|  | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM** |
| --- | --- |

**ĐỒ ÁN MÔN HỌC**

**<TÌM HIỂU VỀ SCRAPY FRAMWORD >**

Ngành: **<KHOA HỌC DỮ LIỆU >**

Chuyên ngành: **<MÃ NGUỒN MỞ TRONG KHOA HỌC DỮ LIỆU >**

Giảng viên hướng dẫn :ThS. LÊ NHẬT TÙNG

Sinh viên thực hiện :

2286400003-NGUYỄN HOÀNG ANH

2286400041-BẠCH QUANG TÙNG

2286400004-NGUYỄN TUẤN ANH

Lớp: 22DKHA1

TP. Hồ Chí Minh, <2024>

1.1

**Giới thiệu tổng quan:** Scrapy là một framework mã nguồn mở mạnh mẽ dành cho việc thu thập dữ liệu web. Với khả năng thu thập thông tin từ nhiều trang web khác nhau, Scrapy được ứng dụng rộng rãi trong các lĩnh vực như phân tích dữ liệu, SEO, và nghiên cứu thị trường

**1.2 Nhiệm vụ của đề tài**

**1.2.1 Tính cấp thiết của đề tài**

Trong kỷ nguyên số hiện nay, dữ liệu trên mạng xã hội và các nền tảng trực tuyến đang gia tăng một cách chóng mặt. Việc thu thập và phân tích các xu hướng, từ khóa và chủ đề nổi bật không chỉ giúp hiểu rõ hành vi người dùng mà còn hỗ trợ trong việc ra quyết định cho doanh nghiệp và tổ chức. Do đó, việc phát triển công cụ "Scrapy Fameword" trở nên cần thiết để đáp ứng nhu cầu này, giúp người dùng có cái nhìn sâu sắc hơn về các xu hướng đang diễn ra.

**1.2.2 Ý nghĩa khoa học và thực tiễn của đề tài**

**Ý nghĩa khoa học:**

* + Đề tài góp phần nâng cao hiểu biết về công nghệ thu thập dữ liệu và xử lý ngôn ngữ tự nhiên, thúc đẩy nghiên cứu trong lĩnh vực này.
  + Cung cấp những phương pháp mới trong phân tích dữ liệu lớn, từ đó mở rộng hướng nghiên cứu cho các lĩnh vực liên quan như khoa học dữ liệu và phân tích ngữ nghĩa.

**Ý nghĩa thực tiễn:**

* + Công cụ phát triển sẽ giúp doanh nghiệp theo dõi và phân tích nhanh chóng các xu hướng, từ đó đưa ra chiến lược marketing hiệu quả hơn.
  + Hỗ trợ nhà nghiên cứu và các tổ chức trong việc nắm bắt thông tin thị trường, cải thiện sản phẩm và dịch vụ dựa trên dữ liệu thu thập được.

**1.3 Mục tiêu**

**1.3.1 Mục tiêu tổng quát**

Mục tiêu tổng quát của đề tài "Scrapy Fameword" là phát triển một công cụ thu thập và phân tích dữ liệu từ các nền tảng trực tuyến nhằm xác định và đánh giá các xu hướng nổi bật trong cộng đồng mạng.

**1.3.2 Mục tiêu cụ thể**

* Thiết kế và triển khai spider sử dụng Scrapy để thu thập dữ liệu từ các trang mạng xã hội và blog.
* Phân tích và xử lý dữ liệu thu thập được để xác định các từ khóa và chủ đề nổi bật.
* Tạo báo cáo trực quan hóa kết quả phân tích để cung cấp thông tin hữu ích cho người dùng.

**1.4 Đối tượng và phạm vi**

**1.4.1 Đối tượng**

Đối tượng nghiên cứu của đề tài bao gồm các trang mạng xã hội, diễn đàn và blog nơi có sự trao đổi thông tin và ý kiến giữa người dùng, cụ thể là các bài viết, bình luận và phản hồi.

**1.4.2 Phạm vi**

Phạm vi nghiên cứu sẽ giới hạn ở các nền tảng mạng xã hội phổ biến và các blog liên quan đến các lĩnh vực như công nghệ, giải trí, thời trang, và sức khỏe trong khoảng thời gian nhất định.

