TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO GIỮA KÌ MÔN**

**XỬ LÝ DỮ LIỆU LỚN**

*Người hướng dẫn*: **TS. Bùi Thanh Hùng**

*Người thực hiện*: **Trần Quốc Tâm – 51800721**

*Lớp:* **18050402**

Khóa: **22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO GIỮA KÌ MÔN**

**XỬ LÝ DỮ LIỆU LỚN**

*Người hướng dẫn*: **TS. Bùi Thanh Hùng**

*Người thực hiện*: **Trần Quốc Tâm – 51800721**

*Lớp:* **18050402**

Khóa: **22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

# LỜI CẢM ƠN

Em xin gởi lời cảm ơn đến với thầy Bùi Thanh Hùng, giảng viên lý thuyết kiêm người hướng dẫn em trong quá trình làm báo cáo môn Xử lý dữ liệu lớn.

Nhờ có sự dẫn dắt từ thầy trong thời gian qua mà em đã có được rất nhiều kiến thức bổ ích về môn học này. Xin cảm ơn thầy đã hỗ trợ để em có thể hoàn thành bài báo cáo này.

Do còn ít kinh nghiệm nên bài báo cáo đôi chỗ vẫn còn sai sót, em mong có được sự góp ý của thầy để em có thể hoàn thiện hơn nữa khả năng viết báo cáo của mình sau này. Một lần nữa, em xin chân thành cảm ơn.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH  
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của giảng viên Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Khóa luận/Đồ án tốt nghiệp còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Khóa luận/Đồ án tốt nghiệp của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Trần Quốc Tâm*

# PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

# TÓM TẮT

# MỤC LỤC

[LỜI CẢM ƠN i](#_Toc99732490)

[ĐỒ ÁN ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG ii](#_Toc99732491)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc99732492)

[TÓM TẮT iv](#_Toc99732493)

[MỤC LỤC 1](#_Toc99732494)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 2](#_Toc99732495)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 3](#_Toc99732496)

[CHƯƠNG 1 - Thu Thập Dữ Liệu 4](#_Toc99732497)

[1.1 Cấu trúc của trang web 4](#_Toc99732498)

[TÀI LIỆU THAM KHẢO 5](#_Toc99732499)

# DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

**CÁC CHỮ VIẾT TẮT**

# DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC BẢNG**

**DANH MỤC HÌNH**

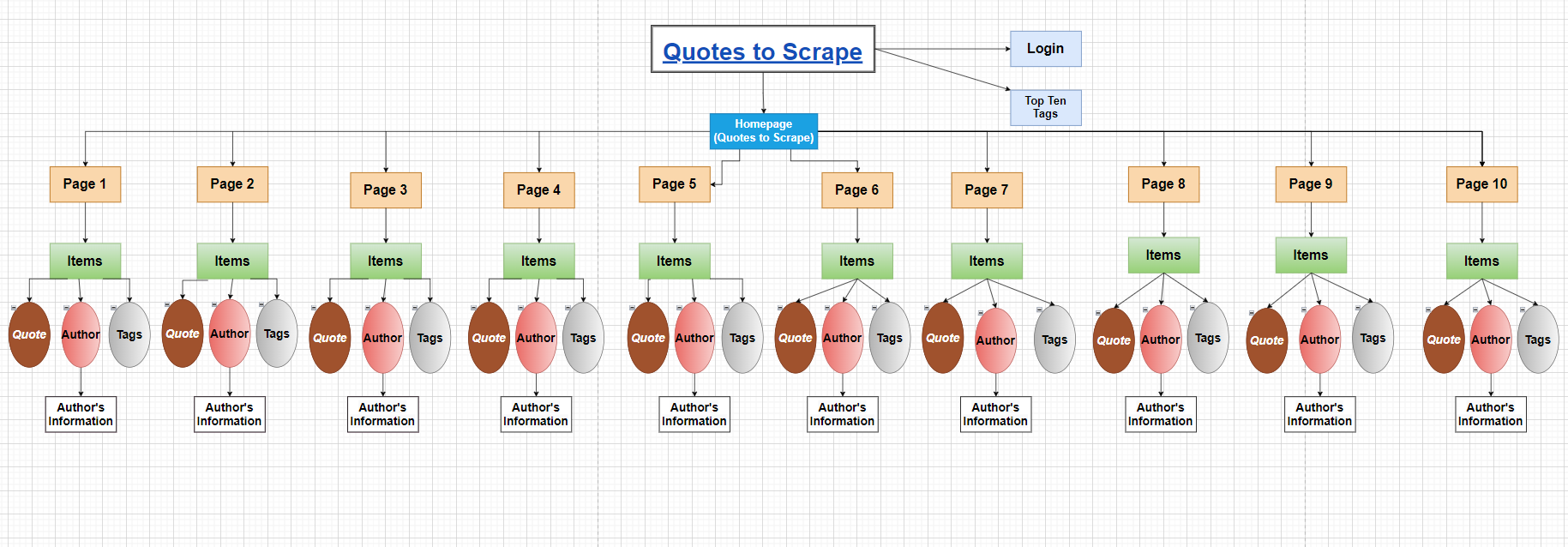
[Hình 1. 1: Sơ đồ cấu trúc của trang web 4](#_Toc99916302)

CHƯƠNG 1 - Thu Thập Dữ Liệu

## Cấu trúc của trang web

Trang web “<https://quotes.toscrape.com>” được cấu tạo gồm 4 phần: Trang chủ (Quotes to Scrape), phần đăng nhập (Login), phần mười thẻ hàng đầu (top ten tag) và phần các câu quote. Trang web này có 10 trang, ở mỗi trang sẽ có trang chủ, phần login, phần top ten tags và một danh sách gồm 10 item, trong đó mỗi item sẽ chứa một cấu quote, tác giả của câu quote đó và các tag thuộc vào nội dung của câu quote đó. Mỗi item sẽ cho phép người dùng tìm hiều về thông tin của tác giả của từng câu quote bằng việc nhấp vào (about) ở phía sau tên của mỗi tác giả và cũng như tìm kiếm các câu quote của cùng tag ở mỗi item.

Sơ đồ cấu trúc của trang web “<https://quotes.toscrape.com>”:



Hình 1. 1: Sơ đồ cấu trúc của trang web

# TÀI LIỆU THAM KHẢO

<https://quotes.toscrape.com>