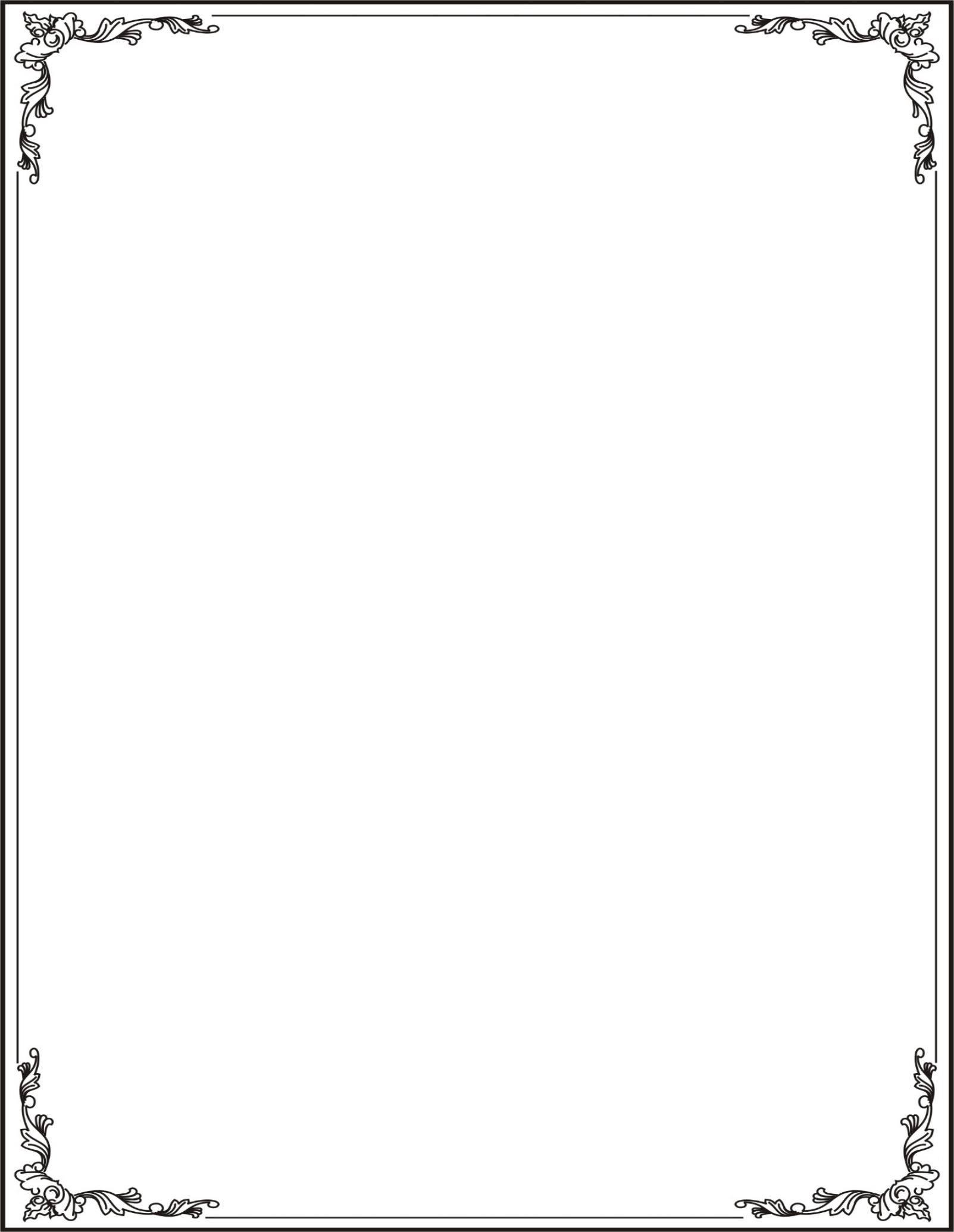
**KHOA TOÁN – CƠ – TIN HỌC**

-----🙞🙜🕮🙞🙜-----



**BÁO CÁO ĐỀ TÀI:**

**Xây dựng web crawler bằng**

**ngôn ngữ python trên máy VPS**

***Môn: Đồ Án Phần Mềm***

**Người thực hiện:** Lê Văn Anh Tú

**Người hướng dẫn**: Đinh Nho Hào

Hà Nội, tháng 8 năm 2018

# Giới thiệu

## Nội dung thực tập:

## Python 3.x

* Là ngôn ngữ lập trình kịch bản (script language**),** thông dịch (interpreter language), hướng đối tượng (object-oriented) được ra mắt lần đầu vào năm 2008. Được xây dựng, bổ sung trên phiên bản Python 2.x
* Python được sử dụng vào các mục đích:
  + Phát triển ứng dụng web (phía server)
  + Phát triển phẩn mềm
  + Xử lý dữ liệu lớn (big data) và thực hiện các biểu thức toán học phức tạp
* Đặc điểm:
  + Cú pháp ngắn gọn, dễ hiểu, cho phép các lập trình viên viết chương trình với số dòng lệnh ít hơn so với viết chương trình bằng các ngôn ngữ khác
  + Chạy trên nhiều nền tảng (Window, Linux, Mac, Raspberry Pi, …)
  + Thư viện dựng sẵn lớn

