**Phân tích Dữ liệu: Phân tích điểm của học sinh 2020**

Phần 1: Thu thập dữ liệu - Data collection

- Ở bài này, mình dùng lệnh curl và python để lấy dữ liệu từ web điểm thi (diemthi.hcm.edu.vn) về máy. Ở cuối bài này, mình sẽ thu được file raw\_data.txt chứa dữ liệu (rất rối, chưa làm sạch) của toàn bộ học sinh. Trang web của bộ giáo dục sớm muộn gì cũng thay đổi, khi nó thay đổi, bạn xem video mình tham khảo, sau đó lấy file raw\_data.txt về máy rồi thực hiện các bài học sau (làm sạch dữ liệu và phân tích dữ liệu)

**- Bôi đen code => Nhấn ctrl+/ để comment đoạn code trong python**

**Bài 1.1: Cách dùng lệnh curl lấy dữ liệu web**

Lưu ý: Trang diemthi.hcm.edu.vn nay đã thành điểm của năm 2021 nên có chút thay đổi, hồi mình làm là năm 2020, cách lấy điểm vẫn giống hệt các bạn nhé, các bạn muốn dùng data của mình có thể tải file raw\_data.csv về để làm tiếp

Clip này mình hướng dẫn qua cách dùng lệch curl trên terminal để lấy giữ liệu web.

curl + tên trang web: Thực hiện lệnh GET request để lấy code html của trang web về

curl -F "tên form data + giá trị" + endpoint: Thực hiện lệch POST request để thay người dùng điền form và lưu giá trị trả về.

curl là lệch có sẵn trong mọi máy

Nếu trang web diemthi.hcm.edu.vn không còn hoạt động, đừng lo, bạn cứ xem tiếp tham khảo, sau đó cuối unit này mình sẽ up file raw\_data.txt (file chứa dữ liệu chưa làm sạch của toàn bộ học sinh)

Bài 1.1 Đây là file raw\_data.txt - dữ liệu web lấy trực tiếp từ website bộ giáo dục, chứa 74718 dòng: Để trong thư mục PTDL\_Python/raw.txt

**Bài 1.2: Dùng python để chạy lệnh trên terminal**

Dùng python, thư viện subprocess để thực hiện lệnh curl và lưu kết quả vào 1 variables (biến).

Nếu có lỗi lúc dùng subprocess gọi curl, bạn thêm shell=True như này:

subprocess.run('curl -F "SoBaoDanh=02004318" [diemthi.hcm.edu.vn/Home/Show](http://diemthi.hcm.edu.vn/Home/Show?fbclid=IwAR1KJLrnBjhPcz3JKc0aO2cn5SwGI9fNKUPVuItRzuySVmQWuxH1S6Usk8c)

', shell=True)

Bài 1.3:

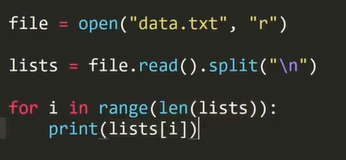
Mình tạo đoạn code này để tổng hợp kiến thức đọc ghi file. Code này sẽ tạo file text (data.txt), viết vào file đó và rồi đọc file đó rồi print ra terminal.



Mode: w => Tạo file mới, nội dung mới sẽ chèn vào nội dung cũ

Mode: a=> Tạo file mới (nếu chưa tồn tại), nếu file tồn tại thì ghi bổ sung nội dung vào file cũ

Mode : r => đọc nội dung file



- Đọc nội dung của 1 file text => Kết quả đọc ra 1 danh sách => hàm len tính độ dài của danh sách



Đối trong hàm write phải là kiểu string

Bài 1.4

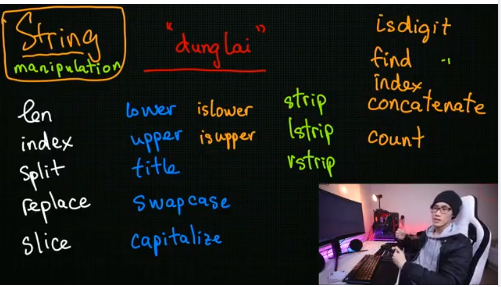
Bài này, mình dùng vòng lặp để dùng code lấy hết data học sinh về, tạo ra file raw\_data.txt chứa toàn bộ dữ liệu chưa làm sạch

**Phần 2**

**Làm sạch dữ liệu - Data cleaning**

Sau khi đã có file raw\_data.txt rất lộn xộn ở bài trước, mình sẽ dùng các phương pháp xử lý string để dọn dẹp và đưa nó về file .csv (có thể mở bằng excel) chứa các cột, mỗi cột lần lượt là số báo danh, tên, ngày tháng năm sinh, điểm toán văn anh, vân vân. Cuối bài này, mình sẽ thu được file clean\_data.csv chứa 74444 thông tin của học sinh đã làm sạch.

