

XỬ LÝ BỘ DỮ LIỆU ĐIỂM THI ĐẠI HỌC NĂM 2020

Trần Nhật Nam-19521872¹ and Trần Thành Luân-19521810²

¹ Đại học CNTT – ĐH QGTPHCM, tp.HCM

² Đại học CNTT – ĐH QGTPHCM, tp.HCM

19521872@gm.uit.edu.vn

Tóm tắt: Kỳ thi THPT năm 2021 sắp diễn ra trong tình hình dịch bệnh đang căng thẳng việc thu thập bộ dữ liệu 2020 nhằm góp phần rất lớn trong việc làm đề thi, đánh giá chất lượng học sinh. Từ các kết quả góp phần trong việc tìm ra những con số đánh giá được khối ngành hot cũng như ngành nào ít được thí sinh đăng kí. Sau quá trình được trao đổi kiến thức với bộ môn Thu Thập và Tiền xử lý dữ liệu với sự hướng dẫn của thầy Lưu Thanh Sơn. Chúng tôi đã hoàn thiện bộ dữ liệu Điểm thi Đại học năm 2020 này nhằm phục vụ nghiên cứu. Bài báo cáo được chia làm 5 phần. Phần đầu tiên chúng tôi sẽ giới thiệu tổng quát và mục tiêu của việc thu thập. Phần tiếp theo nêu lên nguồn thu thập và cách thức thu thập. Phần sau đó là quá trình là sạch dữ liệu. Phần kế tiếp là CodeBook và phần cuối cùng là dựa theo nghiên cứu và quán sát của chúng tôi sẽ đưa ra các vấn đề và cách giải quyết chúng. Toàn bộ bài báo cáo được chúng tôi thực hiện trên ngôn ngữ python.

Các mục chính: Giới thiệu, Phương pháp thu thập, Tiền xử lý dữ liệu, Bộ dữ liệu tidy data, Kết luận và hướng pháp triển.

1 Giới thiệu.

1.1 Tổng quát.

Hiện nay, kì thi THPT là kì thi quan trọng bậc nhất ở Việt Nam. Đây là cơ sở dữ liệu quan trọng để đánh giá chất lượng giáo dục phổ thông, đồng thời là cơ sở để các trường đại học, cao đẳng xác định ngưỡng điểm xét tuyển. Các thí sinh, phụ huynh có thể căn cứ vào điểm từng môn thi và tổ hợp xét tuyển để đăng ký nguyện vọng xét tuyển cho phù hợp. Do đó việc thu thập được bộ dữ liệu điểm thi THPT đặc biệt là năm gần nhất là năm 2020 là vô cùng quan trọng.

1.2 Mục tiêu.

Nhận thấy được tầm quan trọng của bộ dữ liệu này cùng với những kiến thức đã được học tập và trao đổi trong bộ môn Thu thập và tiền xử lý dữ liệu. Chúng tôi bắt đầu lên kế hoạch thực hiện các bước để hoàn thiện bộ dữ liệu này. Từ đó có thể dựa vào bộ dữ liệu chúng tôi có thể đưa ra các vấn đề và giải quyết chúng. Hy vọng bài viết này giúp ích được cho mọi người.

1.3 Công cụ sử dụng.

Việc thu thập dữ liệu có thể sử dụng rất nhiều các ngôn ngữ khác nhau để thực hiện. Trong bài toán này chúng tôi chọn python vì đây là ngôn ngữ phổ biến, đơn giản, và các công cụ mà nó cung cấp rất mạnh và đa dạng. Và để sử dụng python một cách đơn giản và kết quả hiển thị một cách trực quan hơn IDE mà chúng tôi lựa chọn là Pycharm, nó có giao diện khá đơn giản và dễ sử dụng.

2 Phương pháp thu thập.

2.1 Xác định nguồn thu thập.

Hiện nay việc quản lý điểm thi được bộ giáo dục quản lý. Và đây cũng là nguồn uy tín nhất chúng ta có thể dùng để thu thập dữ liệu: diemthi.hcm.edu.vn

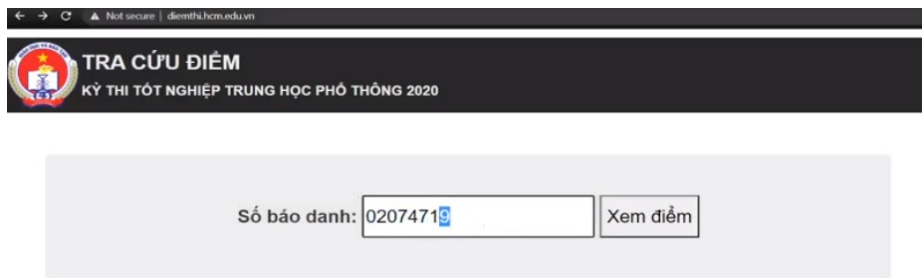


Fig. 1. Giao diện trang web.

Có thể thấy giao diện rất đơn giản chúng ta chỉ cần nhập và số báo danh và được trả về kết quả như sau.

