TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN

KHOA CÔNG NGHỆ THÔNG TIN

--------- oOo ---------

BÀI TẬP THỰC HÀNH

LẬP TRÌNH PYTHON NÂNG CAO

TRÌNH ĐỘ: ĐẠI HỌC CHÍNH QUY

NGÀNH ĐÀO TẠO: KHOA HỌC MÁY TÍNH

Hưng Yên – Tháng 11 năm 2021

BÀI THỰC HÀNH SỐ 3: THAO TÁC VỚI TỆP DỮ LIỆU

A. MỤC TIÊU BÀI THỰC HÀNH

Sau bài thực hành này sinh viên có thể:

- Hiểu được cách thao tác với các loại tệp tin CSV, JSON, XML bằng Pandas
- Sử dụng thành thạo các hàm thao tác với dữ liệu mảng, DataFrame trong Pandas
- Vận dụng các kỹ thuật xử lý dữ liệu với Pandas để phân tích dữ liệu trong các bài toán thực tế

B. ĐIỀU KIỆN THỰC HÀNH

Với đặc thù của môn Lập trình Python nâng cao, mục này sẽ liệt kê một số công cụ sử dụng để làm bài thực hành. Trong bài thực hành này, sinh viên cần kiểm tra và chắc chắn các phần mềm sau trên máy tính còn hoạt động tốt:

1. Anacoda phiên bản 3.0 trở lên
2. Jupyter Notebook

C. TÀI NGUYÊN THAM CHIẾU

Để hoàn thành tốt bài thực hành này sinh viên nên tham khảo các tài nguyên sau:

STT	Tên tài nguyên	Mô tả tài nguyên
1	Practice 03.pdf	Tài liệu hướng dẫn thực hành bài số 1
2	Lesson 04 – Working with CSV, JSON, XML file formats.pdf	Slide bài giảng về thao tác với các tệp dữ liệu CSV, JSON, XML sử dụng Pandas

D. YÊU CẦU BÀI THỰC HÀNH

Bài 1. Dữ liệu 288 bệnh nhân nhiễm Covid-19 ở Việt Nam Pandas (05-2020) được lưu trong tệp Vietnam_Covid19_Patients.csv. Bạn hãy lập trình thực hiện các yêu cầu sau:

a) Đọc toàn bộ dữ liệu trong tệp tin ra DataFrame và hiển thị ra màn hình

b) Viết các lệnh truy xuất lấy về các dữ liệu sau:

- Bệnh nhân trở về từ "**Wuhan(China)**"

- Bệnh nhân ngoài cộng đồng không rõ nguồn lây nhiễm
- Bệnh nhân nhiễm Covid trong khoảng từ ngày "**01-03-2020**" đến ngày "**14-03-2020**"
- Các bệnh nhân có quốc tịch "**USA**" nhiễm Covid ở Việt Nam vào Tháng 3 năm 2020

c) Viết các lệnh thống kê:

- Số bệnh nhân Nam, Nữ
- Số bệnh nhân theo các quốc tịch
- Số bệnh nhân theo các tỉnh thành phát hiện
- Số bệnh nhân được điều trị tại các bệnh viện

d) Kết xuất ra tệp CSV:

- Danh sách các bệnh nhận được điều trị tại **"Bệnh viện Chợ Rẫy"**
- Danh sách các bệnh nhân có người nước ngoài bị nhiễm tại Việt Nam trong Tháng 3 năm 2020

Bài 2. Dữ liệu 2490 sản phẩm của Lazada Vietnam được lưu trong tệp LazadaProducts.json.

Bạn hãy lập trình thực hiện các yêu cầu sau:

a) Đọc toàn bộ dữ liệu trong tệp tin ra DataFrame và hiển thị ra màn hình

b) Viết các lệnh truy xuất lấy về các dữ liệu sau:

- Các sản phẩm thuộc loại "Máy vi tính & Laptop"
- Các sản phẩm có điểm đánh giá trung bình > 4.0

c) Viết các lệnh thống kê:

- Số lượng sản phẩm theo mỗi loại
- Top 10 sản phẩm có có điểm đánh giá trung bình cao nhất
- Top 100 sản phẩm có đánh giá nhưng đạt điểm thấp nhất

d) Chuẩn hóa và kết xuất dữ liệu ra tệp JSON

- Chuyển giá trị trong các trường giá (p_price) và số lượt đánh giá (p_number_reviews) về dạng số
- Thay giá trị các **null** của các trường đánh giá (p_rate1star, … p_rate5star, p_rating) bằng giá trị "0"
- Ghi dữ liệu sau khi đã chuẩn hóa ở trên ra tệp LazadaProducts_Cleaned.json

Bài 3: Dữ liệu 8807 video clips của NetFlix được lưu trữ trong tệp "netflix_data.xml".

Bạn hãy lập trình thực hiện các yêu cầu sau:

a) Đọc toàn bộ dữ liệu trong tệp tin ra DataFrame và hiển thị ra màn hình

b) Sử Xpath lấy dữ liệu sau:

- Các bộ phim (Movie) được phát hành vào năm "2020"
- Các chương trình "TV Show" sản xuất tại "United States" được phát hành năm "2021"

c) Viết các lệnh thống kê:

- Số bộ phim được sản xuất tại "United States" vào năm 2009
- Số video clips được sản xuất theo các nước
- Các chương trình "TV Show" có từ 2 mùa trở lên

d) Lọc và kết xuất dữ liệu ra tệp XML

- Lấy về các bộ phim có thời gian <= 120 phút và lưu ra tệp ShortVideo.XML

E. HƯỚNG DẪN THỰC HIỆN

Sinh viên tạo tệp **Practice03_HoVaTen.ipynb** trên Jupyter Notebook và thực hiện viết mã lệnh để giải quyết các bài tập thực hành.

Sinh viên giải các bài tập theo ví dụ mẫu trên lớp.