**1.5 Phương pháp nghiên cứu**

**1.5.1 Phương pháp nghiên cứu sơ bộ**

Tiến hành khảo sát sơ bộ về các công cụ và kỹ thuật thu thập dữ liệu hiện có để xác định những gì phù hợp cho đề tài.

**1.5.2 Phương pháp nghiên cứu tài liệu**

Tham khảo các tài liệu, báo cáo và nghiên cứu trước đó liên quan đến Scrapy, thu thập dữ liệu web và phân tích dữ liệu ngữ nghĩa.

**1.5.3 Phương pháp nghiên cứu thống kê**

Sử dụng các phương pháp thống kê để phân tích dữ liệu thu thập được, nhằm xác định các xu hướng và mẫu.

**1.5.4 Phương pháp thực nghiệm**

Tiến hành các thí nghiệm thu thập dữ liệu thực tế để kiểm nghiệm tính hiệu quả và độ chính xác của công cụ phát triển.

**1.5.5 Phương pháp đánh giá**

Đánh giá kết quả thu thập và phân tích dữ liệu thông qua các chỉ số đo lường và phản hồi từ người dùng.

**1.6 Những đóng góp nghiên cứu của đề tài**

**1.6.1 Trong lĩnh vực học thuật**

* Góp phần làm phong phú thêm kiến thức về thu thập dữ liệu và phân tích ngữ nghĩa trong nghiên cứu khoa học.
* Mở ra hướng nghiên cứu mới cho các vấn đề liên quan đến dữ liệu lớn và hành vi người dùng trên mạng.

**1.6.2 Trong thực tiễn kinh doanh**

* Cung cấp một công cụ hữu ích cho doanh nghiệp trong việc theo dõi xu hướng thị trường, từ đó tối ưu hóa chiến lược marketing.
* Hỗ trợ trong việc ra quyết định nhanh chóng dựa trên dữ liệu thực tế, nâng cao khả năng cạnh tranh trên thị trường.

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .**

TPHCM,Ngày………tháng……năm 2024

**Giáo viên hướng dẫn**

( Ký tên , đóng dấu )

### LỜI CAM ĐOAN

Nhóm của tôi xin cam đoan rằng:

Mọi thông tin và kết quả nghiên cứu được trình bày trong bài báo cáo này đều trung thực và khách quan, được thu thập và phân tích một cách cẩn trọng dựa trên các nguồn chính thống và đáng tin cậy.

Tôi khẳng định rằng tất cả các thông tin hoặc ý kiến được trích dẫn từ các nguồn khác đều được nêu rõ nguồn gốc và trích dẫn theo đúng quy định. Tôi cam đoan rằng không có bất kỳ sự sao chép hoặc sử dụng thông tin không đúng đắn nào từ các nguồn khác.

Bài báo cáo này là công trình nghiên cứu độc lập của tôi, chưa từng được công bố ở bất kỳ nơi nào khác. Tôi cam kết đã tuân thủ đầy đủ các quy tắc và quy định của môn học, bao gồm cả việc tham khảo và sử dụng các công cụ nghiên cứu như Beautiful Soup một cách trung thực và khoa học.

Tôi hy vọng rằng bài báo cáo này sẽ cung cấp một cái nhìn tổng quan rõ ràng và toàn diện về chủ đề “Sử dụng Beautiful Soup trong thu thập dữ liệu website” và sẽ đóng góp một phần nhỏ vào lĩnh vực nghiên cứu này.

TPHCM,Ngày………tháng……năm 2024

Sinh viên 1 Sinh viên 2 Sinh viên 3

**DANH MỤC CÁC KÝ HIỆU , CÁC CHỮ VIẾT TẮT VÀ TỪ KHÓA**

**MỤC LỤC**

**1. TỔNG QUAN**

**2. CƠ SỞ LÝ THUYẾT**

**3. PHƯƠNG PHÁP THỰC NGHIỆM**

**4. KẾT QUẢ THỰC NGHIỆM**

**5. KẾT LUẬN VÀ KIẾN NGHỊ**

**DANH SÁCH BẢNG**

**DANH SÁCH HÌNH VẼ**