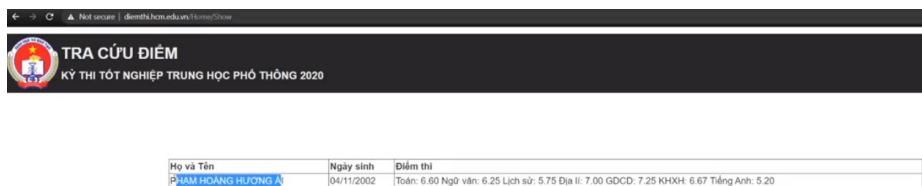


Fig. 1. Kết quả giao diện trả về.

Như chúng tôi tìm hiểu được thì số báo danh của thí sinh được bắt đầu từ 02000001 và kết thúc 02074718.

Từ đó chúng tôi có ý tưởng sẽ sử dụng công cụ của python thu thập dữ liệu trang web dưới dạng html để từ đó có thể là sạch.

2.2 Cách thức thu thập.

Sau khi xem trang source code sau trang web ta có thể thấy file html sau đây.

```
<section id="loginForm">
<form action="/Home/Show" method="post">
<div style="text-align:center">
<label>Số báo danh:</label>
<input data-val="true" data-val-required="The Số báo danh field is required." id="SoBaoDanh" name="SoBaoDanh"
type="text" value="" />
<input type="submit" value="Xem điểm" />
<span class="field-validation-valid text-danger" data-valmsg-for="SoBaoDanh" data-valmsg-replace="true"></span>
```

Fig. 1. Source code sau trang web.

Phần này có liên quan một đến một phần kiến thức là web. Ở đây chúng tôi sẽ nói qua về phần method = "post". Post ở đây là chúng ta có thể gửi 1 thông điệp nào đó đến trang web là trang web sẽ trả về kết quả. Có thể thấy name = "SoBaoDanh" là nơi chúng ta nhập thông tin và action = "/Home/Show" là đích tới của hành động.

```
curl -F "soboadanh=02000001" diemthi.hcm.edu.vn/Home/Show
```

Trong python ta có cú pháp như sau dùng để thu thập dữ liệu đăng sau trang web.

Và để phục vụ trong quá trình làm sạch bằng python chúng ta cần sử dụng thư viện subprocess

Tiếp theo chúng ta sẽ tiến hành tạo 1 file txt chứa source code sau trang web của tất cả các số báo danh.

```
import subprocess

start = 2000001
end = 2074719

file = open("raw_data.txt", "w") # Tao một file txt mới hoàn toàn

for sbd in range(start, end):
    command = 'curl -F "SoBaoDanh=' + str(sbd) + '" diemthi.hcm.edu.vn/Home/Show'
    result = subprocess.check_output(command)
    file.write(str(result) + "\n")
```

Fig. 1. Code đọc source code vào file txt.

Ta đã có cú pháp lấy dữ liệu từ web cùng với thư viện được cung cấp sẵn trong python chúng ta sẽ tiến hành tạo file và đọc code html của từng mã số báo danh vào file txt. Cuối cùng ta thu được một file như sau.

Fig. 1. File txt sau khi lấy dữ liệu từ web.

Mỗi dòng sẽ là source code của một số báo danh.

Sau khi có file draw-data là source của của 74719 thí sinh chúng ta sẽ tiến hành làm sạch dữ liệu.

3 Tiền xử lý dữ liệu.

Do số lượng các dòng trong file draw-data là quá lớn. Mà chúng đều là source code dc viết từ html cho nên các cú pháp cũng giống nhau vì vậy chúng ta sẽ thử làm sạch dữ liệu của 1 dòng trước rồi từ đó áp dụng cho tất cả số báo danh.

3.1 Làm sạch một dòng dữ liệu.

Bước 1: Quan sát một dòng dữ liệu.

Bước đầu tiên ta sẽ tiến hành mở file dữ liệu thô và đọc dòng dữ liệu đầu tiên để quan sát. Chúng tôi sẽ tạo một file test riêng để đọc chúng dễ hơn.

```
file = open("raw_data.txt", "r")

# Read first student
data = file.readline()

print(data)
```

```
# make data becomes a list
data = data.split("\n")

# write data to test.txt
file = open("test.txt", "w")
for i in range(len(data)):
    file.write(data[i] + "\n")
```

Fig. 1. Code để đọc một dòng txt đầu tiên.



Fig. 1. Kết quả của dòng đầu tiên.

