ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO BÀI TẬP THỰC HÀNH 01

ĐỀ BÀI: **TIỀN XỬ LÝ DỮ LIỆU**

KHAI THÁC DỮ LIỆU & ỨNG DỤNG - CSC14004

NGƯỜI HƯỚNG DẪN

GS. LÊ HOÀI BẮC
ThS. LÊ NHỰT NAM
ThS. NGUYỄN NGỌC ĐỨC

THÔNG TIN SINH VIÊN

MSSV	Họ và Tên	Email
23127187	Lý Nhật Hào	lnhao23@clc.fitus.edu.vn
23127384	Huỳnh Lê Duy Khánh	hldkhanh23@clc.fitus.edu.vn
22127322	Lê Phước Phát	lpphat22@clc.fitus.edu.vn

THÀNH PHỐ HỒ CHÍ MINH - THÁNG 10, 2025

LỜI NÓI ĐẦU

Trong bối cảnh bùng nổ dữ liệu số, việc khai thác và phân tích dữ liệu hiệu quả trở thành yếu tố then chốt giúp doanh nghiệp nâng cao năng lực quản trị, tiếp thị và cải thiện trải nghiệm khách hàng. Chính vì vậy, việc tiền xử lý dữ liệu là một bước căn bản cực kỳ quan trọng trong bất kỳ hệ thống khai thác dữ liệu tự động và học máy, học sâu, hay các công nghệ trí tuệ nhân tạo hiện đại ngày nay. Thông thường những dữ liệu thô ngoài thực tế thường không có giá trị sâu sắc. Chúng thường có nhiều dạng khác nhau từ các đầu vào khác nhau như các cảm biến, ứng dụng, ... Đối với từng dạng và cấu trúc dữ liệu thô đầu vào, chúng yêu cầu các kỹ thuật tiền xử lý khác nhau.
Xuất phát từ nhu cầu hiện thực và cấp thiết đó, thông qua bài tập thực hành này, bài tập này sẽ cung cấp các kiến thức và kỹ năng cần thiết để xử lý các dạng dữ liệu sau thông qua các chương sau:

- **CHƯƠNG 01. Tiền xử lý dữ liệu hình ảnh**
  - Ở cấp độ này, chúng ta cần phải tiền xử lý dữ liệu với đầu vào là những điểm pixel có thể chuyển hoá được.
- **CHƯƠNG 02. Tiền xử lý dữ liệu dạng bảng**
  - Ở cấp độ này, chúng ta sẽ thực hiện việc xử lý dữ liệu đối với dữ liệu có cấu trúc (dạng bảng) với những cột và hàng với nhiều loại dữ liệu khác nhau.
- **CHƯƠNG 03. Tiền xử lý dữ liệu dạng văn bản**
  - Ở cấp độ này, chúng ta sẽ thực hiện việc xử lý dữ liệu không có cấu trúc cố định mà dữ liệu này yêu cầu sự hiểu ngôn ngữ một cách tự nhiên.

Tóm lại, thông qua bài tập này, chúng ta sẽ phát triển những kỹ năng thực hành bằng cách áp dụng các kỹ thuật tiền xử lý thích hợp đối với nhiều dạng dữ liệu khác nhau. Đồng thời, chúng ta có thể hiểu và vận dụng những kỹ thuật này cho những tác vụ đòi hỏi việc phân tích và khai thác dữ liệu phức tạp.

Thành phố Hồ Chí Minh, Mùa hè 2025.

LỜI CAM ĐOAN

Nhóm số 05 thực hiện bài tập thực hành số 01 về chủ đề **Tiền xử lý dữ liệu** gồm các thành viên **Lê Phước Phát**, **Lý Nhật Hào**, và **Huỳnh Lê Duy Khánh** đều là sinh viên thuộc khoa Công nghệ Thông tin Chất lượng cao, thuộc trường Đại học Khoa học Tự nhiên, ĐHQG-HCM. Nhóm cam đoan rằng bài tập nghiên cứu này là do các thành viên trong nhóm tìm hiểu, nghiên cứu và thực hiện dưới sự giám sát và hướng dẫn của thầy **GS. LÊ HOÀI BẮC**, thầy **ThS. LÊ NHỰT NAM**, và thầy **ThS. NGUYỄN NGỌC ĐỨC**. Các dữ liệu được nêu trong đồ án là hoàn toàn trung thực, phản ánh đúng kết quả mô phỏng thực tế. Tất cả các tài liệu được sử dụng trong nghiên cứu này được các thành viên trong nhóm thu thập bằng cách tự thân và từ các nguồn khác nhau, và những tài liệu này được liệt kê đầy đủ trong phần tài liệu tham khảo. Tất cả đều được trích dẫn đúng đắn. Trong trường hợp có vi phạm bản quyền, các thành viên trong nhóm sẽ chịu trách nhiệm cho hành động đó. Do đó, trường **Đại học Khoa học Tự nhiên, ĐHQG-HCM** không chịu trách nhiệm về bất kỳ vi phạm bản quyền nào được thực hiện trong bài tập nghiên cứu này.

