|  |  |  |
| --- | --- | --- |
|  | **TRƯỜNG ĐẠI HỌC BÁCH KHOA**  **KHOA CÔNG NGHỆ THÔNG TIN** |  |

TIỂU LUẬN CUỐI KỲ

HỌC PHẦN: KHOA HỌC DỮ LIỆU

**TÊN ĐỀ TÀI**

|  |  |
| --- | --- |
| Nhóm | 9 |
| Họ Và Tên Sinh Viên | Lớp Học Phần |
| … | 2011A |
| … |
| … |

ĐÀ NẴNG, 06/2023

**TÓM TẮT**

Sinh viên mô tả tóm tắt tiểu luận trong 01 đoạn văn bản, từ 7 đến 10 dòng, nêu 03 ý lớn: vấn đề cần giải quyết, phương pháp giải quyết và kết quả đạt được. Phần này trình bày trên 01 trang riêng, ngay sau trang bìa.

**BẢNG PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện | Các nhiệm vụ | Tự đánh giá theo 3 mức  (Đã hoàn thành/Chưa hoàn thành/Không triển khai) |
| Lê Nguyễn Tiến Anh | -…   * … | -…  -… |
| Dương Quốc Bảo |  |  |
| Hoàng Nhật Bảo Long | - Crawl dữ liệu bằng API của Tiki | - Đã hoàn thành |

**MỤC LỤC**

**CÁC YÊU CẦU VIẾT BÁO CÁO**

**VỀ HÌNH THỨC:**

* Định dạng: căn chỉnh lề kiểu Justify; font chữ 12, times new roman; giãn dòng 1.5, cách trước 6pt; hình ảnh phải đánh số, có chú thích bên dưới hình; bảng phải đánh số, có chú thích bên trên bảng.
* Báo cáo có độ dài từ 15 đến 20 trang và không đóng bìa gương (chỉ cần bìa mềm có màu) gồm: trang bìa (trang 1 của file hướng dẫn này), tóm tắt, bảng phân công nhiệm vụ,mục lục, và các trang nội dung.

**VỀ NỘI DUNG:**

Báo cáo gồm 06 phần sau.

# 1. Giới thiệu

Trong phần này, sinh viên giới thiệu vấn đề cần giải quyết và đề xuất giải pháp tổng quan cần thực hiện.

# 2. Thu thập và mô tả dữ liệu

**2.1. Thu thập dữ liệu**

|1. Nguồn dữ liệu: Nguồn dữ liệu được sử dụng là Tiki - một trang web thương mại điện tử hàng đầu tại Việt Nam, cung cấp hàng ngàn sản phẩm đa dạng từ đồ điện tử, gia dụng, thời trang, sách, ...

2. Công cụ thu thập: Công cụ được sử dụng để thu thập dữ liệu là ngôn ngữ lập trình Python cùng với các thư viện: requests, pandas, csv, retrying. Công cụ này sử dụng API của Tiki để truy xuất và lấy dữ liệu từ trang web này thông qua các yêu cầu HTTP,

3. Các bước sử dụng công cụ

Các bước để thu thập dữ liệu từ Tiki qua API bao gồm:

* Tạo session để sử dụng cho tất cả các yêu cầu API của Tiki
* Đặt số lần retry mặc định là 10 để xử lý các lỗi liên quan đến mạng hoặc API phản hồi lỗi
* Xác định các headers, params, và body cần thiết cho từng yêu cầu API
* Sử dụng hàm retrying để xử lý các lỗi xảy ra trong quá trình truy xuất khác nhau ΑΡΙ
* Tiến hành lấy dữ liệu bằng cách xác định các key chính của các đối tượng trả về, sau đó lưu trữ dữ liệu vào file CSV
* Thực hiện các bước trên cho tất cả các sản phẩm được hiển thị trên trang web Tiki qua các yêu cầu API khác nhau

4. Đầu vào và đầu ra của quá trình thu thập

Đầu vào: Các yêu cầu API của Tiki được cấu trúc thông qua các headers, params, và body để truy xuất dữ liệu.

- Đầu ra: Các dữ liệu thu thập được sẽ được lưu trữ trên file CSV, các trường bao gồm id, tên sản phẩm, thương hiệu, giá gốc, giá giảm, giá hiện tại, tỷ lệ giảm giá, số lượng bán, đánh giá trung bình, số lượng đánh giá, mã nhà bán hàng, tên nhà bán hàng, ngày tham gia của nhà bán hàng, đánh giá trung bình của nhà bán hàng, số lượng đánh giá của nhà bán hàng và tổng số người theo dõi.

5. Ví dụ minh hoạ

Mã nguồn đã đưa ra ở câu trả lời trước cung cấp một ví dụ minh họa về cách sử dụng công cụ Python để lấy dữ liệu từ Tiki thông qua API. Ví dụ này đưa ra các bước trình bày trên để minh họa việc thu thập dữ liệu.

**2.2. Mô tả dữ liệu**

SV mô tả các thống kê tổng quan về tập dữ liệu (số mẫu, số đặc trưng của 1 mẫu, kiểu dữ liệu của mỗi đặc trưng, số mẫu dữ liệu trống của mỗi đặc trưng,...) và xuất ra các thống kê mô tả trực quan về các đặc trưng (ví dụ: dùng boxplot, histogram, scatter plot về độ tương quan,…).

# 3. Trích xuất đặc trưng

SV trình bày việc lựa chọn đặc trưng, làm sạch và chuẩn hóa dữ liệu, giảm chiều dữ liệu, …; và trực quan hoá kết quả/tác dụng của các quá trình trên (trước và sau khi xử lý).

# 4. Mô hình hóa dữ liệu

SV trình bày quá trình xây dựng, xác thực và kiểm thử mô hình gồm:

- Chọn các mô hình phù hợp (ít nhất 2 mô hình) và mô tả ngắn gọn cơ sở lý thuyết của các mô hình đã lựa chọn, mô tả bộ tham số của mỗi mô hình (ví dụ: số lớp của mạng nơ-ron, số node của mỗi lớp, hàm activation được sử dụng,…) thông qua API sử dụng để xây dựng mô hình.

- Chia dữ liệu thành các tập Huấn luyện/Xác thực/Kiểm thử theo tỉ lệ phù hợp.

- Trình bày các tham số của quá trình huấn luyện mô hình (ví dụ: loss function, learning rate, number of epochs,…).

- Trình bày các đồ thị thể hiện hiệu suất (đánh giá bằng Loss hoặc Accuracy) của các mô hình trên các tập Huấn luyện/Xác thực/Kiểm thử.

- Đưa ra các độ đo (metrics) dùng để đánh giá các mô hình (ví dụ: RMSE hoặc MAPE), và so sánh hiệu quả của các mô hình bằng bảng số liệu (hoặc đồ thị) dựa trên các độ đo.

# 5. Kết luận

Sinh viên tổng kết các việc đã làm và các kết quả đạt được. Nêu hướng phát triển để cải thiện kết quả hiện có.

# 6. Tài liệu tham khảo

SV liệt kê các TLTK đã trích dẫn (cite) trong báo cáo tại đây.