Có thể thấy là dữ liệu khá là rối chúng ta sẽ tiến hành các bước là sạch như sau:

Bước 2: Làm sạch đến dòng cần sử dụng.

```
# 1. Xóa kí tự đặc biệt.
for i in range(len(data)):
    data[i]=data[i].replace("\r", "")
    data[i]=data[i].replace("\t", "")
    data[i]=data[i].strip()

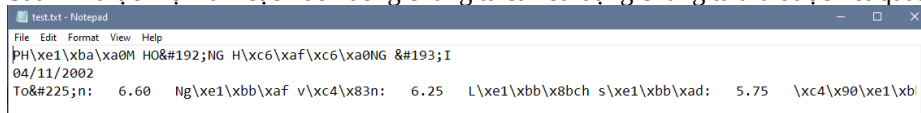
# 2. Xóa kí tự trong tag..
for i in range(len(data)):
    tags=[]
    for j in range(len(data[i])):
        if data[i][j]=="<":
            begin=j
        if data[i][j]==">":
            end=j
    tags.append(data[i][begin:end+1])
```

```

    for tag in tags:
        data[i]=data[i].replace(tag, "")
# 3. Xoa ki tu thua
for i in range (len(data)):
    data[i]=data[i].strip()
unempty_line=[]
for i in range (len(data)):
    if data[i] != "":
        unempty_line.append(data[i])
data= unempty_line
# 4. Chon du lieu can su dung
names=data[7]
dob=data[8]
scores=data[9]
data=[names, dob, scores]

```

Sau khi thực hiện làm sạch đến dòng chúng ta cần sử dụng chúng ta thu được kết quả:



```

test.txt - Notepad
File Edit Format View Help
PH\xe1\xba\x80 H\xca\xaf\xca\x80NG &#193;I
04/11/2002
To&#225;n: 6.60 Ng\xe1\xba\xaf v\xca\x83n: 6.25 L\xe1\xba\x8bch s\xe1\xba\xad: 5.75 \xc4\x90\xe1\xba

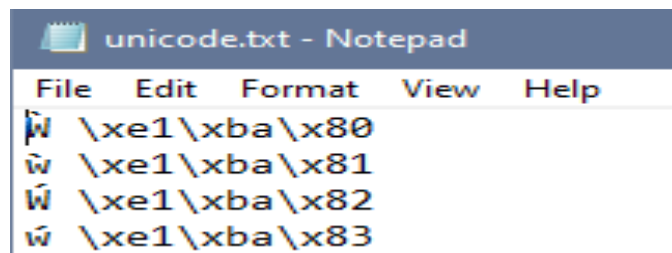
```

Fig. 1. Các dòng dữ liệu thô cần sử dụng.

Khi đã làm sạch tới đây thì có thể thấy dữ liệu thô ban đầu trở nên ngắn hơn. Quan sát có thể thấy chúng ta sẽ lấy được thông tin như là họ tên, ngày sinh, điểm thi.

Bước 3: Chuyển kí tự unicode và kí tự đặc biệt về tiếng việt.

Để thực hiện bước này chúng ta cần tạo một bảng kí tự unicode tương đương với các kí tự tiếng việt. Chúng ta có thể dễ dàng tìm thấy trên google.



```

unicode.txt - Notepad
File Edit Format View Help
W \xe1\xba\x80
w \xe1\xba\x81
W \xe1\xba\x82
w \xe1\xba\x83

```

Fig. 1. Minh họa về file unicode.

Tiếp theo chúng ta sẽ tiến hành chuyển đổi hết các kí tự này và kí tự đặc biệt về tiếng việt.

```

# 5. load unicode table
chars =[]
codes =[]

```

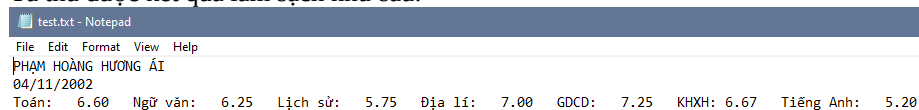
```

file=open("unicode.txt",encoding="utf8")
unicode_table=file.read().split("\n")
for code in unicode_table:
    x=code.split(" ")
    chars.append(x[0])
    codes.append(x[1])
# 6. Chuyen ki tu dac biet thanh ten chu cai tieng viet
# 6.1 Chuyen ki tu unicode.
for i in range(len(unicode_table)):
    names=names.replace(codes[i],chars[i])
    scores=scores.replace(codes[i],chars[i])
# 6.2 Chuyen ki tu char.
for i in range(len(names)):
    if names[i:i+2]=="&#":
        names=names[:i]+chr(int(names[i+2:i+5]))+
names[i+6:]
for i in range(len(scores)):
    if scores[i:i+2]=="&#":
        scores=scores[:i]+chr(int(scores[i+2:i+5]))+
scores[i+6:]

```

Fig. 1. Làm sạch kí tự đặc biệt.

Ta thu được kết quả làm sạch như sau:



test.txt - Notepad

File Edit Format View Help

PHẠM HOÀNG HƯƠNG ÁI

04/11/2002

Toán: 6.60 Ngữ văn: 6.25 Lịch sử: 5.75 Địa lí: 7.00 GDCD: 7.25 KHXH: 6.67 Tiếng Anh: 5.20

Fig. 1. Dữ liệu sau khi chuyển qua tiếng việt.

Có thể thấy là dữ liệu đã rõ ràng và dễ hiểu. Từ đó dựa vào đây chúng ta sẽ tạo nên dòng tidy đầu tiên.

Bước 4: Tạo dòng tidy data hoàn chỉnh.