## PHP

* Là ngôn ngữ lập trình kịch bản mã nguồn mở (open source), hướng đối tượng. Ra mắt lần đầu vào năm 1995
* PHP được sử dụng vào các mục đích:
  + Sử dụng trong hệ thống server
  + Sử dụng để truy cập, thay đổi, tạo mới cookie
  + Mã hóa dữ liệu
* Đặc điểm:
  + Đơn giản, Hiệu quả, An toàn, linh động
  + Chạy trên nhiều nền tảng (Window, Linux,…)
  + Tương thích với nhiều server được sử dụng phổ biến (Apache, IIS,…)
  + Hỗ trợ nhiều

## C#

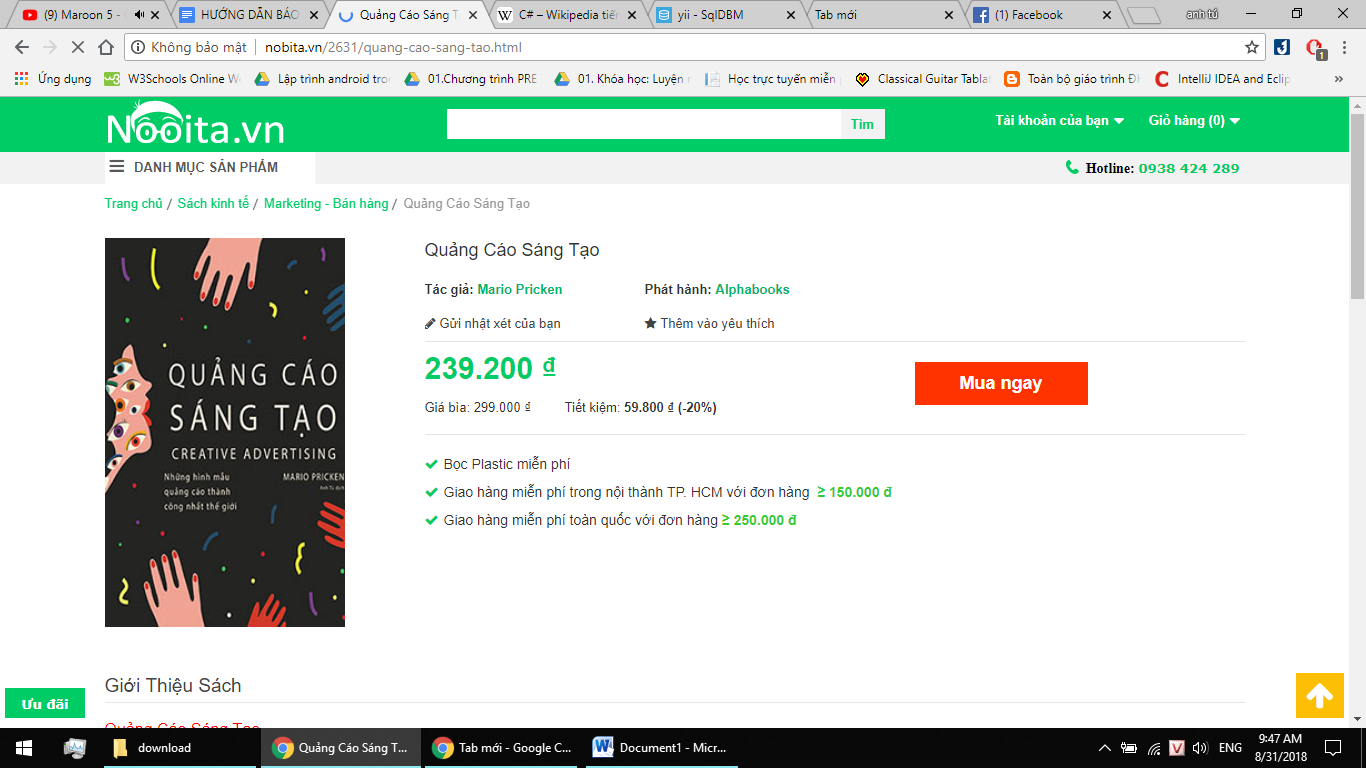
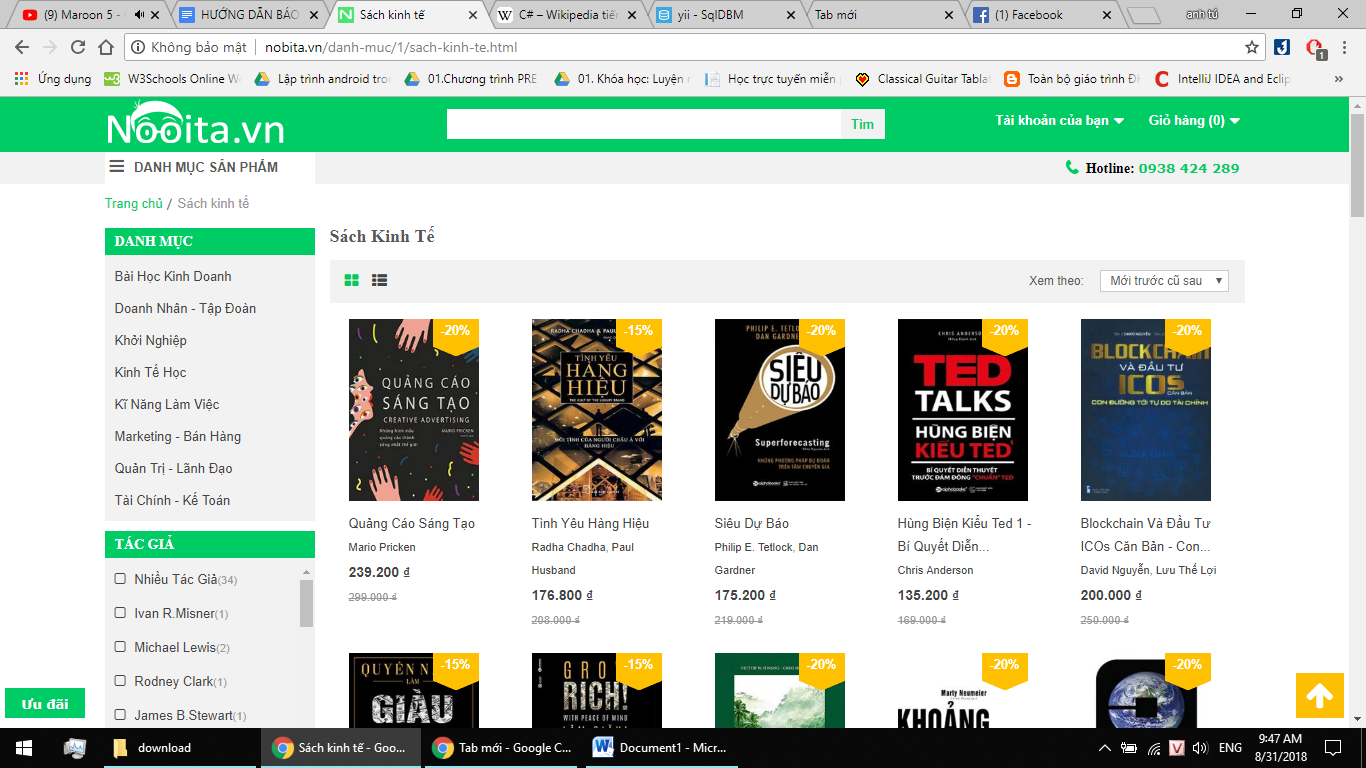
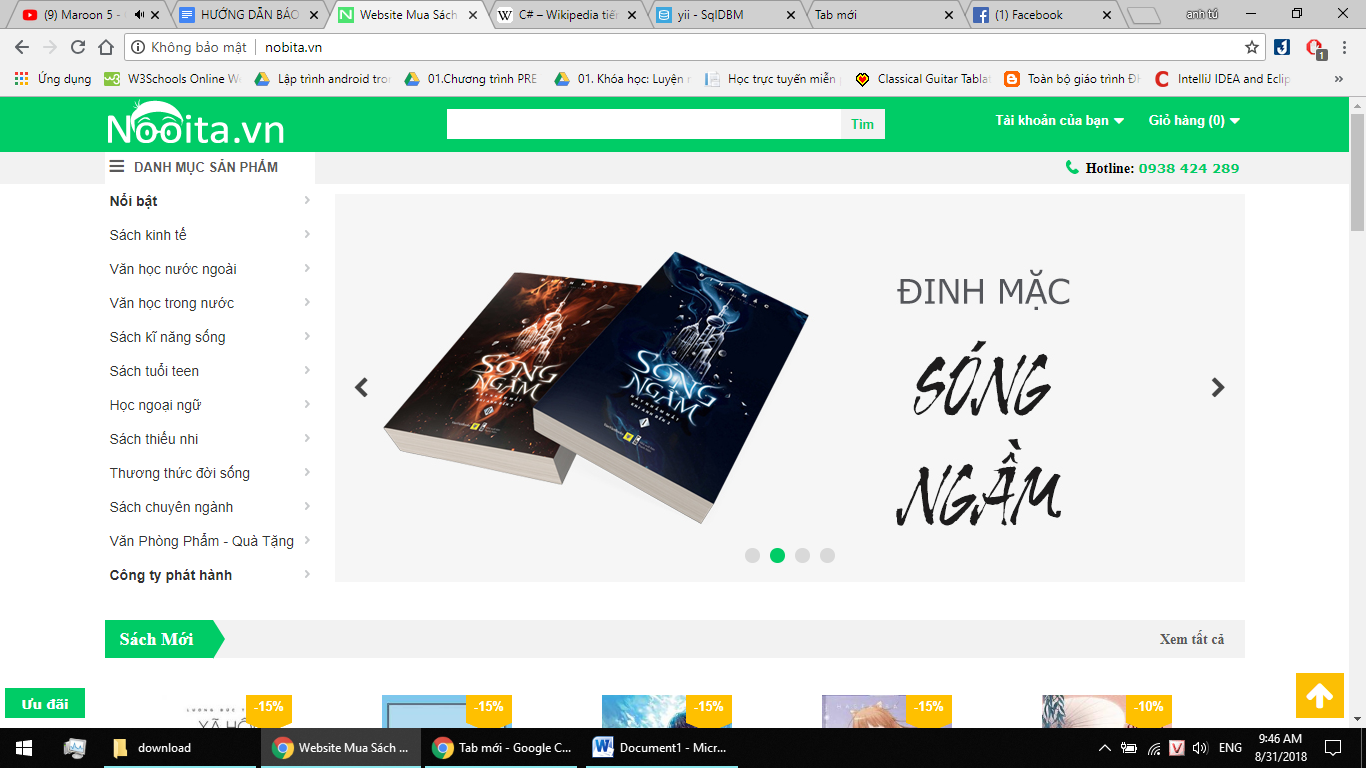
* Là ngôn ngữ lập trình hướng đối tượng, sử dụng vào nhiều mục đích (general-purpose). Được phát triển bới Microsoft và ra mắt lần đầu vào năm 2000
* Đặc điểm: .
  + Tạo được những phần mềm hiệu quả
  + Là ngôn ngữ có cấu trúc

