BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

Báo Cáo Đồ Án Khai Khoáng Dữ Liệu

Đề tài

Crawler Website Danh mục ngành và Chỉ Tiêu Tuyển Sinh 2020

Giáo viên hướng dẫn:
TS. LƯU TIẾN ĐẠO

Sinh viên thực hiện:
NGUYỄN VĂN LINH
Mã số: B1609778
LÊ NGUYỄN ĐỨC DUY
Mã số: B1611126
Khóa: 42

Cần Thơ, 06/2020

NHẬN XÉT CỦA GIẢNG VIÊN

LỜI CẢM ƠN

Để có được đồ án này, em xin được bày tỏ lòng biết ơn chân thành và sâu sắc đến Thầy Lưu Tiến Đạo – người đã trực tiếp tận tình hướng dẫn, giúp đỡ em. Trong suốt quá trình thực hiện đồ án, nhờ những sự chỉ bảo và hướng dẫn quý giá đó mà bài niên luận này được hoàn thành một cách tốt nhất.

Cần Thơ, ngày 1 tháng 6 năm 2020

Người viết

Nguyễn Văn Linh & Lê Nguyễn Đức Duy

MỤC LỤC

TÓM TẮT

Khi chúng ta xem một website nào đó, mà ta cần lấy nội dung của trang web thì chỉ cần sao chép dữ liệu đó. Nếu như website ấy thây đổi nội dung sau một khoản thời gian và chúng lại lấy những dữ liệu liên tục vậy chúng ta phải vào trang web ấy lấy dữ liệu theo một chu kì thời gian sao?

Câu trả lời là không. Tai sao chúng ta phải mất nhiều thơi gian cho việc làm đó.

Trong bài báo cáo này tôi sẽ giới thiệu cho các bạn một phương pháp đơn giản mà ai cũng có thể làm được và áp dụng cho nhiều website khác nhau đó là crawler (một thuật ngữ để chỉ việc lấy dữ liệu từ một trang web).

Để thực hiện lời nói của mình rằng ai cũng có thể crawler dữ liệu thì tôi sẽ sử dụng thư viện Beautiful Soup của ngôn ngữ lập trình Python để crawler dữ liệu từ trang web danh mục ngành và chỉ tiêu tuyển sinh đại học 2020 của đai học Cần Thơ.

PHẦN GIỚI THIỆU

1. Đặt vấn đề

Dữ liệu được xem là một phần không thể thiếu của mỗi trang web bất kỳ khi thiết lập. Tuy nhiên, để giải quyết vấn đề người dùng thì ít mà kho dữ liệu thì nhiều khá khó khăn. Vậy tầm quan trọng của web crawler là gì đối với những trang web mới?

Dữ liệu đã và đang trở thành một phần chính trong chiến lược tăng trưởng của mọi công ty. Các dữ liệu này bao gồm:

- Dữ liệu công khai.
- Dữ liệu thông tin người dùng.
- Dữ liệu của đối thủ cạnh tranh.
- Dữ liệu thị trường chứng khoán.
- Dữ liệu sản phẩm.
- Dữ liệu công việc.
- Dữ liệu dự báo thời tiết v…v

2. Lịch sử giải quyết vấn đề

Scraping dữ liệu không nhất thiết phải liên quan đến web. Scraping có thể đề cập đến việc trích xuất thông tin từ một hệ thống cục bộ, cơ sở dữ liệu chung hoặc thậm chí từ internet. Web Scaping cũng thực hiện việc tìm kiếm và thu thập thông tin nhưng khác với Web Crawling, Web Scraping không thu thập toàn bộ thông tin của một trang web mà chỉ thu thập những thông tin cần thiết, phù hợp với mục đích của người dùng. Trong WebScraping chúng ta cũng phần nào sử dụng WebCrawler để thu thập dữ liệu, kết hợp với Data Extraction (trích xuất dữ liệu) để tập trung vào các nội dung cần thiết.

