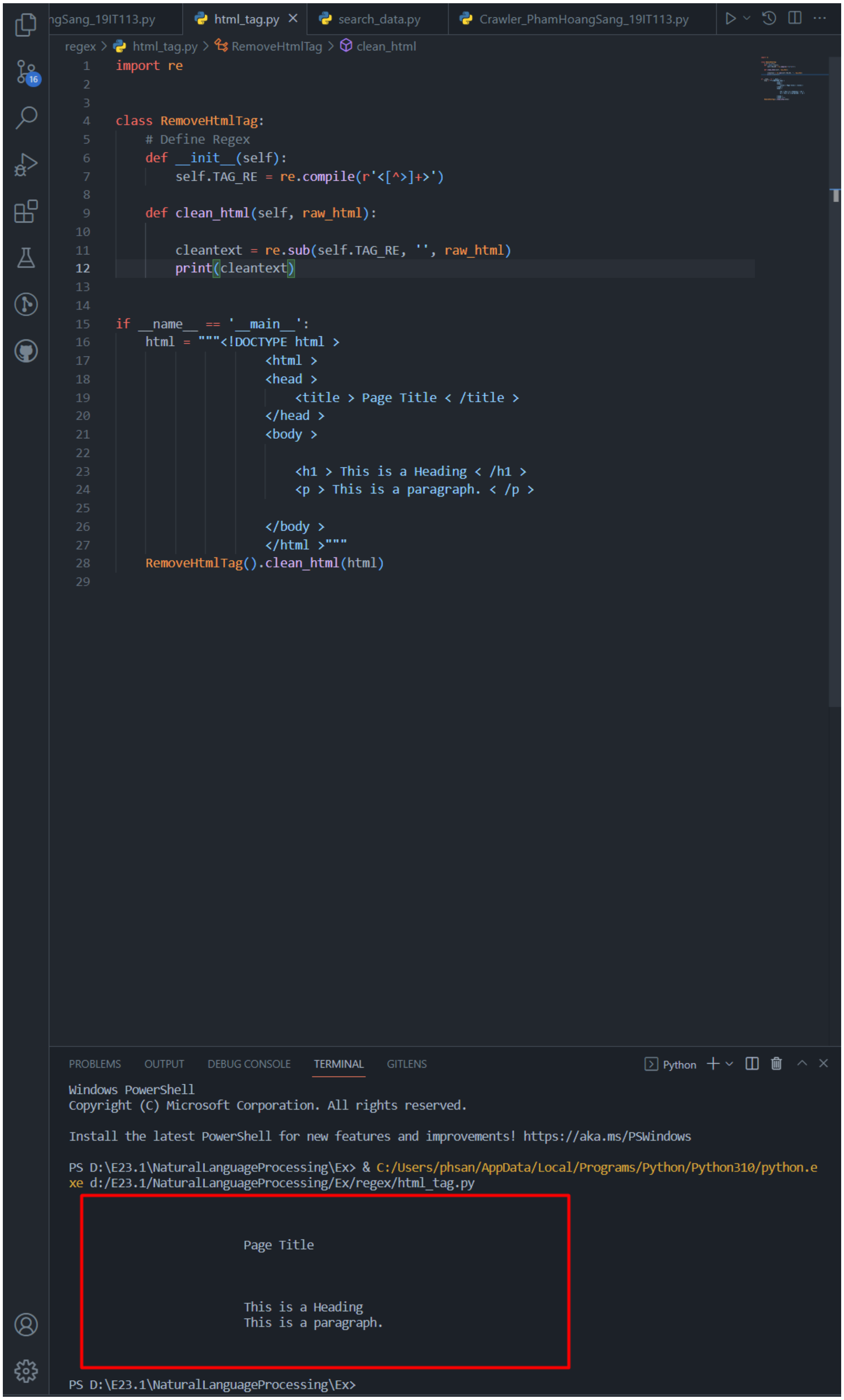
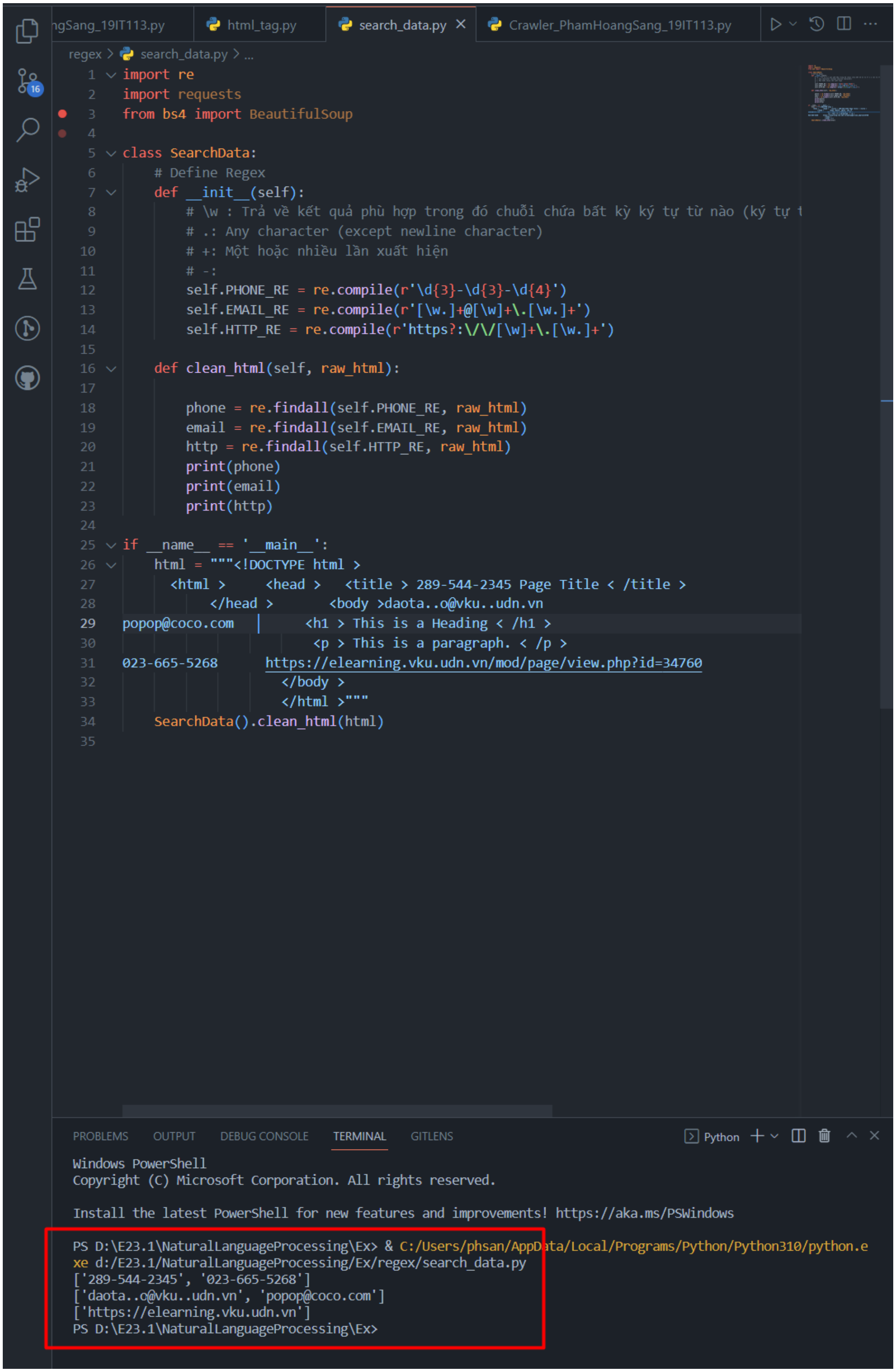
Bài 1



```python
import re


class RemoveHtmlTag:
    # Define Regex
    def __init__(self):
        self.TAG_RE = re.compile(r'<[^>]+>')

    def clean_html(self, raw_html):

        cleantext = re.sub(self.TAG_RE, '', raw_html)
        print(cleantext)


if __name__ == '__main__':
    html = """<!DOCTYPE html >
                    <html >
                    <head >
                        <title > Page Title < /title >
                    </head >
                    <body >

                        <h1 > This is a Heading < /h1 >
                        <p > This is a paragraph. < /p >

                    </body >
                    </html >"""
    RemoveHtmlTag().clean_html(html)