Do có người thi môn này môn kia. Vì thế trong bước này chúng ta sẽ xử lý các dòng thi theo thứ tự, ai không thi môn nào thì đánh dấu là -1. Tự đó đồng bộ mọi thứ.

Dựa vào quát sát tôi liệt kê ra được danh sách các môn theo thứ tự như sau : Toán, ngữ văn,khxx,khtn,lịch sử,địa lý,gdcd,sinh học,vật lí, hóa học, tiếng anh.

```

# 6.3 Change lower case
names=names.lower()
scores=scores.lower()
# 6.4 split dob.
dob_list=dob.split("/")
dd=int(dob_list[0])
mm=int(dob_list[1])
yy=int(dob_list[2])

```

```
# 6.5 split scores
scores=scores.replace(":", "")
scores=scores.replace("khxh ", "khxh ")
scores=scores.replace("khtn ", "khtn ")
scores=scores.replace(" 10", " 10")

scores_list=scores.split(" ")

data=[sbd_list, names.title(), str(dd), str(mm), str(yy)]

# 6.6 add scores in data.
for subject in ['toán', 'ngữ văn', 'khxh', 'khtn', 'lịch
sử', 'địa lí', 'gdcđ', 'sinh học', 'vật lí', 'hóa học', 'tiếng
anh']:
    if subject in scores_list:
        data.append(str(float(scores_list[scores_list.index(subject)+1])))
    else:
        data.append('-1')
```

Fig. 1. Code tạo ra dòng tidy đầu tiên.

Kết quả chúng ta thu được như sau:

```
02000001
Phạm Hoàng Hương Ái
4
11
2002
6.6
6.25
6.67
-1
5.75
7.0
7.25
-1
-1
-1
5.2
```

Fig. 1. Dữ liệu tidy cả thí sinh đầu tiên.

Bước 5: Làm sạch tất cả dòng dữ liệu thô.

Sau khi biết các làm sạch 1 dòng do cấu trúc từng dòng raw data đều viết bằng html. Chúng ta sử dụng vòng lặp for cho tất cả các dòng và thay vì lưu vào file txt chúng ta sẽ lưu thành file csv để dễ dàng sử dụng.

```
# in ra file csv.
with
open("clean_data.csv",mode="a",encoding="utf8",newline='')
) as file_csv:
    write=csv.writer(file_csv)
    write.writerow(data)
```

Fig. 1. Lưu từng dòng data vào file clean_data.

Cuối cùng chúng ta sẽ thu được bộ dữ liệu hoàn chỉnh về điểm thi THPT quốc gia năm 2020.

File Edit Selection Find View Goto Tools Project Preferences Help

clean_data.csv

raw_data.txt

1034

02001034,Nguyễn Thiên Phúc,10,7,2002,7.4,7.0,-1,5.75,-1,-1,-1,5.0,6.5,5.75,8.0

1035

02001035,Trần Phúc,31,10,2002,7.0,6.75,-1,6.0,-1,-1,-1,5.0,7.25,5.75,4.8

1036

02001036,Đỗ Lý Kim Phụng,20,1,2002,8.0,6.75,-1,7.08,-1,-1,-1,6.75,7.0,7.5,8.0

1037

02001037,Đỗ Thị Như Phụng,9,10,2002,8.2,7.5,-1,6.42,-1,-1,-1,5.75,5.75,7.75,5.2

1038

02001038,Thái Nguyễn Đan Phụng,22,9,2002,6.0,7.25,-1,5.83,-1,-1,-1,5.0,6.0,6.5,4.6

1039

02001039,Nguyễn Minh Phước,25,1,2002,5.8,5.75,-1,4.58,-1,-1,-1,4.75,4.25,4.75,4.8

1040

02001040,Lê Ngọc Minh Phương,18,9,2002,9.0,6.25,-1,6.25,-1,-1,-1,4.5,7.25,7.0,7.4

1041

02001041,Nguyễn Bá Phương,26,1,2002,8.8,6.5,-1,8.25,-1,-1,-1,8.0,8.5,8.25,9.4

1042

02001042,Nguyễn Nhi Nam Phương,2,6,2002,7.8,7.0,-1,5.75,-1,-1,-1,4.5,6.0,6.75,7.0

1043

02001043,Nguyễn Phương Ngọc Phương,29,12,2002,7.8,6.5,-1,7.42,-1,-1,-1,7.0,7.5,7.75,6.2

Fig. 1. Dữ liệu sạch sau khi đã hoàn tất làm sạch.

Sau các bước làm sạch chúng tôi đã đưa ra được bộ dữ liệu hoàn chỉnh. Từ kết quả này giúp ích rất nhiều trong các bài toán chúng tôi sẽ đưa ra tiếp theo.

4 Bộ dữ liệu sạch tidy data.

Sau quá trình làm sạch ở phần tiếp theo này chúng tôi sẽ tiến hành mở file csv ở dạng excel để thầy cùng các bạn có thể hình dung rõ kết quả chúng tôi đã hoàn thành.