Ví dụ như đối với trang amazon.com, Web Crawling sẽ thu thập toàn bộ nội dung của trang web này (tên các sản phẩm, thông tin chi tiết, bảng giá, hướng dẫn sử dụng, các reviews và comments về sản phẩm,…). Tuy nhiên Web Scaping có thể chỉ thu thập thông tin về giá của các sản phẩm để tiến hành so sánh giá này với các trang bán hàng online khác

Data scraping	Data Crawling
Involves extracting data from various sources including web	Refers to downloading pages from the web
Can be done at any scale	Mostly done at a large scale
Deduplication is not necessarily a part	Deduplication is an essential part
Needs crawl agent and parser	Needs only crawl agent

3. Mục tiêu đề tài

Lấy được thông tin về thông tin tuyển sinh năm 2020 , từ trang web danh mục ngành và chỉ tiêu tuyển sinh đại học chính quy năm 2020.

Sau khi lấy được dữ liệu , dữ liệu sẽ được lưu vào cơ sở dữ liệu Postgresql.

4. Phạm vi nghiên cứu

- Phạm vi nghiên cứu : Đại Học Cần thơ ,
- Đối tượng nghiên cứu : sinh viên

5. Phương pháp nghiên cứu

Nghiên cứu thư viện Beautiful Soup trong python và cấu trúc của trang web danh mục ngành.

Về thư viện Beautiful Soup trong python , nắm rõ cách thức sử dụng của thư viện này . Còn về trang web cần crawl phải hiểu rõ các thẻ html được định nghĩa trong trang web , để khi crawl dữ liệu không bị sai lệch.

6. Bố cục

Phần giới thiệu

Giới thiệu tổng quát về đề tài .

Phần nội dung

Chương 1 : Mô tả bài toán .

Chương 2 : Thiết kế, cài đặt .

Chương 3 : Kiểm thử hệ thống và đánh giá .

Phần kết luận

Trình bày kết quả đạt được và hướng phát triển hệ thống .

PHẦN NỘI DUNG

CHƯƠNG 1

MÔ TẢ BÀI TOÁN

1. Mô tả chi tiết bài toán

- Sử dụng thư viện Beautiful Soup trong python để thu thập dữ liệu từ trang web danh mục ngành và chỉ tiêu tuyển sinh đại học chính quy năm 2020, của Đại học Cần Thơ .

- Hình ảnh về trang web cần thu thập dữ liệu:

DANH MỤC NGÀNH VÀ CHỈ TIÊU TUYỂN SINH ĐẠI HỌC CHÍNH QUY NĂM 2020

Mã trường: TCT; Tổng chỉ tiêu tuyển sinh: 8.900

1. Chương trình đào tạo đại trà

(*) ngành đào tạo giáo viên chỉ xét tuyển theo phương thức 1 và 2

Mã ngành	Tên Ngành - *chuyên ngành*	Mã tổ hợp xét tuyển (Phương thức 2 và 3)	Chỉ tiêu	Tham khảo điểm trúng tuyển		
				2019	2018	2017
Nhóm ngành Công nghệ						
7510401	Công nghệ kỹ thuật hóa học	A00, A01, B00, D07	170	15,00	17,25	21,25
7520114	Kỹ thuật cơ điện tử	A00, A01	100	16,25	17,00	20,50
7520103	Kỹ thuật cơ khí, *có 2 chuyên ngành:* - *Cơ khí chế tạo máy* - *Cơ khí ôtô*	A00, A01	240	18,75	17,50	20,50
7520201	Kỹ thuật điện	A00, A01, D07	140	16,00	16,50	20,50
7520207	Kỹ thuật điện tử - viễn thông	A00, A01	100	15,00	15,00	18,25
7520216	Kỹ thuật điều khiển và tự động hóa	A00, A01	100	16,00	16,50	19,00
7480106	Kỹ thuật máy tính	A00, A01	100	15,00	15,25	16,50
7580201	Kỹ thuật xây dựng	A00, A01	180	16,00	16,00	19,25
7520309	Kỹ thuật vật liệu	A00, A01, B00, D07	60	14,00	14,00	14,00
7580205	Kỹ thuật xây dựng công trình giao thông	A00, A01	60	14,00	14,00	18,00
7580202	Kỹ thuật xây dựng công trình thủy	A00, A01	60	14,00	14,00	15,50

