Assigment 3: Scraping dữ liệu từ web

Tổng quan: Ở bài tập này chúng ta sẽ lần lượt thực hành các bước cho việc trích xuất text từ web: sraping dữ liệu từ web, làm sạch dữ liệu, lưu trữ dữ liệu để cho quá trình xử lý tính toán. Dữ liệu của bài toán này là một web được trình bày dưới dạng bảng. Bài tập yêu cầu các kiến thức về lập trình Python với các thư viện BeautifulSoup, numpy, pandas, urlopen

```
import pandas as pd
import numpy as np
from urllib.request import urlopen
from bs4 import BeautifulSoup
```

```
# Tạo biến chứa url
url = "http://www.hubertiming.com/results/2017GPTR10K"
html = urlopen(url)
```

```
# Khởi tạo đối tượng BeautifulSoup
soup = BeautifulSoup(html, 'lxml')
type(soup)
```

```
bs4.BeautifulSoup
```

```
# Get the title
title = soup.title
print(title)
```

```
<title>Race results for the 2017 Intel Great Place to Run \ Urban Clash Games!</title>
```

```
# Print out the text
text = soup.get_text()
print(soup.text)
```

```
Race results for the 2017 Intel Great Place to Run \ Urban Clash Games!
```

```
2017 Intel Great Place to Run 10K \ Urban Clash Games
Hillsboro Stadium, Hillsboro, OR
June 2nd, 2017

Email
                        timing@hubertiming.com with results questions. Please include your bib

Huber Timing Home

10K:
```

```
# Truy cập tất cả các link
soup.find_all('a')
```

```
[<a href="timing@hubertiming.com">timing@hubertiming.com</a>,
 <a href="https://www.hubertiming.com/">Huber Timing Home</a>,
 <a class="btn btn-primary btn-lg" href="/results/2017GPTR" role="button" style="margin: 0px 0px 
 <a class="btn btn-primary btn-lg" href="/results/team/2017GPTR" role="button" style="margin: 0px
 <a class="btn btn-primary btn-lg" href="/results/team/2017GPTR10K" role="button" style="margin: 
 <a class="btn btn-primary btn-lg" href="/results/summary/2017GPTR10K" role="button" style="margi
 <a id="individual" name="individual"></a>,
 <a data-url="/results/2017GPTR10K" href="#tabs-1" id="rootTab" style="font-size: 18px">10K Result
 <a href="https://www.hubertiming.com/"><img height="65" src="https://www.hubertiming.com//sites/
 <a href="https://facebook.com/hubertiming/"><img src="https://www.hubertiming.com/results/FB-f-L
 <a class="small" id="bestFeatureEver" style="color:#007bff">Dark Mode</a>]
```

```
# Print out các link liên kết
all_links = soup.find_all("a")
for link in all_links:
```

```
    print(link.get("href"))
```

```
timing@hubertiming.com
https://www.hubertiming.com/
/results/2017GPTR
/results/team/2017GPTR
/results/team/2017GPTR10K
/results/summary/2017GPTR10K
None
#tabs-1
https://www.hubertiming.com/
https://facebook.com/hubertiming/
None
```

```
# Print the first 10 rows for sanity check
rows = soup.find_all('tr')
print(rows[:10])
```

```
[<tr colspan="2">
<b>10K:</b>
</tr>, <tr>
<td>Finishers:</td>
<td>577</td>
</tr>, <tr>
<td>Male:</td>
<td>414</td>
</tr>, <tr>
<td>Female:</td>
<td>163</td>
</tr>, <tr class="header">
<th>Place</th>
<th>Bib</th>
<th>Name</th>
<th>Gender</th>
<th>City</th>
<th>State</th>
<th>Chip Time</th>
<th>Chip Pace</th>
<th>Gun Time</th>
<th>Team</th>
</tr>, <tr data-bib="814">
<td>1</td>
<td>814</td>
<td>

                    JARED WILSON

                </td>
<td>M</td>
<td>TIGARD</td>
<td>OR</td>
<td>36:21</td>
<td>5:51</td>
<td>36:24</td>
<td></td>
</tr>, <tr data-bib="573">
<td>2</td>
<td>573</td>
<td>
```

```
                        NATHAN A SUSTERSIC

                    </td>
     <td>M</td>
     <td>PORTLAND</td>
     <td>OR</td>
     <td>36:42</td>
     <td>5:55</td>
     <td>36:45</td>
     <td>
     <img class="lazy teamThumbs" data-src="/teamLogoThumbnail/logo?teamName=INTEL%20TEAM%20F&amp;ra
                                INTEL TEAM F
                            </td>
     </tr>, <tr data-bib="687">
     <td>3</td>
     <td>687</td>
```

```
# Print out nội dung của các hàng trong bảng
for row in rows:
    row_td = row.find_all('td')
print(row_td)
type(row_td)
```

```
     [<td>577</td>, <td>443</td>, <td>

                        LIBBY B MITCHELL

                    </td>, <td>F</td>, <td>HILLSBORO</td>, <td>OR</td>, <td>1:41:18</td>, <td>16:20</t
     bs4.element.ResultSet
```

```
str_cells = str(row_td)
# print out nội dung text
cleantext = BeautifulSoup(str_cells, "lxml").get_text()
print(cleantext)
```

```
     [577, 443,

                        LIBBY B MITCHELL

                    , F, HILLSBORO, OR, 1:41:18, 16:20, 1:42:10, ]
```

```
import re

def clean_rows(rows_):
    list_rows_ = []
    for row in rows_:
        cells = row.find_all('td')
        str_cells = str(cells)
        str_cells = str_cells.replace("\r", "") # loại bỏ các ký tự "\r"
        str_cells = str_cells.replace("\n", "") # loại bỏ các ký tự "\n"
        # thay thế pattern "<.*?>" thành " "
        clean = re.compile('<.*?>')
        clean2 = (re.sub(clean,'',str_cells))
        list_rows_.append(clean2)
    return list_rows_
```

```
# Đưa list_rows vào pd.DataFrame()
df = pd.DataFrame(list_rows)
df.head(10)
```