4.1 Quan sát dữ liệu bằng excel.

Do chúng ta lưu dưới dạng csv. Ở dạng file này chúng ta có thể được dưới dạng file excel. Vì vậy, để rõ hình dung tôi sẽ sử dụng excel thử.

clean_data.csv

File Origin: 65001: Unicode (UTF-8) | Delimiter: Comma | Data Type Detection: Based on first 200 rows

SBD	Tên	dd	mm	yy	toán	ngữ văn	khxh	khtn	lịch sử	địa lí	gdcđ	sinh học	vật lí	hóa học	tiếng anh
2000001	Phạm Hoàng Hương Ái	4	11	2002	6.6	6.25	6.67	-1	5.75	7	7.25	-1	-1	-1	5.2
2000002	Đặng Huỳnh Vĩnh An	13	12	2002	8.2	7.75	7.58	-1	7	7.25	8.5	-1	-1	-1	7
2000003	Lâm Nguyễn Mộng Thủy An	6	4	2001	6.8	6.75	6.92	-1	4.75	7.75	8.25	-1	-1	-1	6
2000004	Lê Tiểu Hoàng An	18	11	2002	7.8	6.25	-1	6.25	-1	-1	-1	7	5.5	6.25	5.6
2000005	Lư Thuận An	14	1	2002	6.4	6.5	-1	6.17	-1	-1	-1	5.5	6.75	6.25	8.2
2000006	Mai Bình An	14	6	2002	6.8	7.5	7.58	-1	6.75	7.5	8.5	-1	-1	-1	-1
2000007	Mai Xuân An	16	3	2002	8.4	8.25	-1	6.67	-1	-1	-1	5.25	7.5	7.25	6.4
2000008	Nguyễn Huỳnh Khánh An	28	7	2002	6.8	7	-1	4.33	-1	-1	-1	3.5	5	4.5	4.4
2000009	Nguyễn Trần Hòa An	14	11	2002	7.2	8	7.92	-1	6.75	8.25	8.75	-1	-1	-1	8.4
2000010	Nguyễn Vương Thùy An	14	8	2002	8.4	7.75	-1	6.33	-1	-1	-1	4	8	7	6.4
2000011	Phạm Thị Hồng An	11	3	2002	6.4	7.75	7.08	-1	6.25	6.5	8.5	-1	-1	-1	5.2
2000012	Tỷ Thiệu Thuận An	28	9	2002	8	5.25	8.17	-1	7.25	8	9.25	-1	-1	-1	8.2
2000013	Võ Thiên An	27	3	2002	8.4	5.75	-1	7.08	-1	-1	-1	5.75	7.75	7.75	5.8
2000014	Vũ Thanh An	19	11	2002	8.4	7.5	7.67	-1	8	7.25	7.75	-1	-1	-1	9.2
2000015	Bùi Nguyễn Minh Anh	18	3	2002	8.2	6.75	-1	6.08	-1	-1	-1	5.25	6.25	6.75	8.6
2000016	Bùi Thụy Quỳnh Anh	24	7	2002	8.6	6.75	-1	6.17	-1	-1	-1	5.5	6.25	6.75	7.2
2000017	Bùi Trần Lan Anh	11	3	2002	8.6	6.75	-1	7.42	-1	-1	-1	7.75	7.25	7.25	8.8
2000018	Cao Ngọc Phương Anh	6	6	2002	8	6.5	5.92	-1	3.75	6.5	7.5	-1	-1	-1	8
2000019	Châu Xuân Anh	30	9	2002	8.2	4.5	-1	7.33	-1	-1	-1	7.75	6	8.25	7.2
2000020	Chung Vũ Thủy Anh	8	3	2002	8.4	6.25	-1	5.75	-1	-1	-1	5.25	6.75	5.25	7.4
2000021	Dương Hoàng Tuấn Anh	11	9	2002	7.4	6.25	-1	6.67	-1	-1	-1	5.5	6	8.5	7
2000022	Đào Quỳnh Anh	11	6	2002	8.8	6	-1	5.92	-1	-1	-1	5.75	8	4	6.4
2000023	Đặng Minh Phi Anh	24	11	2002	7.4	6.5	-1	5.58	-1	-1	-1	5	5.75	6	4.8
2000024	Hà Lê Kiều Anh	3	10	2002	7.2	8	-1	5.75	-1	-1	-1	6.25	7.75	3.25	5.4
2000025	Hà Quỳnh Anh	29	6	2002	8.4	8	6.83	-1	5.75	7.25	7.5	-1	-1	-1	8.4
2000026	Hoàng Kỳ Anh	26	6	2002	8.2	7	6.92	-1	6.25	7	7.5	-1	-1	-1	6.6
2000027	Hồ Minh Anh	15	12	2002	7.6	7	-1	6.58	-1	-1	-1	5	6.25	8.5	6.4
2000028	Hồ Thủy Anh	11	12	2002	7.6	7	-1	6	-1	-1	-1	5.25	5.5	7.25	6.6
2000029	Huỳnh Duy Anh	22	11	2002	7.8	5.25	-1	7.67	-1	-1	-1	7.5	7.25	8.25	7
2000030	Huỳnh Đoàn Minh Anh	13	9	2002	8.2	7.5	-1	7.92	-1	-1	-1	7.5	7.75	8.5	-1
2000031	Lê Minh Anh	19	8	2002	7.8	7.5	6.17	-1	5.25	5	8.25	-1	-1	-1	6.8
2000032	Lê Đan Quỳnh Anh	3	2	2002	7.8	8	6.92	-1	5	7.75	8	-1	-1	-1	7.6
2000033	Lê Hoàn Minh Anh	4	1	2002	8.4	7.25	7.33	-1	7	7	8	-1	-1	-1	7
2000034	Lê Ngọc Quỳnh Anh	27	6	2002	7.8	6.5	-1	5.92	-1	-1	-1	3	7.5	7.25	-1
2000035	Lê Nguyễn Duy Anh	21	2	2002	6.8	5.5	6.08	-1	4	7	7.25	-1	-1	-1	5
2000036	Lê Nữ Hoàng Anh	20	8	2002	7.6	6	-1	7.17	-1	-1	-1	6.75	7.5	7.25	7.4
2000037	Ngô Quế Anh	16	1	2002	7.2	5.5	6.08	-1	5	5.5	7.75	-1	-1	-1	5
2000038	Nguyễn Cao Phương Anh	11	12	2002	7.8	7	7.42	-1	7.25	7.5	7.5	-1	-1	-1	6.8
2000039	Nguyễn Đức Ngọc Anh	20	5	2002	7.2	7	6.08	-1	4	6.5	7.75	-1	-1	-1	5.2
2000040	Nguyễn Đặng Minh Anh	2	7	2002	8.2	6.75	-1	6.75	-1	-1	-1	4.75	8.25	7.25	5.2