# Python Web Crawler

## Tổng quan:

Mục đích:

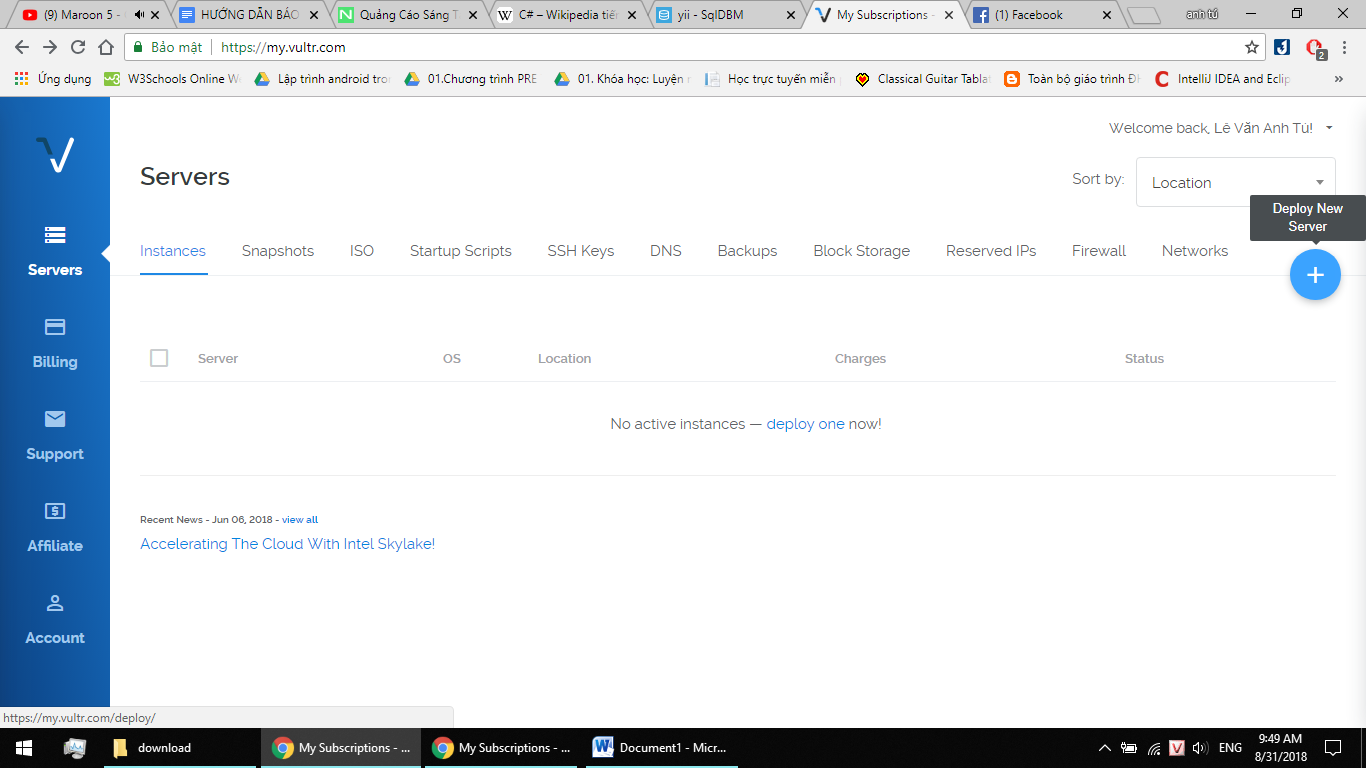
* Lấy dữ liệu được hiển thị trên trang web và sử dụng lại dữ liệu đó để tạo trang web riêng sau này.
* Nguồn dữ liệu được lấy : Nobita.vn

