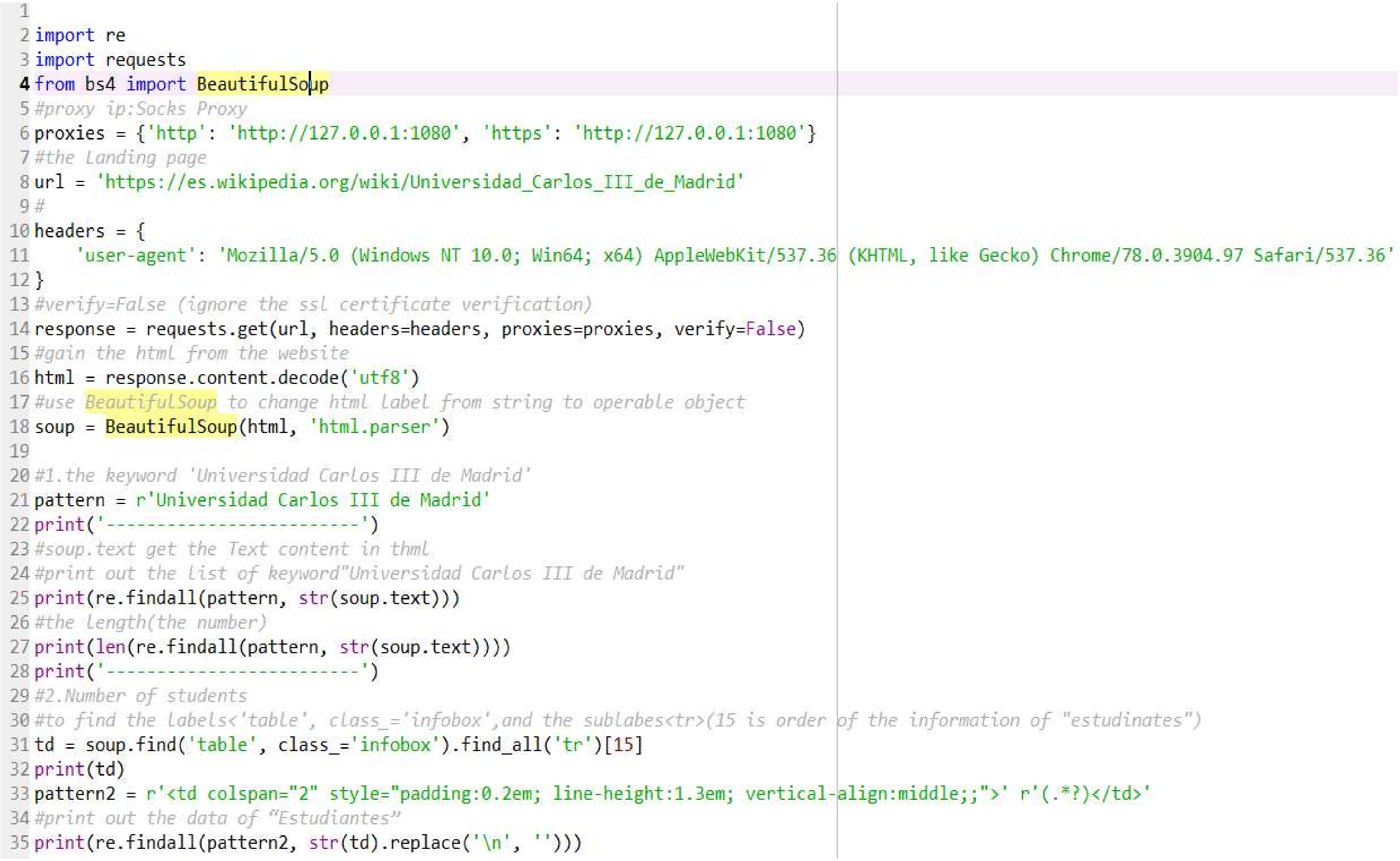
Exercise1: Since my ability in python programming is not enough, I could not crawl the whole Wikipedia, here crawling only one page of the keyword in Spanish in Wikipedia.



