CHƯƠNG 1: MỞ ĐẦU

CHƯƠNG 2:

CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU

Để phục vụ cho một bài toán lớn là việc xây dựng lên một hệ thống thu thập, phân tích và đánh giá tin tức liên quan đến chủ đề An ninh quốc gia - Trật tự an toàn xã hội. Nhóm chúng em đã lựa chọn một chiến lược nghiên cứu đó chính là chia nhỏ bài toán lớn ấy ra thành những bài toán nhỏ hơn để giải quyết vấn đề theo từng chức năng. Cụ thể được chia thành 3 bài toán con như sau:

* Bài toán thu thập tin tức tự động từ trang mạng xã hội Facebook
* Phân loại và đánh giá những tin tức đã thu thập được
* Xây dựng hệ thống có chức năng lưu trữ, quản lý, thống kê và báo cáo về những tin tức đã được phân loại.

Như vậy, việc phân chia một bài toán lớn thành các module (mô-đun) nhỏ hơn giúp cho việc nghiên cứu trở lên tường minh và cụ thể. Trong đó ta có thể thấy output ( dữ liệu đầu ra) của bài toán con thứ nhất chính là input (dữ liệu đầu vào) của bài toán con thứ hai và tiếp tục như thế, dữ liệu của bài toán thứ hai sau khi được giải quyết sẽ trở thành dữ liệu để triển khai bài toán thứ ba. Cụ thể các phương pháp nghiên cứu sẽ được trình bày một cách chi tiết như sau:

1.Bài toán thu thập tin tức tự động từ trang mạng xã hội Facebook

1.1Thu thập tin tức tự động (hay còn gọi là quét dữ liệu tự động) là gì?

Hiện nay, nhu cầu về việc thu thập những tin tức, bài báo từ trên internet là rất nhiều. Nhu cầu này chủ yếu đến từ các doanh nghiệp, công ty và các tổ chức về tài chính. Mục đích sau khi thu thập dữ liệu là để phân tích về kinh doanh, nhu cầu của người dùng, phân tích về xu hướng trên thị trường. Nắm bắt được nhu cầu đó, rất nhiều những công cụ thu thập tin tức đã ra đời trong đó có thể kể đến như Cisco, Reputa. Nhưng hiện tại các công cụ như thế vẫn đang hoạt động chính ở trên các website của các trang báo điện tử và giá thành để thuê dịch vụ cũng tương đối lớn. Vậy thì thu thập tin tức tự động nghĩa là gì?

Thu thập tin tức tự động (quét dữ liệu tự động) là một thuật ngữ trong ngành Công nghệ thông tin, đây được xem là một quá trình khá quan trọng trong việc thu thập và lấy dữ liệu từ một trang web rồi phân tích dữ liệu lấy được và bóc tách thông tin dữ liệu theo yều cầu đã được đặt ra. Có thể hiểu như thế này, chúng ta muốn lấy thông tin từ một trang web nào đó thì phần quét dữ liệu sẽ cung cấp chức năng lấy được dữ liệu khi đang kết nối internet, sau đó tất cả dữ liệu thu thập được sẽ được lưu vào database một cách tự động mà không phải thông qua bất cứ một chi tiết nhập liệu nhập liệu nào cả - phương pháp này được coi là ưu điểm của giúp tiết kiệm thời gian cho người sử dụng. Nói một cách dễ hiểu, quét dữ liệu là phương pháp giúp chúng ta lấy thông tin khi biết đường link của website mà chúng ta muốn lấy thông tin, phần còn lại sẽ diễn ra một cách tự động.

* 1. Các cơ chế, phương pháp thu thập tin tức tự động

Đối với các kĩ thuật được sử dụng để thu thập dữ liệu tự động, có một vấn đề chung được đặt ra đó chính là các website chúng ta muốn lấy thông tin đều không có API kết nối trực tiếp để lấy dữ liệu từ database của webiste đó. Vì vậy cơ chế duy nhất đó chính là phân tích cấu trúc code, cấu trúc HTML của website đó để lấy được dữ liệu mong muốn.

Về phương pháp triển khai, mục tiêu đặt ra của đề tài là thu thập được tin tức được đăng trên các trang mạng xã hội (cụ thể là mạng xã hội Facebook). Do đó công đoạn đầu tiên là phân tích cấu trúc trang web Facebook

1.3

CHƯƠNG 4: XÂY DỰNG VÀ TRIỂN KHAI HỆ THỐNG

1. Xây dựng database lưu trữ tin tức
2. Xây dựng công cụ thu thập – phân loại – đánh giá tin tức tự động
3. Xây dựng Web API
4. Xây dựng Web-site quản lý, thống kê, báo cáo về tin tức