ĐẠI HỌC KHOA HỌC TỰ NHIÊN
TRUNG TÂM TIN HỌC

Địa chỉ: 227 Nguyễn Văn Cừ, Q.5, TP.HCM
Tel: (028) 38351056
Email: tuvan@csc.hcmus.edu.vn
Website: https://csc.edu.vn/

Database SQL and Data Collection for Data Science

(Truy vấn và thu thập dữ liệu cho Khoa học dữ liệu)

Thời gian: 5 tuần

Thời lượng: 40 giờ

Học phí: 5.200.000 đ

1. MỤC TIÊU

- Hiểu được tầm quan trọng của Database và truy vấn SQL trong Khoa học dữ liệu
- Hiểu các nguyên tắc cơ bản của cơ sở dữ liệu quan hệ và SQL
- Viết các truy vấn SQL phức tạp để truy xuất và thao tác dữ liệu
- Sử dụng FugueSQL để tương tác với dữ liệu và thực hiện các thao tác dữ liệu
- Thu thập dữ liệu từ các trang web bằng BeautifulSoup, Selenium
- Tích hợp dữ liệu đã thu thập vào quy trình phân tích dữ liệu Python
- Xử lý các lỗi thường gặp trong việc thu thập dữ liệu.
- Hoàn thành dự án thu thập dữ liệu bằng cách sử dụng các công cụ và kỹ thuật đã học
- Là khóa học thứ ba trong chương trình “Data Science and Machine Learning Certificate”

2. ĐỐI TƯỢNG HỌC

- Đã học qua “Fundamentals of Python” và “Data Manipulation and Visualization with Python”

3. KẾT QUẢ ĐẠT ĐƯỢC

Sau khi hoàn thành khóa học, học viên sẽ đạt được các kỹ năng:

- Hiểu và vận dụng ngôn ngữ truy vấn SQL trong việc truy xuất dữ liệu phục vụ cho Khoa học dữ liệu
- Hiểu và vận dụng các kỹ thuật trích xuất dữ liệu để thu thập dữ liệu từ các trang web

4. NỘI DUNG KHOÁ HỌC

- **Giới thiệu về Database và SQL**
  - Giới thiệu về cơ sở dữ liệu
  - Các loại cơ sở dữ liệu (Quan hệ, NoSQL)
  - Tổng quan về SQL và tầm quan trọng của nó trong Khoa học dữ liệu
  - Cài đặt và Thiết lập cơ sở dữ liệu (MySQL + WorkBench, Big Query console).
- **Truy vấn SQL cơ bản**
  - Truy vấn dữ liệu với SELECT, FROM
  - Lọc dữ liệu với WHERE
  - Giới hạn dữ liệu với LIMIT
  - Sắp xếp dữ liệu với ORDER BY
  - Gom nhóm và Lọc dữ liệu nhóm với GROUP BY, HAVING
  - Sử dụng các hàm trong SQL (string, number, datetime, aggregate)
- **Truy vấn SQL nâng cao**
  - Truy vấn con
  - Truy vấn trên nhiều bảng (cross join)
  - Kết hợp truy vấn với UNION
  - Sử dụng truy vấn con tạo bảng dẫn xuất
  - Common Table Expressions (CTEs)
  - Window Functions
- **Truy cập Database với Python**
  - Giới thiệu tổng quan, Ưu điểm và mô hình truy cập database từ Python
  - Kết nối database thông qua DB-API (Local Database, Cloud Database)
  - Tạo Connection và Cursor để kết nối và thực hiện truy vấn với Database (MySQL)
  - Magic SQL
  - Tổ chức lưu trữ dữ liệu từ truy vấn vào DataFrame
- **Làm việc với FugueSQL**
  - Giới thiệu FugueSQL
  - Cài đặt và cấu hình FugueSQL.
  - Đọc/Ghi dữ liệu từ tập tin

- Truy vấn dữ liệu với FugueSQL.
- Aggregate Functions và Window Functions trong FugueSQL.
- Hàm tự định nghĩa trong FugueSQL

- **Thu thập dữ liệu với BeautifulSoup**
  - Tổng quan về Web Scraping
  - Cấu trúc cơ bản của trang HTML.
  - Cài đặt BeautifulSoup
  - Trích xuất dữ liệu từ trang web với BeautifulSoup.
  - Các xử lý trích xuất dữ liệu nâng cao trong BeautifulSoup: các trang HTML sử dụng nội dung động với Javascript và Ajax, trang web có phân trang
  - Xử lý dữ liệu trích xuất và lưu vào file (CSV, Excel, JSON)
- **Thu thập dữ liệu với Selenium**
  - Tổng quan về Selenium.
  - Cài đặt và thiết lập môi trường Selenium.
  - Điều hướng các trang web với Selenium
  - Tương tác với các thành phần web (click, input).
  - Xử lý popup, alert và iframe, chụp ảnh màn hình
  - Xử lý Cookie và Sessions.
  - Trích xuất dữ liệu từ các trang web động với Selenium: Javascript và Ajax, trang web có phân trang.