prototype cơ bản** - Viết tool CLI/Web nhỏ nhận input mã nguồn - Tính toán đặc trưng và xuất kết quả - Gợi ý cảnh báo nếu nghi ngờ là mã AI-generated - Báo cáo tiến độ tuần 5 với GVHD |  |
| **Tuần 7** | 19/05 – 25/05 | **Kiểm thử – tinh chỉnh thuật toán** - Thử nghiệm với nhiều mẫu code - Tối ưu ngưỡng cảnh báo - Loại bỏ false positive nhiều nhất có thể - Báo cáo tiến độ tuần 6 với GVHD |  |
| **Tuần 8** | 26/05 – 01/06 | **Phát triển giao diện người dùng**- Web app nhỏ để nhập mã và hiển thị kết quả - Giao diện đơn giản (Next.js/FastAPI) - Hoàn thiện chức năng toàn hệ thống - Báo cáo tiến độ tuần 7 GVHD |  |
| **Tuần 9** | 02/06 – 08/06 | **Kiểm thử toàn hệ thống + viết khóa luận** - Test thử toàn diện - Viết các phần chính của khóa luận: Cơ sở lý thuyết, Phương pháp tiếp cận, Phân tích dữ liệu - Báo cáo tiến độ tuần 8 với GVHD |  |
| **Tuần 10** | 09/06 – 11/06 | **Chỉnh sửa – Viết báo cáo hoàn chỉnh** - Viết phần đánh giá, kết luận, hướng phát triển - Gửi GVHD xem bản nháp |  |
| **Tuần 11** | 12/06 – 14/06 | **Sửa báo cáo theo phản hồi GVHD** - Hoàn thiện quyển khóa luận, chuẩn bị slide báo cáo - Gửi khóa luận hoàn chỉnh |  |
| **Tuần 12** | 15/06 – 16/06 | **Bảo vệ, nộp đồ khóa luận** - Chuẩn bị trình bày - Nộp source code, tài liệu, slide |  |

*Trà Vinh, ngày tháng 4 năm 2025*

**GIẢNG VIÊN HƯỚNG DẪN SINH VIÊN THỰC HIỆN**