	0
0	[]
1	[Finishers:, 577]
2	[Male:, 414]
3	[Female:, 163]
4	[]
5	[1, 814, JARED WILSON ...
6	[2, 573, NATHAN A SUSTERSI...
7	[3, 687, FRANCISCO MAYA ...
8	[4, 623, PAUL MORROW ...
9	[5, 569, DEREK G OSBORNE ...

```
# Tách cột "0" ở vị trí dấu "," thành nhiều cột
df1 = df[0].str.split(',', expand=True)
df1.head(10)
```

	0	1	2	3	4	5	6	7	8	9
0	[]	None	None	None	None	None	None	None	None	None
1	[Finishers:	577]	None	None	None	None	None	None	None	None
2	[Male:	414]	None	None	None	None	None	None	None	None
3	[Female:	163]	None	None	None	None	None	None	None	None
4	[]	None	None	None	None	None	None	None	None	None
5	[1	814	JARED WILSON	M	TIGARD	OR	36:21	5:51	36:24	]
6	[2	573	NATHAN A SUSTERSIC ...	M	PORTLAND	OR	36:42	5:55	36:45	INTEL TEAM F ...
7	[3	687	FRANCISCO MAYA ...	M	PORTLAND	OR	37:44	6:05	37:48	]
8	[4	623	PAUL MORROW	M	BEAVERTON	OR	38:34	6:13	38:37	]
9	[5	569	DEREK G OSBORNE	M	HILLSBORO	OR	39:21	6:20	39:24	INTEL

```
# Bỏ '[' ở cột đầu tiên
df1[0] = df1[0].str.strip('[')
df1.head(10)
```