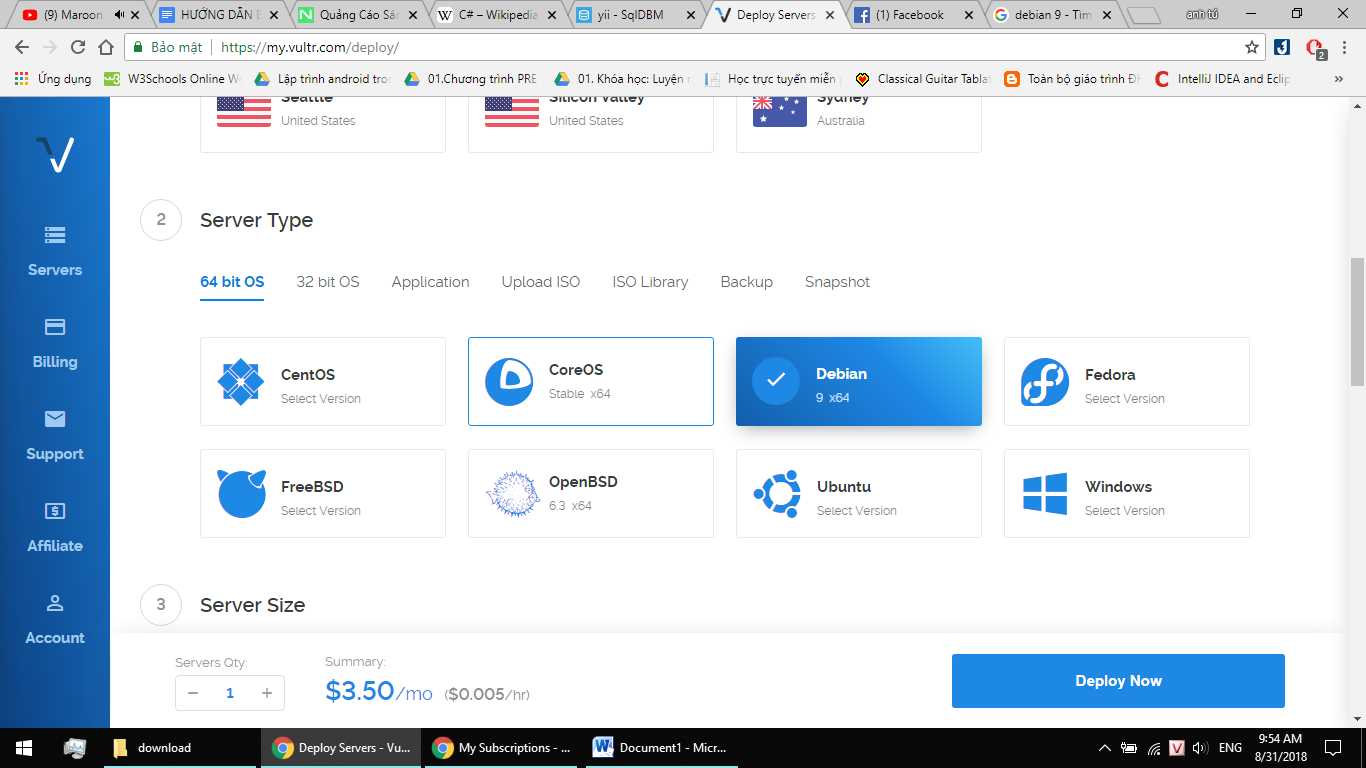


## Những công cụ cần sử dụng

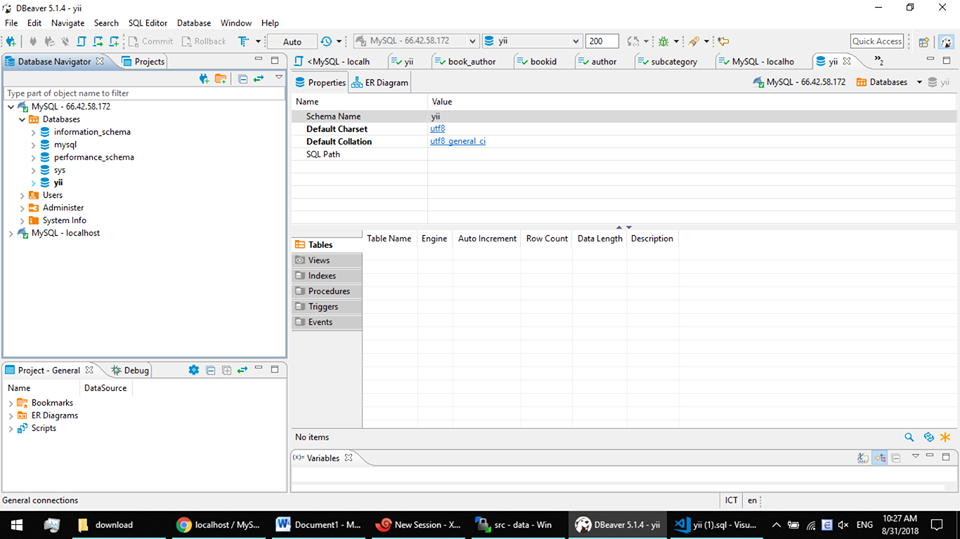
* Python 3
* MySQL
* VPS : vultr

## Bước chuẩn bị trước:

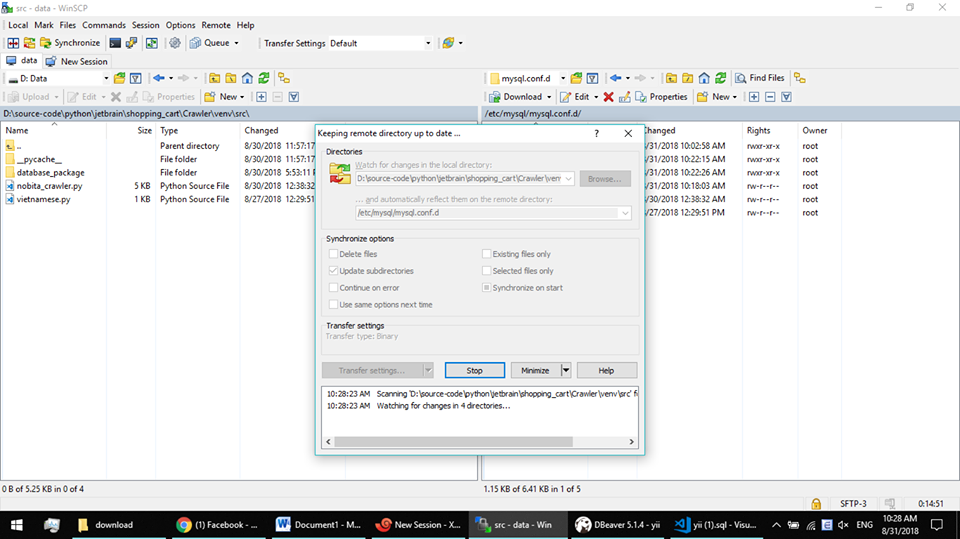