2. Vấn đề và giải pháp liên quan đến bài toán

Từ trang web ta có thể thấy trang web này được viết theo cách thuần html, tức là web tĩnh . Để hiểu rõ hơn về trang web mà chúng ta cần thu thập dữ liệu , tôi sẽ nói rõ hơn về như thế nào là web tĩnh , web động

Website tĩnh là gì ?

- Là trang web sử dụng hoàn toàn ngôn ngữ HTML, sau khi tải trang HTML từ máy chủ xuống, trình duyệt sẽ biên dịch mã và hiển thị nội dung trang web, người dùng hầu như không thể tương tác với trang web.

- Nội dung website ít khi cập nhật và ít nên bạn muốn tiết kiệm chi phí.

- Website bạn nhỏ và bạn thuê luôn người chuyên về web để quản trị Nếu bạn là doanh nghiệp muốn tự mình làm website thì bạn có thể học các kiến thức căn bản và tự làm một Web tĩnh cho mình.

Ưu điểm của website tĩnh:

- Tốc độ truy cập nhanh bởi nó chỉ là những file HTML.

- Chi phí đầu tư thấp bởi bạn không phải trả tiền nhiều cho Coder.

- Về giao diện Designer có thẻ thiết kế theo kiểu mới lạ.

- Thân thiện với bộ máy tìm kiếm bởi bạn có thể đặt tên file tùy ý

Nhược điểm của website tĩnh:

- Khó quản lý nội dung.

- Khó nâng cấp bảo trì...

Website động là gì ?

Là một tập hợp các dữ liệu số hóa được tổ chức thành cơ sở dữ liệu, các dữ liệu số hóa được gọi ra trình diễn trên các trang web dưới dạng văn bản, âm thanh, hình ảnh.nó có thêm các phần xử lý thông tin và truy xuất dữ liệu còn website tĩnh thì không.

Khác với web tĩnh, web động luôn luôn có thông tin mới do các thông tin này được cập nhật bởi phần mềm quản trị web do các công ty thiết kế website cung cấp. Các thông tin mới này được lưu vào cơ sở dữ liệu của website và đưa ra sử dụng dựa theo yêu cầu của người dùng.

Trang web động được các chuyên gia lập trình, sử dụng các ngôn ngữ lập trình tạo ra mã nguồn dựa theo yêu cầu của trang web.

Bạn làm web tin tức, blog cá nhân.

Web bạn tầm cỡ lớn.

Bạn làm website thương mại điện tử bán hàng.

Bạn làm web giới thiệu sản phẩm công ty

Ưu điểm của website động:

- Dễ dàng nâng cấp và bảo trì.

- Có thể xây dựng được web lớn.

- Thường sử dụng tương tác với người dùng cao.

- Dễ dàng quản lý nội dung

Nhược điểm của website động:

- Nếu web lớn có thể cần thêm nhân sự chuyên ngành.
- Chi phí xây dựng cao

Như vậy là là ta có thể hiểu được sự khác biệt giữa web tĩnh và động khác nhau như thế nào và trang web mà ta thu thập dữ liệu sẽ là trang web tĩnh.

Và như chúng ta đã biết một trang web tĩnh thì cấu trúc của các thẻ trong trang sẽ rất khác nhau và không đồng nhất , việc lấy dữ liệu từ một trang web như vậy sẽ rất khó khăn trong việc lấy dữ liệu.