Fig. 1. Mở file clean_data.csv trong ứng dụng excel.

4.2 Codebook.

Table 1. Codebook mô tả bộ dữ liệu.

Thông tin	Nội dung
Tên Bộ dữ liệu	Dữ liệu điểm thi THPTQG năm 2020
Nguồn thu thập và cách thức thu thập	<p>1.Thu thập từ trang web :Diemthi.hcm.edu.vn</p> <p>2. Cách thức thu thập: Bằng cách truy cập vào source code bên trong trang web nhờ vào công cụ python. Từ đó tiếp sử dụng python trong việc làm sạch.</p>
Số điểm dữ liệu	74444 điểm dữ liệu tương với số lượng thí sinh
Số thuộc tính	16 thuộc tính.

Thông tin các tên thuộc tính	<ul style="list-style-type: none"> • SBD: Số báo danh. • Ten: Họ và tên thí sinh. • Mm/dd/yy: Ngày, tháng, năm sinh của thí sinh. • 11 thuộc tính tiếp theo: tương ứng với 11 môn thí sinh có thể thi.
Thông tin người thực hiện.	<p>Trần Nhật Nam - Sinh viên ngành Khoa học dữ liệu thuộc đại học CNTT-ĐH QGTPHCM. 19521872@gm.uit.edu.vn.</p> <p>Trần Thành Luân - Sinh viên ngành Khoa học dữ liệu thuộc đại học CNTT-ĐH QGTPHCM. 19521810@gm.uit.edu.vn.</p>

5 Kết luận và hướng giải quyết.

Với sự hỗ trợ của ngôn ngữ python cũng với những kiến thức đã được học ở bộ môn Thu Thập và tiền xử lý dữ liệu. Dưới sự hướng dẫn của thầy Lưu Thanh Sơn đã tận tình chỉ dẫn chúng em đã hoàn thiện bộ dữ liệu Điểm thi THPTQG năm 2020. Hy vọng bộ dữ liệu có thể giúp ích nhiều trong việc nghiên cứu cũng như phân tích dự đoán trong tương lai. Và để góp phần xây dựng vấn đề liên quan chúng em xin đề ra một số vấn đề và phương hướng giải quyết chúng như sau:

5.1 Đặt vấn đề.

Do là bộ dữ liệu liên quan đến điểm thi chúng em đã tự tìm hiểu và đưa ra được một số câu hỏi như:

1. Có 11 môn thi vậy thì số lượng thí sinh thường thi bao nhiêu môn ?
2. Vậy thì điểm thi trung bình của từng học sinh ứng với số môn mà họ thi là bao nhiêu ?
3. Thí sinh thường thi môn nào nhiều nhất ?

Và còn rất nhiều vấn đề liên quan chúng ta có thể giải quyết từ bộ dữ liệu này. Nhưng sau đây là phương hướng giải quyết vấn đề mà chúng em đưa ra.

5.2 Hướng giải quyết vấn đề.

Vấn đề 1,2:

Theo hướng giải quyết của chúng em đưa ra thì ban đầu chúng ta cần tạo ra 1 list các số tương ứng với số môn thi là từ 0 đến 11. Vì sẽ có trường hợp thí sinh đăng kí nhưng bỏ thi và chúng ta có 11 môn nên biến của chúng ta sẽ có 12 biến.