* Tạo máy VPS: Hệ điều hành sử dụng Debian9-Stretch
* 



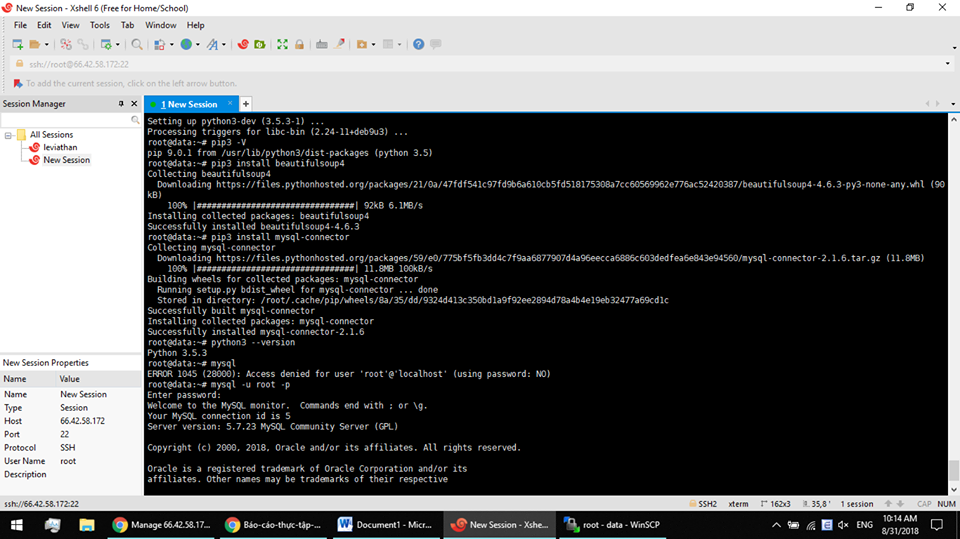
* Sử dụng XShell6, WinSCP để kết nối đến VPS, cài đặt các chương trình cần thiết như python và mysql. Sử dụng DBeaver để kết nối đến cơ sở dữ liệu trong VPS



