

Bài 1

ngSang_19IT113.pyhtml_tag.py ×search_data.pyCrawler_PhamHoangSang_19IT113.py

regex > html_tag.py > RemoveHtmlTag > clean_html

```
1 import re
2
3
4 class RemoveHtmlTag:
5     # Define Regex
6     def __init__(self):
7         self.TAG_RE = re.compile(r'<[^>]+>')
8
9     def clean_html(self, raw_html):
10
11         cleantext = re.sub(self.TAG_RE, '', raw_html)
12         print(cleantext)
13
14
15 if __name__ == '__main__':
16     html = """<!DOCTYPE html >
17         <html >
18         <head >
19             <title > Page Title </title >
20         </head >
21         <body >
22
23             <h1 > This is a Heading </h1 >
24             <p > This is a paragraph. </p >
25
26         </body >
27         </html >"""
28     RemoveHtmlTag().clean_html(html)
29
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL GITLENS

Python + ~ [] [] ^ ×

Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! <https://aka.ms/PSWindows>

PS D:\E23.1\NaturalLanguageProcessing\Ex> & C:/Users/phsan/AppData/Local/Programs/Python/Python310/python.exe
xe d:/E23.1/NaturalLanguageProcessing/Ex/regex/html_tag.py

Page Title

This is a Heading
This is a paragraph.

PS D:\E23.1\NaturalLanguageProcessing\Ex>

Bài 2

ngSang_19IT113.pyhtml_tag.pysearch_data.pyX Crawler_PhamHoangSang_19IT113.py

regex > search_data.py > ...

```
1 import re
2 import requests
3 from bs4 import BeautifulSoup
4
5 class SearchData:
6     # Define Regex
7     def __init__(self):
8         # \w : Trả về kết quả phù hợp trong đó chuỗi chứa bất kỳ ký tự từ nào (ký tự
9         # . : Any character (except newline character)
10        # + : Một hoặc nhiều lần xuất hiện
11        # - :
12        self.PHONE_RE = re.compile(r'\d{3}-\d{3}-\d{4}')
13        self.EMAIL_RE = re.compile(r'[\w.]+@[ \w]+\.[\w.]+')
14        self.HTTP_RE = re.compile(r'https?:\/\/[ \w]+\.[\w.]+')
15
16    def clean_html(self, raw_html):
17
18        phone = re.findall(self.PHONE_RE, raw_html)
19        email = re.findall(self.EMAIL_RE, raw_html)
20        http = re.findall(self.HTTP_RE, raw_html)
21        print(phone)
22        print(email)
23        print(http)
24
25    if __name__ == '__main__':
26        html = """<!DOCTYPE html >
27        <html > <head > <title > 289-544-2345 Page Title < /title >
28        </head > <body >daota..o@vku..udn.vn
29        popop@coco.com <h1 > This is a Heading < /h1 >
30        <p > This is a paragraph. < /p >
31        023-665-5268 https://elearning.vku.udn.vn/mod/page/view.php?id=34760
32        </body >
33        </html >"""
34        SearchData().clean_html(html)
35
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL GITLENS

Python + - [] [] ^ X

Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! <https://aka.ms/PSWindows>

PS D:\E23.1\NaturalLanguageProcessing\Ex> & C:\Users\phsan\AppData\Local\Programs\Python\Python310\python.exe
xe d:/E23.1/NaturalLanguageProcessing/Ex/regex/search_data.py
['289-544-2345', '023-665-5268']
['daota..o@vku..udn.vn', 'popop@coco.com']
['https://elearning.vku.udn.vn']
PS D:\E23.1\NaturalLanguageProcessing\Ex>