Xử lý dữ liệu text



**Bài 2.3: Làm sạch, bỏ kí tự xuống dòng**

Bài này, mình tạo sườn code để xử lý học sinh đầu tiên, làm sạch 1 dòng rồi sẽ làm được tự động toàn bộ.

Đọc dòng đầu của file raw\_data.txt

Bỏ kí tự xuống dòng \n - newline

Lưu kết quả vào file test.txt để tiện debug sửa lỗi

**Bài 2.4 - Xóa tag - Phân tích**

Giải thích việc cần xóa tag để làm sạch

Phân tích logic để xóa tag

Vì chỉ cần nội dung của trang web nên ta xóa toàn bộ nội dung của trang web.

Ví dụ code xóa tag này sẽ đổi

<title>So giao duc</title>

thành

So giao duc

Code:

s = "<title>So giao duc</title>"

tags = []

for i in range(len(s)):

if s[i] == "<":

begin = i

if s[i] == ">":

end = i

tags.append(s[begin:end+1])

for tag in tags:

s = s.replace(tag, "")

print(s)

Bài 2.6

**Bài 2.6: Xóa tag - Áp dụng thực hành**

Dùng ví dụ bài 2.5 vào raw\_data.txt

**Bài 2.7: Xóa kí tự thừa, dòng trống**

**Bài 2.8: Xử lý kí tự đặc biệt - Chuẩn bị bảng mã**

có nhiều kí tự đặc biệt cần được đổi thành kí tự tiếng việt, mình lấy các bảng mã trên mạng tạo ra file unicode.txt, đọc file này để chuẩn bị làm sạch raw\_data.txt

**Bài 2.9: Đổi kí tự web - Phân tích**

Kí tự như &#225 cần được đổi thành chữ á

**Code bài 2.9: Làm sạch mã unicode**

Tiếp nối code bài 2.8 sau khi đã có 2 danh sách (list) kí tự và mã code tương ứng. Code này sẽ đổi toàn bộ unicode ở dòng đầu thành kí tự

**Bài 2.10: Đổi kí tự web - Thực hành**

Code phần đổi kí tự dạng &#225

Đọc dòng đầu file raw\_data.txt

Đổi mã code dạng &#225 thành kí tự

Ghi kết quả ra file test.txt

**Bài 2.11Xủ lý điểm thi**

Điểm thi hiện là 1 dòng có dạng: Toán : 5.50 Ngữ Văn: 6.60 ,....

Có người thi môn này có người thi môn kia, có người thi nhiều môn có người thi ít môn.

Bài này mình xử lý cái dòng điểm thi trên thành danh sách điểm thi theo thứ tự, ai không có điểm môn nào thì đánh dấu -1 vào đó, đồng bộ mọi thứ.

**Bài 2.13: Xử lý làm sạch toàn bộ**

Kết nối các công đoạn làm sạch để xử lý toàn bộ raw\_data.txt để tạo ra file data sạch clean\_data.csv

**Data bài 2: clean\_data.csv (dữ liệu đã làm sạch)**

Dữ liệu này sẽ được dùng ở bài sau để phân tích, giải đáp các câu hỏi, vẽ biểu đồ.

file này mở bằng sublime text sẽ ra dạng chữ, mở bằng excel sẽ ra kí tự đặc biệt, cần làm theo các bước mình nói để xem đúng dạng

PHần 3: Phân tích dữ liệu

A. **Môn nào được chọn nhiều nhất - Barchart**

Đếm số người không đăng kí hoặc bỏ thi từng môn

**Bài 3.1: So sánh Excel vs Python**

Excel: Xử lí các bước cơ bản thì nhanh, tốc độ chậm, không xử lí được quá nhiều

Python: Kiểm soát được nhiều hơn, làm được các phép xử lý phức tạp, biến đổi dữ liệu, vẽ biểu đồ đa dạng

Với dự án lớn thì cần dùng python để dễ chuẩn bị data cho bộ phận khác hoặc tích hợp nhiều công nghệ khác như machine learning, AI, deep learning.