Tiếp theo chúng sẽ lọc qua tất cả các thí sinh là đếm số môn của từng thí sinh từ đó công vào biến tương ứng với số môn họ thi.

```
# Mở file clean.
with open("clean_data.csv", encoding="utf8") as file:
    data = file.read().split("\n")
# Bỏ dòng đầu là các dòng chỉ tên thuộc tính
header = data[0]
students = data[1:] #Thí sinh tiếp theo bắt đầu ở dòng
tiếp theo
# Tổng Số lượng học sinh.
total_student = len(students)
# turn each student to a list
for i in range(len(students)):
    students[i] = students[i].split(",")

# remove empty list (end of file)
students.pop()

#Số lượng thí sinh tương ứng với 0,1,...11 môn thi
num_of_exam_taken = [0,0,0,0,0,0,0,0,0,0,0,0]

for s in students:
    count = 0
    for i in range(11):
        if s[i+5] != "-1":
            count += 1

    num_of_exam_taken[count] += 1

print(num_of_exam_taken)
```

Fig. 1. Code để tạo ra danh sách số lượng ứng với số môn thi.

Kết quả mà chúng ta thu được là: [0, 80, 120, 2600, 4329, 300, 2507, 64507, 0, 0, 0, 1].

Để hình dùng ra hơn, để vẽ ra được biểu đồ cho tất cả mọi người có thể hình dùng. Chúng ta có thể dùng code python để vẽ trực quan hoặc có thể sử dụng công cụ khác. Ở đây, chúng tôi sử dụng một công cụ vẽ online trên trang:

<https://www.rapidtables.com/tools/pie-chart.html>

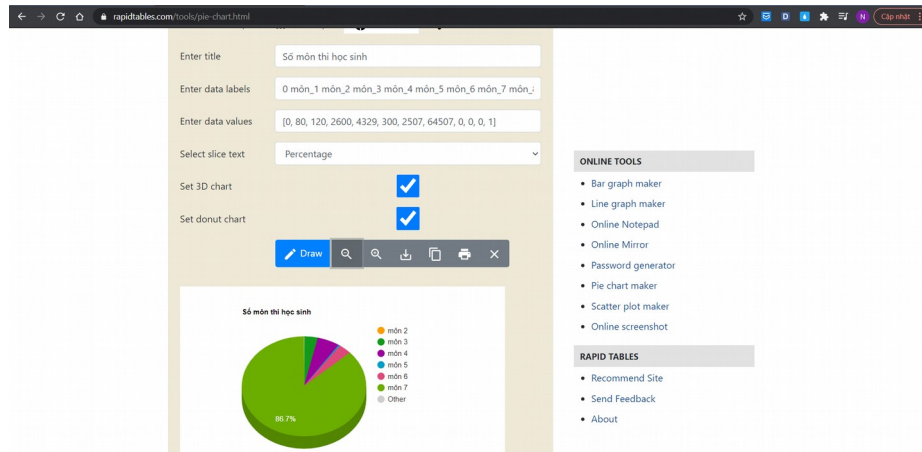


Fig. 1. Giao diện và kết quả trực quan bài toán đầu tiên.

Để giải quyết bài toán thứ 2 là tìm số điểm tương ứng thì dễ thấy ta chỉ cần tổng số điểm của tổng số thí sinh tương ứng với số môn họ thi rồi chia cho số lượng thí sinh thi 0,1,...,11 môn.

```
average = [0,0,0,0,0,0,0,0,0,0,0,0,0]
for s in students:
    count = 0
    total = 0
    for i in range(11):
        if s[i+5] != "-1":
            total += float(s[i+5])
            count += 1

    num_of_exam_taken[count] += 1
    average[count] += total/count

for i in range(12):
    if num_of_exam_taken[i] != 0:
        average[i]=round(average[i]/num_of_exam_taken[i],2)

print(num_of_exam_taken)
print(average)
```

Fig. 1. Code tìm điểm trung bình của các thí sinh ứng với số môn họ thi.

Ta thu được kết quả như sau : [0, 5.68, 6.91, 6.59, 5.82, 6.52, 6.59, 6.6, 0, 0, 0, 4.27].

Kết quả là điểm trung bình ứng với số môn mà dự thi.