	0	1	2	3	4	5	6	7	8	9
0	]	None	None	None	None	None	None	None	None	None
1	Finishers:	577]	None	None	None	None	None	None	None	None
2	Male:	414]	None	None	None	None	None	None	None	None
3	Female:	163]	None	None	None	None	None	None	None	None
4	]	None	None	None	None	None	None	None	None	None
5	1	814	JARED WILSON	M	TIGARD	OR	36:21	5:51	36:24	]
6	2	573	NATHAN A SUSTERSIC ...	M	PORTLAND	OR	36:42	5:55	36:45	INTEL TEAM F ...
7	3	687	FRANCISCO MAYA ...	M	PORTLAND	OR	37:44	6:05	37:48	]

```
# Truy cập tiêu đề (header) của bảng
col_labels = soup.find_all('th')


all_header = []
col_str = str(col_labels)
# Trích xuất text giữa html tags cho header của bảng
cleantext2 = BeautifulSoup(col_str, "lxml").get_text()
all_header.append(cleantext2)
print(all_header)
```

```
['[Place, Bib, Name, Gender, City, State, Chip Time, Chip Pace, Gun Time, Team]']
```

```
# Chuyển danh sách trong headers vào một dataframe
df2 = pd.DataFrame(all_header)
df2.head()
```

	0
0	[Place, Bib, Name, Gender, City, State, Chip T...

```
# ta cần tách cột "0" thành nhiều cột
df3 = df2[0].str.split(',', expand=True)
df3.head()
```

	0	1	2	3	4	5	6	7	8	9
0	[Place	Bib	Name	Gender	City	State	Chip Time	Chip Pace	Gun Time	Team]

```
# Ghép 2 dataframe
frames = [df3, df1]

df4 = pd.concat(frames)
df4.head(10)
```

	0	1	2	3	4	5	6	7	8	9
0	[Place	Bib	Name	Gender	City	State	Chip Time	Chip Pace	Gun Time	Team]
0	]	None	None	None	None	None	None	None	None	None
1	Finishers:	577]	None	None	None	None	None	None	None	None
2	Male:	414]	None	None	None	None	None	None	None	None
3	Female:	163]	None	None	None	None	None	None	None	None
4	]	None	None	None	None	None	None	None	None	None
5	1	814	JARED WILSON	M	TIGARD	OR	36:21	5:51	36:24	]
6	2	573	NATHAN A SUSTERSIC ...	M	PORTLAND	OR	36:42	5:55	36:45	INTEL TEAM F ...

```
# Gán hàng đầu tiên thành header
df5 = df4.rename(columns=df4.iloc[0])
df5.head()
```

	[Place	Bib	Name	Gender	City	State	Chip Time	Chip Pace	Gun Time	Team]
0	[Place	Bib	Name	Gender	City	State	Chip Time	Chip Pace	Gun Time	Team]
0	]	None	None	None	None	None	None	None	None	None
1	Finishers:	577]	None	None	None	None	None	None	None	None
2	Male:	414]	None	None	None	None	None	None	None	None
3	Female:	163]	None	None	None	None	None	None	None	None

```
# Kiểm tra dữ liệu có missing values hay không
df5.info()
df5.shape
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 583 entries, 0 to 581
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   [Place      583 non-null    object
 1    Bib        581 non-null    object
 2    Name       578 non-null    object
 3    Gender     578 non-null    object
 4    City       578 non-null    object
 5    State      578 non-null    object
 6    Chip Time  578 non-null    object
 7    Chip Pace  578 non-null    object
 8    Gun Time   578 non-null    object
 9    Team]      578 non-null    object
dtypes: object(10)
memory usage: 50.1+ KB
(583, 10)
```

```
# Bỏ đi các hàng dữ liệu có miss value
```

```
df6 = df5.dropna(axis=0, how='any')


df7 = df6.drop(df6.index[0])
df7.head()
```

	[Place	Bib	Name	Gender	City	State	Chip Time	Chip Pace	Gun Time	Team]
5	1	814	JARED WILSON	M	TIGARD	OR	36:21	5:51	36:24	]
6	2	573	NATHAN A SUSTERSIC ...	M	PORTLAND	OR	36:42	5:55	36:45	INTEL TEAM F ...
7	3	687	FRANCISCO MAYA ...	M	PORTLAND	OR	37:44	6:05	37:48	]
8	4	623	PAUL MORROW	M	BEAVERTON	OR	38:34	6:13	38:37	]