CHƯƠNG 2
THIẾT KẾ VÀ CÀI ĐẶT

1. Thiết kế hệ thống

- Mô hình hệ thống thu thập dữ liệu :

2. Xây dựng hệ thống

a) Thực hiện

- Các bước thực hiện

+ Bước 1 :

Xây dựng hệ thống kết nối với heroku thông qua Flask :

```
from bs4 import BeautifulSoup
from flask import Flask
import  requests as rq
import psycopg2


app = Flask(__name__, template_folder='template')
app.secret_key = 'replace later'

@app.route('/')
def index():
    print(crawler_chitiet())
    print(crawler_tennganh())
    connect()
    print(demo())

    return 'successfully'
```

+ Bước 2 :

Sau khi đã kết nối tới heroku , thì chúng ta bắt đầu thu thập dữ liệu bằng beautifulsoup trong python

```
def crawler_tennganh ():
    html = rq.get
    ("https://tuyensinh.ctu.edu.vn/chuong-trinh-dai-tra/
    841-danh-muc-nganh-va-chi-tieu-tuyen-sinh-dhcq.html
    /gioi-thieu-nganh/551-cong-nghe-ky-thuat-hoa-hoc").text
    soup = BeautifulSoup(html, 'html5lib')
    new_feed = soup.find('section', class_='article-content clearfix').find_all('p')


    # Công nghệ kỹ thuật hóa học
    TDGT = new_feed[6].text
    GT = new_feed[7].text
    GT1 = new_feed[8].text
    GT2 = new_feed[9].text
```

+ bước 3 : Sau khi đã thu thập được dữ liệu , tiến hành lưu dữ liệu : bằng cách sử dụng thư viện psycopg2 và lấy thông tin cần thiết.

```
user="xdjbqgulyhzhck",
password="d8aa4a5f3335d5489970c10428d5adbcd7c7ad5d215ec54b8d0fd8afdaf9b763",
host="ec2-52-72-65-76.compute-1.amazonaws.com",
port="5432",
database="db3kr1a2ptuutr")
```

- Và cuối cùng ta lưu dữ liệu thu thập vào cơ sở dữ liệu PostgreSQL của heroku

```
bảng danh mục ngành

[('7510401', 'Công nghệ kỹ thuật hóa học', 'A00, B00, A01, D07', '170', '15,00', '17,25', '21,25'),
 ('7520114', 'Kỹ thuật cơ điện tử', 'A00, A01', '100', '16,25', '17,00','20,50')]
```

CHƯƠNG 3
KIỂM THỬ VÀ ĐÁNH GIÁ

- Ta thấy rằng dữ liệu khi lưu vào không được hoàn chỉnh và đồng nhất do như đã nói trên , trang web mà chúng ta thu thập là trang web tĩnh vì vậy việc các thẻ được thiết kế không đồng nhất . Dẫn tới việc thu thập dữ liệu rất khó khăn và bị sai sót.

PHẦN KẾT LUẬN

- Hiện nay với sự phát triển của công nghệ thông tin nói chung , ta có thể dễ dàng tìm thấy những giải pháp thu thập dữ liệu mới .

- Điển hình như có một phần mềm mã nguồn mở với tên gọi parsehub , đây là ứng dụng thu thập dữ liệu sử dụng giao diện đồ họa rất tiện ít và giúp ngắn thời gian thu thập dữ liệu lại rất nhiều.

- Đây là trang chủ của parsehub các bạn có thể tham khảo thêm cách này : https://www.parsehub.com/docs/ref/api/v2/#introduction

- Như ta thấy thì với ứng dụng này việc thu thập dữ liệu sẽ nhanh chóng hơn rất nhiều so với thu thập truyền thống và đặc biệt hơn với sự hỗ trợ của API của ứng dụng này . Các bạn sẽ không cần làm việc với cơ sở dữ liệu .

TÀI LIỆU THAM KHẢO

- https://elcit.ctu.edu.vn/course/view.php?id=3752
- https://vi.wikipedia.org/wiki/Wikipedia
- https://devcenter.heroku.com/articles/git