Data Science – Phân tích điểm thi đại học tại thành phố hồ chí minh năm 2020

1. Data mining :

* Công cụ sử dụng : python
* Web site : <http://diemthi.hcm.edu.vn/>
* Sử dụng thư viện : subprocess
* Số lượng data : 75,000 dòng (02000001 - > … )

Start = 2000001

End = 2074719

import subprocess

start = 2000001

end = 2074719

res\_minging\_data = []

for sbd in range(start,end):

    command ='curl -F ' + '\"SoBaoDanh=0' + str(sbd) + '\"' + ' diemthi.hcm.edu.vn/Home/Show'

    data = subprocess.check\_output(command)

    res\_minging\_data.append(data)

with open("raw\_data.txt", "a", *newline*="") as file\_rawData:

    writer = csv.writer(file\_rawData)

    writer.writerow(res\_minging\_data)

1. Clean HTML data manually:
2. Sử dụng string manipulation
3. Sử dụng file unicode ( utf-8) lấy từ <https://www.utf8-chartable.de/> để encode cho dữ liệu

* Sử dụng file: decode.txt để lấy char và code để xử lý unicode trong bài => được file unicode.txt

#load unicode table

f = open('decode.txt', *encoding*= "utf8")

unicode\_table = f.read().split("\n")

for i in range(len(unicode\_table)):

    unicode\_table[i] = unicode\_table[i].split("\t")

f.close()

# print(unicode\_table[0].encode('utf-8'))

print(len(unicode\_table[0]))

fWrite = open('unicode.txt', 'wb+')

for code in unicode\_table:

    fWrite.write((code[1] + " " + code[2] + "\n").encode("utf-8"))

fWrite.close()

1. Kết quả :

File clean\_data.csv có dạng

sbd,name,dd,mm,yy,toán,ngữ văn,khxh,khtn,lịch sử,địa lí,gdcd,sinh học,vật lí,hóa học,tiếng anh,tiếng nhật,tiếng pháp,tiếng trung,tiếng đức,tiếng nga