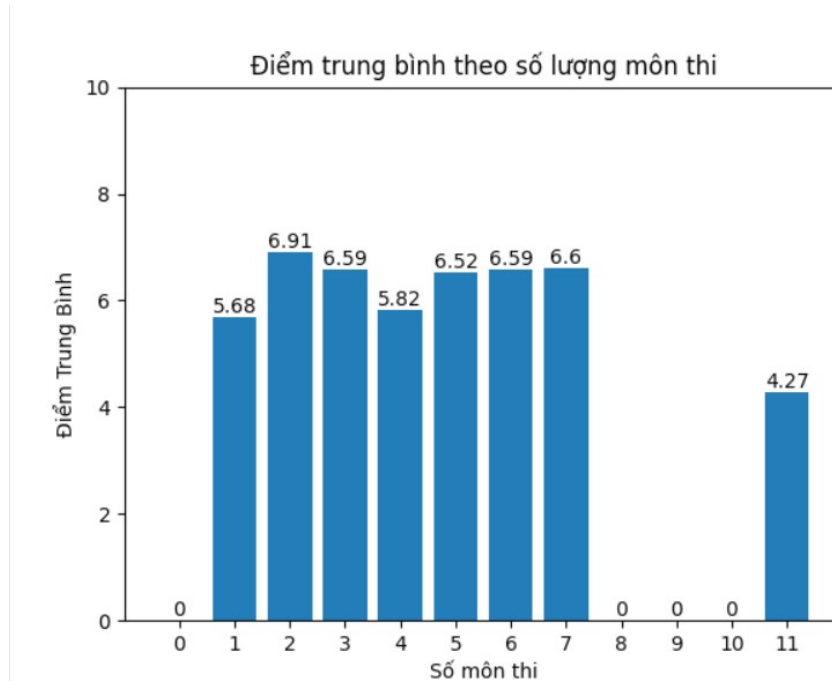


Fig. 1. Điểm thi trung bình theo số lượng môn thi.

Từ 2 vấn đề trên chúng ta có thể đưa ra nhận xét là hầu hết thí sinh đều đăng kí thi 7 môn và lượng số môn còn lại là rất ít. Và có thể kết luận là điểm thi trung bình nằm trong khoảng 6.6.

Vấn đề 3.

Từ vấn đề nêu trên chúng ta lại có thắc mắc là môn thi, hay nhóm ngành nào được dự thi nhiều nhất hoặc là môn nào bỏ thi nhiều nhất.

Và để làm rõ điều này đầu tiên chúng ta cũng cần tạo một list ứng với số môn học, xong lọc qua để tìm số lượng thí sinh bỏ thi môn đó. Sau đó chúng ta cần tìm thêm phần trăm thí sinh bỏ thi môn đó.

```
not_take_exam = [0,0,0,0,0,0,0,0,0,0,0,0]
```

```
# Lọc qua tất cả học sinh.
for s in students:
    # Đếm trong tất cả môn học.
    for i in range(5,16):
        if s[i] == "-1":
            not_take_exam[i-5] += 1
```

```
not_take_exam_percentage = [0,0,0,0,0,0,0,0,0,0,0,0]

# convert to percentage
for i in range(0,11):
    not_take_exam_percentage[i] =
round(not_take_exam[i]*100/total_student, 2)
print(not_take_exam)
print(not_take_exam_percentage)
```

Fig. 1. Code để tìm ra số học sinh bỏ thi từng môn.

Kết quả thu được : + [246, 1802, 49223, 32496, 44129, 44845, 49206, 31839, 31591, 31303, 8666]

+ [0.33, 2.42, 66.12, 43.65, 59.28, 60.24, 66.1, 42.77,

42.44, 42.05, 11.64]

Từ 2 kết quả này chúng chúng ta thu được kết quả trực quan như sau:

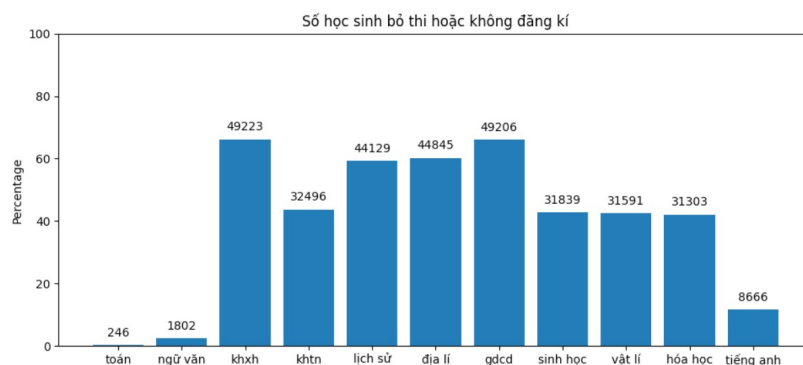


Fig. 1. Số học sinh bỏ thi hoặc không thi.

Có thể thấy thì môn Toán , văn, anh là số lượng thí sinh thi nhiều nhất và thí sinh đăng kí khối A,B nhiều hơn thí sinh đăng kí khối C,D.

5.3 Tổng kết.

Kết thúc bài cáo chúng tôi đã đưa ra được các vấn đề và hướng giải quyết chúng. Hy vọng đây có thể là một tài liệu tham khảo để có thể dựa vào đây giải quyết những vấn đề khác. Thông qua bài báo cáo này thì chúng ta hoàn toàn có thể chủ động trong việc thu thập nguồn dữ liệu của điểm thi những năm sau. Cảm ơn thầy và các bạn và xem qua bài viết. Một lần nữa xin cảm ơn thầy Lưu Thanh Sơn đã đồng hành và chỉ dạy tận tình những kiến thức về thu thập và tiền xử lý dữ liệu để chúng em có thể hoàn thiện bộ dữ liệu này.