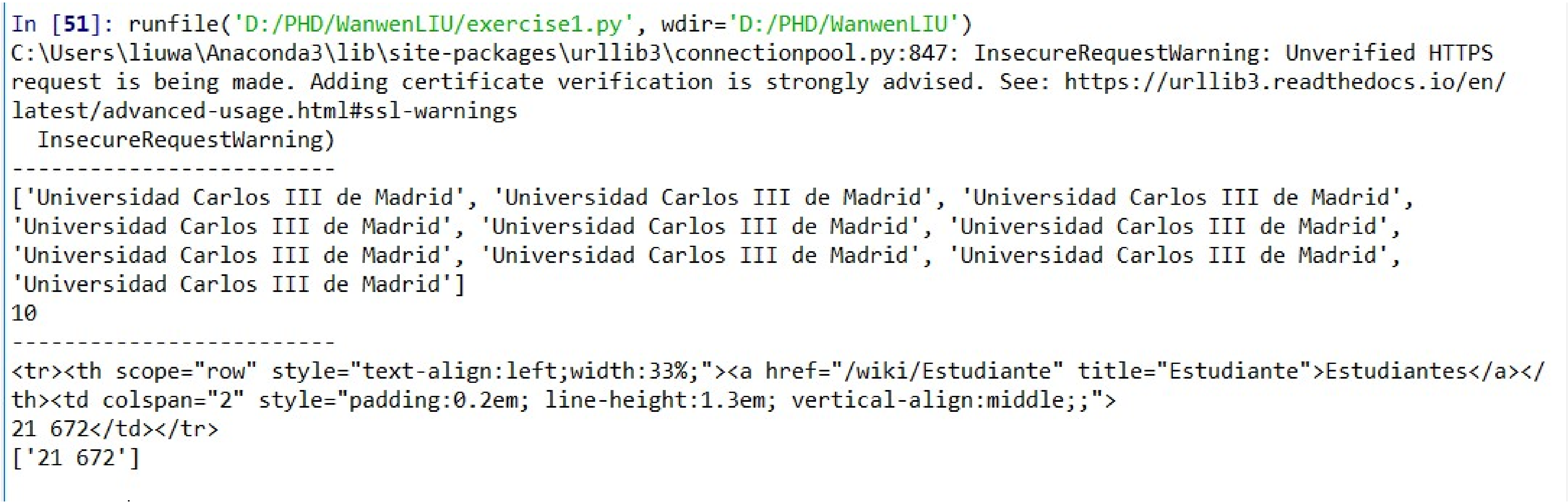
```python

import re
import requests
from bs4 import BeautifulSoup
#proxy ip:Socks Proxy
proxies = {'http': 'http://127.0.0.1:1080', 'https': 'http://127.0.0.1:1080'}
#the Landing page
url = 'https://es.wikipedia.org/wiki/Universidad_Carlos_III_de_Madrid'
#
headers = {
    'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.97 Safari/537.36'
}
#verify=False (ignore the ssl certificate verification)
response = requests.get(url, headers=headers, proxies=proxies, verify=False)
#gain the html from the website
html = response.content.decode('utf8')
#use BeautifulSoup to change html label from string to operable object
soup = BeautifulSoup(html, 'html.parser')

#1.the keyword 'Universidad Carlos III de Madrid'
pattern = r'Universidad Carlos III de Madrid'
print('-------------------------')
#soup.text get the Text content in thml
#print out the list of keyword"Universidad Carlos III de Madrid"
print(re.findall(pattern, str(soup.text)))
#the length(the number)
print(len(re.findall(pattern, str(soup.text))))
print('-------------------------')
#2.Number of students
#to find the labels<'table', class_='infobox',and the sublabes<tr>(15 is order of the information of "estudinates")
td = soup.find('table', class_='infobox').find_all('tr')[15]
print(td)
pattern2 = r'<td colspan="2" style="padding:0.2em; line-height:1.3em; vertical-align:middle;;">' r'(.*?)</td>'
#print out the data of “Estudiantes”
print(re.findall(pattern2, str(td).replace('\n', '')))
```

Crawling the specified page, firstly use BeautifulSoup to gain the text from html, help to operate the search the keyword in the text and count its number of references.

Secondly, use findall() to help to find all the matched string, and return in a list

Thirdly,count the keyword in the text and find the specified information or data.

The result:



```
In [51]: runfile('D:/PHD/WanwenLIU/exercise1.py', wdir='D:/PHD/WanwenLIU')
C:\Users\liuwa\Anaconda3\lib\site-packages\urllib3\connectionpool.py:847: InsecureRequestWarning: Unverified HTTPS
request is being made. Adding certificate verification is strongly advised. See: https://urllib3.readthedocs.io/en/
latest/advanced-usage.html#ssl-warnings
  InsecureRequestWarning)
-------------------------
['Universidad Carlos III de Madrid', 'Universidad Carlos III de Madrid', 'Universidad Carlos III de Madrid',
'Universidad Carlos III de Madrid', 'Universidad Carlos III de Madrid', 'Universidad Carlos III de Madrid',
'Universidad Carlos III de Madrid', 'Universidad Carlos III de Madrid', 'Universidad Carlos III de Madrid',
'Universidad Carlos III de Madrid']
10
-------------------------
<tr><th scope="row" style="text-align:left;width:33%;"><a href="/wiki/Estudiante" title="Estudiante">Estudiantes</a></
th><td colspan="2" style="padding:0.2em; line-height:1.3em; vertical-align:middle;;">
21 672</td></tr>
['21 672']
```

1. the number of references in the Spanish entry page in Wikipedia : https://es.wikipedia.org/wiki/Universidad_Carlos_III_de_Madrid :10
2. the number of students:21672