Ngoài ra nếu có nhiều data đi ra đi vào, python được dùng như 1 cầu nối để giao tiếp với các hệ thống data (data pipeline)

**Bài 3.3: Chuẩn bị data - Số người không thi hoặc bỏ thi**

Chuẩn bị 3 list sau để chuẩn bị vẽ hình:

not\_take\_exam\_percentage: Phần trăm số người không đăng kí hoặc bỏ thi từng môn

subjects: Tên các môn học

not\_take\_exam: Số người bỏ thi hoặc không đăng kí

**Bài 3.4: Vẽ hình - Số người bỏ thi hoặc không đăng kí**

**HƯỚNG DẪN 4**

**Học sinh thường thi mấy môn - Piechart**

**Bài 4.1: Giới thiệu piechart**

Dùng biểu đồ piechart để tìm xem học sinh thường thi bao nhiêu môn

**Bài 4.2: Chuẩn bị dữ liệu cho piechart**

Mình cùng bạn tạo ra list

num\_of\_exam\_taken - bao nhiêu học sinh thi 0 môn, 1 môn, 2 môn, ....

piechart sẽ tự đổi danh sách trên thành phần trăm và vẽ hình

**Code bài 4.2**

Chuẩn bị data cho piechart, có bao nhiêu học sinh thi 0 môn, 1 môn, 2 môn, ...

**Bài 4.3: Vẽ piechart**

Với barchart, thường mình đưa vào danh sách (list) giá trị, nó tự đổi thành phần trăm (%)

**HƯỚNG DẪN 5**

**Điểm trung bình theo số lượng môn thi - Barchart**

**0/3 BÀI VIẾT ĐÃ HOÀN TẤT**

**Bài 5.1: Điểm trung bình theo số lượng môn thi**

**HƯỚNG DẪN 6**

**Điểm trung bình theo nhóm tuổi - Linechart, Barchart**

**0/6 BÀI VIẾT ĐÃ HOÀN TẤT**

**Bài 6.1: Càng lớn càng điểm thấp?**

Phân tích để trả lời câu hỏi: Có phải tuổi càng cao điểm càng thấp không?

**Bài 6.2: Vẽ barchart số học sinh theo nhóm tuổi**

**HƯỚNG DẪN 7**

**Họ nào phổ biến nhất? - Barchart**

**Bài 7.1: Họ nào phổ biến nhất**

**Bài 7.2: Cách tự kiểm tra code**

**Code bài 7.2s**

Chuẩn bị 2 dánh sách:

1. Các họ thí sinh đi thi
2. Có bao nhiêu học sinh có các họ đó

**Bài 7.3: Cách tìm danh sách số lớn nhất**

Bài toán này được áp dụng để tìm danh sách chứa các họ có nhiều học sinh nhất

**Bài 7.4: Tạo danh sách đã sắp xếp**

**Bài 7.5: Vẽ biểu đồ**

**BONUS: 20 tên phổ biến nhất**

Chỗ lastname = s\_name[0] bạn đổi thành s\_name[-1] là lấy được tên của người đó, phần còn lại y hệt, [-1] nghĩa là vị trí cuối cùng

**BONUS: Tên học sinh dài nhất**

Tên dài nhất: 02033237 - ĐOÀN HUỲNH NGUYỄN CHÂU THANH TÚ

Rất đơn giản, mình lặp qua danh sách students, dùng hàm len() để tính độ dài (số chữ cái, tính cả dấu cách), sau đó tìm cái lớn nhất.

**Một vài lối sửa và update**

// Sửa list thành Dic => Độ phức tạp thấp hơn



- oh đúng vậy rồi, mấy đứa 10 điểm có mỗi 2 dấu cách, sửa như bạn Nguyen Hai là ok replace("10"," 10"). Làm data chạy thí nghiệm có nhiều lỗi vặt lắm. Tại không như làm phần mềm thường chỉ có chạy đúng hay sai.

scores = scores.replace(" 10", " 10") thay 1 space thành 2 space bạn.