```

```
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.e
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



```python
import re
import requests
from bs4 import BeautifulSoup

class SearchData:
    # Define Regex
    def __init__(self):
        # \w : Trả về kết quả phù hợp trong đó chuỗi chứa bất kỳ ký tự từ nào (ký tự t
        # .: Any character (except newline character)
        # +: Một hoặc nhiều lần xuất hiện
        # -:
        self.PHONE_RE = re.compile(r'\d{3}-\d{3}-\d{4}')
        self.EMAIL_RE = re.compile(r'[\w.]+@[\w]+\.[\w.]+')
        self.HTTP_RE = re.compile(r'https?:\/\/[\w]+\.[\w.]+')

    def clean_html(self, raw_html):

        phone = re.findall(self.PHONE_RE, raw_html)
        email = re.findall(self.EMAIL_RE, raw_html)
        http = re.findall(self.HTTP_RE, raw_html)
        print(phone)
        print(email)
        print(http)

if __name__ == '__main__':
    html = """<!DOCTYPE html >
      <html >     <head >   <title > 289-544-2345 Page Title < /title >
           </head >       <body >daota..o@vku..udn.vn
popop@coco.com         <h1 > This is a Heading < /h1 >
                        <p > This is a paragraph. < /p >
023-665-5268      https://elearning.vku.udn.vn/mod/page/view.php?id=34760
                    </body >
                    </html >"""
    SearchData().clean_html(html)

```

```
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.e
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>
```