DBeaver

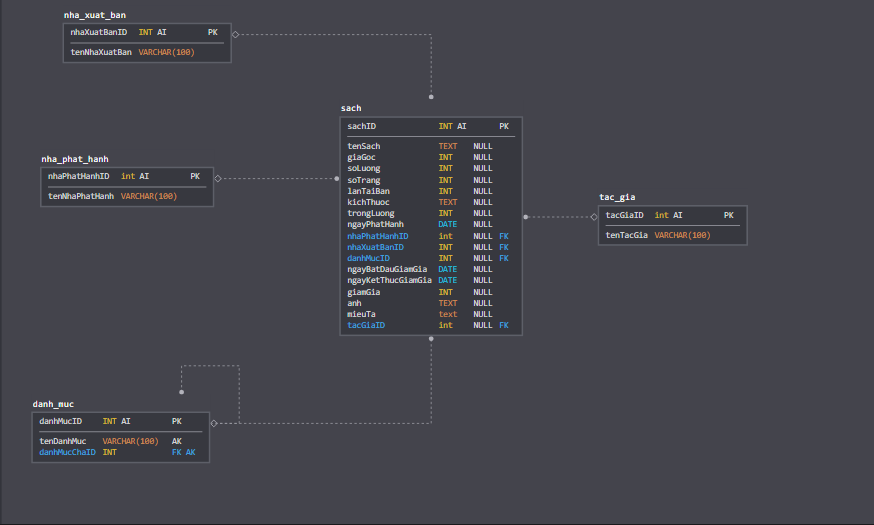


WinSCP



XShell6

## Cấu trúc cơ sở dữ liệu



Cấu trúc cơ sở dữ liệu

Các dữ liệu cần được lấy về:

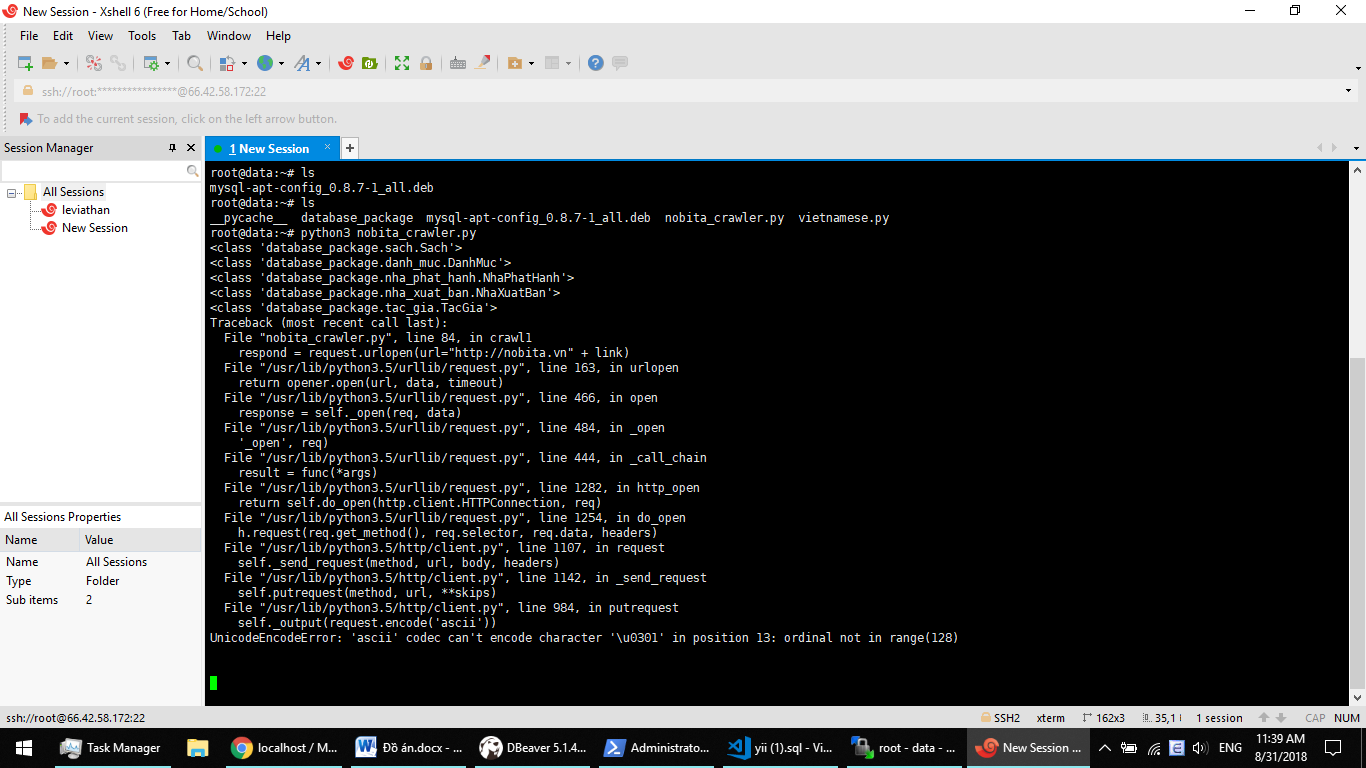
* Thông tin về sách: Tên sách, tác giả, giá
* Thông tin về nhà phát hành, nhà xuất bản, tác giả