Bôi đen trong sublinetext:

 chỗ này bạn bôi đen cái kí tự cuối cùng của dòng đầu, sau đó bấm ctrl+D liên tục trong sublime text là nó sẽ tự bôi đen toàn bộ các kí tự xuống dòng nhé

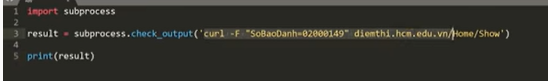
- Thi trước tuổi

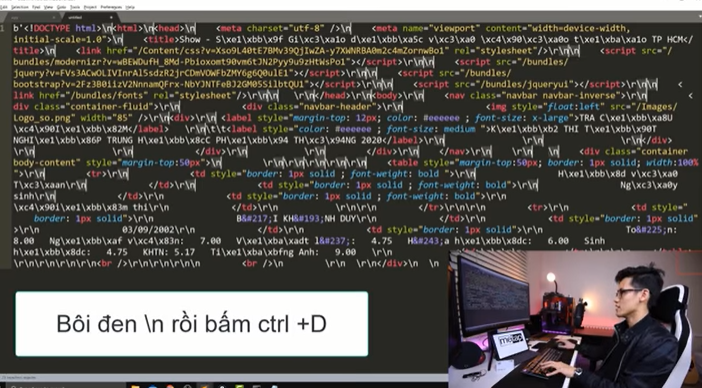
- Thi lớn tuổi

- Họ nào phổ biến

- Môn nào thi phổ biến nhất

- Điểm thi của người lớn tuổi có cao hơn người trẻ không?

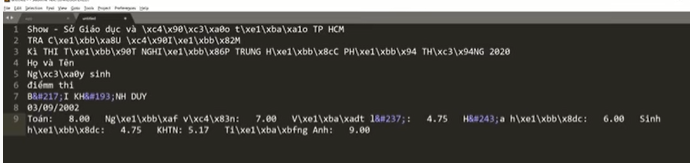


Copy biến result để dán sang file khác

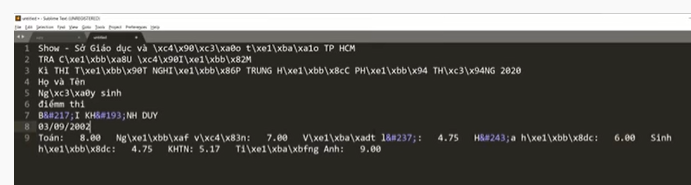
LÀm sạch dữ liệu:

Bước 1: Xóa các dữ liệu không cần thiết – Bôi đen các chữ nên google search -> Mã UTF8

- Xóa các thẻ <> HTML: Chỉ quan tâm đến nội dung

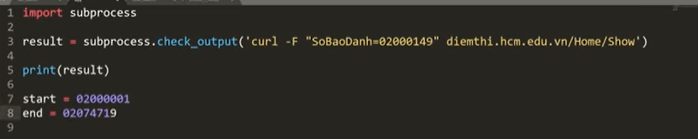
-

- Bước 2: Làm sạch tiếp =>



- Chuyển quy luật thành code

Lấy toàn bộ dữ liệu:



Sử dụng vòng lặp để lấy dữ liệu cho cả tập thay start và end vào chỗ SoBaoDanh

-1 : dữ liệu ko thu thập được

- Làm Sạch dữ liệu => file.CSV  
Bước 3:

Tổng kết

Đây là video cuối cùng rồi, cảm ơn các bạn rất nhiều. Mình xin tổng kết lại các kiến thức đã học.

1. Quy trình 1 bài phân tích đơn giản: Thu thập dữ liệu - Làm sạch - Phân tích - Trình bày kết quả

2. Cách lấy dữ liệu từ web, kĩ năng này gọi là web crawling, cái mình làm rất đơn giản vì trang web của bộ rất đơn giản, sau này bạn có thể học thêm về web crawler, cách vét dữ liệu từ web, kĩ năng này rất dễ kiếm tiền.

2. Các hàm python để xử lý string (string manipulation). Từ nay, mình mong bản sẽ tự tin hơn nhiều khi làm việc với string, chỉ cần có quy luật, mình đều có thể code được. Mọi biến đổi đều có thể làm được với python.

3. Cách dùng vòng for, để lặp qua string, lặp qua list để xử lý dữ liệu

4. Cách dùng matplotlib để làm biểu đồ.

5. Thử qua bài toán sắp xếp 2 list (unit cuối). Bài này yêu cầu mình phải hiểu rõ vòng lặp, thế nào là index của list, giá trị của list.

Cảm ơn các bạn đã cùng Dũng đi đến hết khóa học.

Chân thành cảm ơn và hẹn gặp lại các Bạn.