TP.HCM, ngày 28 tháng 10 năm 2025

Người cam đoan

Nhóm trưởng

LÝ NHẬT HÀO

MỤC LỤC

DANH MỤC KÝ HIỆU & CHỮ VIẾT TẮT

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH MỤC BIỂU ĐỒ

THÔNG TIN NHÓM & BẢNG PHÂN CÔNG

Thông Tin Nhóm

Nhóm số 05 bao gồm 3 thành viên với các thông tin chi tiết sau:

- Lê Phước Phát - 22127322
- Trần Lý Nhật Hào - 23127187
- Huỳnh Lê Duy Khánh - 23127384

Bảng Phân Công Công Việc

Bảng 1.1 Bảng Phân Công Việc

MSSV	Họ và Tên	Công Việc	Mức độ
22127322	Lê Phước Phát	**[Phần thực thi chương trình (code)]** • Tiền xử lý dữ liệu văn bản **(Twitter 15 and 16)** **[Phần viết tài liệu báo cáo]** • Viết báo cáo phần **CHƯƠNG 03: TIỀN XỬ LÝ VĂN BẢN.** • Viết hướng dẫn cài đặt README.md	100%
23127187	Trần Lý Nhật Hào	**[Phần thực thi chương trình (code)]** • Tiền xử lý dữ liệu hình ảnh *(điền tên bộ dataset vào nha ...)* **[Phần viết tài liệu báo cáo]** • Viết báo cáo phần **CHƯƠNG 01. TIỀN XỬ LÝ DỮ LIỆU HÌNH ẢNH.** • Viết hướng dẫn cài đặt README.md	100%
23127384	Huỳnh Lê Duy Khánh	**[Phần thực thi chương trình (code)]** • Tiền xử lý dữ liệu dạng bảng *World Cities Population and Location* **[Phần viết tài liệu báo cáo]** • Viết báo cáo phần **CHƯƠNG 02. TIỀN XỬ LÝ DỮ LIỆU DẠNG BẢNG.** • Viết hướng dẫn cài đặt README.md	100%

Mô tả báo cáo thực hành

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

CHƯƠNG 01

TIỀN XỬ LÝ DỮ LIỆU HÌNH ẢNH

Trong chương này, nhóm chúng em sẽ trình bày về các kỹ thuật tiền xử lý dữ liệu dạng hình ảnh.

ĐIỀN THÔNG TIN TÓM TẮT Ở ĐÂY ...

1.1 Mô tả dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.1.1 Giới tiệu về tập dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.1.2 Hướng dẫn cài đặt dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.1.3 Phân tích tổng quan dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.1.4 Lý do chọn tập dữ liệu này ?

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.2 Cơ sở tiền xử lý

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.3 Triển khai chi tiết

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.4 Phân tích & đánh giá kết quả

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.4.1 Phân tích kết quả định lượng

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.4.2 So sánh trực quan và diễn giải kết quả

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

1.4.3 Phân tích tác động về chất lượng dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

CHƯƠNG 02

TIỀN XỬ LÝ DỮ LIỆU DẠNG BẢNG

ĐIỀN THÔNG TIN CẦN THIẾT Ở ĐÂY ...

Trong chương này, nhóm chúng em sẽ trình bày về các kỹ thuật tiền xử lý dữ liệu dạng bảng.

ĐIỀN THÔNG TIN TÓM TẮT Ở ĐÂY ...

2.1 Mô tả dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.1.1 Giới tiệu về tập dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.1.2 Hướng dẫn cài đặt dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.1.3 Phân tích tổng quan dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.1.4 Lý do chọn tập dữ liệu này ?

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.2 Cơ sở tiền xử lý

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.3 Triển khai chi tiết

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.4 Phân tích & đánh giá kết quả

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.4.1 Phân tích kết quả định lượng

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.4.2 So sánh trực quan và diễn giải kết quả

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

2.4.3 Phân tích tác động về chất lượng dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

CHƯƠNG 03

TIỀN XỬ LÝ DỮ LIỆU VĂN BẢN

ĐIỀN THÔNG TIN CẦN THIẾT Ở ĐÂY ...

Trong chương này, nhóm chúng em sẽ trình bày về các kỹ thuật tiền xử lý dữ liệu dạng văn bản.

ĐIỀN THÔNG TIN TÓM TẮT Ở ĐÂY ...

3.1 Mô tả dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.1.1 Giới tiệu về tập dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.1.2 Hướng dẫn cài đặt dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.1.3 Phân tích tổng quan dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.1.4 Lý do chọn tập dữ liệu này ?

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.2 Cơ sở tiền xử lý

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.3 Triển khai chi tiết

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.4 Phân tích & đánh giá kết quả

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.4.1 Phân tích kết quả định lượng

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.4.2 So sánh trực quan và diễn giải kết quả

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

3.4.3 Phân tích tác động về chất lượng dữ liệu

ĐIỀN THÔNG TIN CHI TIẾT Ở ĐÂY ...

KẾT LUẬN & ĐỀ NGHỊ

Kết luận chung

ĐIỀN THÔNG TIN CẦN THIẾT Ở ĐÂY ...

Hướng phát triển

ĐIỀN THÔNG TIN CẦN THIẾT Ở ĐÂY ...

Kiến nghị và đề xuất

ĐIỀN THÔNG TIN CẦN THIẾT Ở ĐÂY ...

TÀI LIỆU THAM KHẢO