## Thư viện cần được cài đặt thêm:

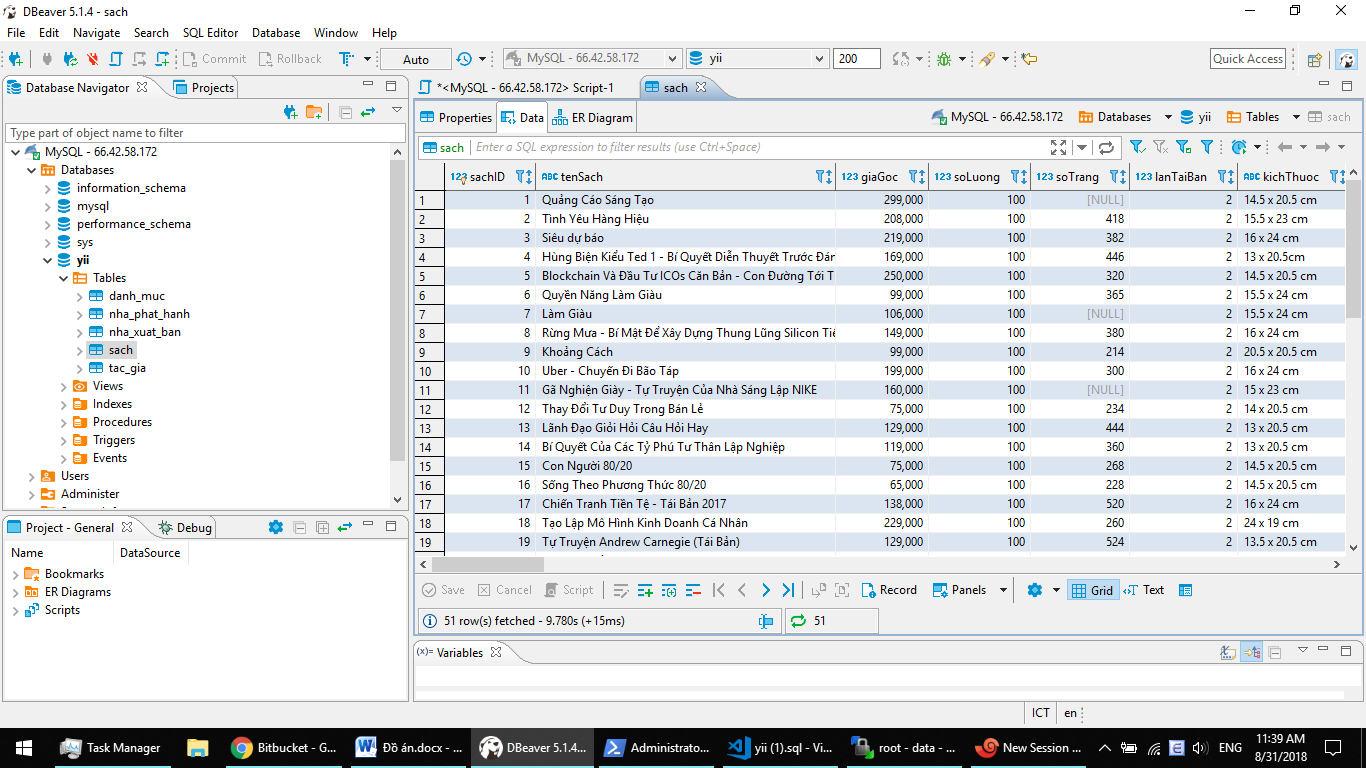
* beautifulsoup4: pip3 install beautifulsoup4
* mysq.connector: pip3 install mysql-connector

## Chạy chương trình

Câu lệnh: python3 nobita\_crawler.py



Chương trình đang chạy



Kết quả

# Tài liệu tham khảo

1. C#

<https://www.tutorialspoint.com/csharp/index.htm>

1. Python

<https://www.tutorialspoint.com/python3/index.htm>

<https://www.w3schools.com/python/default.asp>

1. PHP

<https://www.w3schools.com/php/default.asp>

<https://www.tutorialspoint.com/php/index.htm>

1. VPS Vultr

<https://my.vultr.com/>

1. Source code

<https://app.sqldbm.com/MySQL/Share/hi-qKJl4jluiN3hV-TP270GFrngIE8md_DYjF4jNYw0>

<https://github.com/levananhtu/WebCrawler>

# Nhận xét của công ty

* Công ty House3D.
* Người hướng dẫn: Đinh Nho Hào
* Nhận xét về quá trình thực tập:
  + Xác nhận về công việc được giao và kết quả báo cáo:

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

* + Ý thức và thái độ làm việc:

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

………………………………………………………………………………………………………………

* Điểm số đánh giá:
  + Bằng số:…………………………………… Bằng chữ:……………………………………

Hà Nội, ngày….tháng….năm….

Đại diện công ty