

```
In [1]: 1 import requests          # module use for send HTTP requests
        2 from bs4 import BeautifulSoup # parsing HTML documents
        3 import json
```

executed in 253ms, finished 19:34:48 2021-03-14

## checking for web scrape for chapter 1

## GET request

```
In [2]: 1 # get() method sends a GET request to the specified url and returns a response
        2 response = requests.get("https://novelfull.com/warriors-promise/chapter-1.html")
        3 #page
        4 response.status_code
```

executed in 1.23s. finished 19:34:50 2021-03-14

Out[2]: 200

**display respose content (in html form)**

```
In [3]: 1 response.content
```

executed in 26ms. finished 19:34:50 2021-03-14

```
Out[3]: b'\r\n<!DOCTYPE html><html lang="en-US"><head><meta http-equiv="Content-Type"
content="text/html; charset=UTF-8"/><meta http-equiv="X-UA-Compatible" conten
t="IE=edge"><meta name="viewport">\n                content="width=device-width,
height=device-height, initial-scale=1.0, user-scalable=0, minimum-scale=1.0,
maximum-scale=1.0"><meta name="csrf-param" content="_csrf"><meta name="csrf-t
oken" content="NjM2NGQzZjBmNDk1YjZhYjlkY2Y4ZDNiNmMDEAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAA=="><!-- TODO: Meta --><meta name="title" content="Re
ad Warrior\xe2\x80\x99s Promise - Chapter 1 online free - Novel Full"><meta n
ame="description" content="Read Warrior\xe2\x80\x99s Promise - Chapter 1 onli
ne free from your Mobile, Table, PC... Novel Updates Daily "><meta name="imag
e" content="https://novelfull.com/uploads/thumbs/warriors-promise-568d11545d-
2239c49aee6b961904acf173b7e4602a.jpg"><meta name="keywords" content="Warrior
\xe2\x80\x99s Promise - Chapter 1"><title>Read Warrior\xe2\x80\x99s Promise -
Chapter 1 online free - Novel Full</title><link rel="preload" href="/web/font
s/RobotoCondensed-Regular.woff" as="font" crossorigin="anonymous" /><link rel
="preload" href="/web/fonts/glyphicons-halflings-regular.woff2" as="font" cro
ssorigin="anonymous" /><link rel="shortcut icon" href="/web/images/favicon.ic
o" type="image/x-icon"><link rel="icon" href="/web/images/favicon.ico" type
="image/x-icon"><!--[if lt IE 9]><link rel="stylesheet" href="/web/css/iefix.
css"><script defer src="/web/js/html5shiv.min.js"></script><script defer src
```

## parsing the html content

```
In [4]: 1 soup = BeautifulSoup(response.content, 'html.parser')
```

executed in 168ms. finished 19:34:50 2021-03-14

## count all paragraphs and print it one by one

```
In [5]: 1 count_p = len(soup.find_all('p'))
        2 i = 0
        3 while i < count_p:
        4     print(soup.find_all('p')[i].get_text())
        5     i+=1
```

executed in 122ms, finished 19:34:50 2021-03-14

Chapter 1: The Awakening of Martial Souls

Translator: Transn Editor: Transn

The story started in Sunnywood City, Skymoon Country.

In the mansion of the Sus, a ceremony of Martial Soul awakening was being held in a huge martial practice field.

An altar stood in the center of the martial practice field.

Many elders and disciples of the Sus were clustered in the field, watching the ceremony.

“Rank 2 Human Class Martial Soul!”

“Rank 3 Human Class Martial Soul!”

“Rank 3 Human Class Martial Soul!”

The elder hosting the ceremony continuously announced the ranks of the Su disciples’ awakened Martial Souls.

The level of the Martial Soul indicated the cultivating talent of martial artists. Both Rank 2 and Rank 3 Human Class Martial Soul were common Martial Souls and were considered mediocre.

Many people in the martial practice field fixed their eyes on two young men who were standing in the front row.

~~~~~

```
In [6]: 1 all_p_content = soup.findAll('p') # find all paragraphs
        2 paragraphs = [str(x.text) for x in all_p_content] # list of all paragraphs
```

executed in 12ms, finished 19:34:50 2021-03-14

## separate chapter no., chapter name and content inside chapter

```
In [7]: 1 idx = paragraphs[1].find(':') # give index of ':' in 2nd element of list
        2 chap_no = paragraphs[1][:idx] # chapter no.
        3 chap_name = paragraphs[1][idx+1:] # chapter name
        4 contents = paragraphs[2:-1] # contents inside chapter
```

executed in 76ms, finished 19:34:50 2021-03-14

## Json file creation

```
In [8]: 1 out_file = open("test1.json", "w") # creating json file object as write mode
        2
        3 # dumping all the data of response as {chapter_no. : {chapter_name : chapter_content}}
        4 json_text = json.dump({chap_no : {chap_name : contents}}, out_file, indent=4)
        5
        6 out_file.close() # closing the writing mode
```

executed in 59ms, finished 19:34:50 2021-03-14

# Apply all steps same as above over all 30 chapters iteratively

## Method 1: separate json file for each chapter ( till 30 chapters)

```
In [9]: 1 for index in range(1, 31):
2     response = requests.get("https://novelfull.com/warriors-promise/chapter")
3     soup = BeautifulSoup(response.content, 'html.parser')
4
5     all_p_content = soup.findAll('p')
6     paragraphs = [str(x.text) for x in all_p_content]
7     #for i in tables:
8     #     print(i)
9
10    idx = paragraphs[1].find(':')
11    chap_no = paragraphs[1][:idx]
12    chap_name = paragraphs[1][idx+1:]
13    contents = paragraphs[2:-1]
14
15    json_text = json.dumps({chap_no : {chap_name : contents}})
16    with open('Chapter' + str(index) + '.json', 'w') as json_file:
17        json_file.write(json_text)
```

executed in 30.4s, finished 19:35:21 2021-03-14

## Method 2: all chapters in a single json file

```
In [10]: 1 out_file = open("Novel_chapter_1_to_30.json", "w") # create json file
2
3 for index in range(1, 31):
4     response = requests.get("https://novelfull.com/warriors-promise/chapter")
5     soup = BeautifulSoup(response.content, 'html.parser')
6
7     all_p_content = soup.findAll('p')
8     paragraphs = [str(x.text) for x in all_p_content]
9
10    idx = paragraphs[1].find(':')
11    chap_no = paragraphs[1][:idx]
12    chap_name = paragraphs[1][idx+1:]
13    contents = paragraphs[2:-1]
14
15    json.dump({chap_no : {chap_name : contents}}, out_file, indent = 4, sort_keys=True)
16
17    out_file.close()
```

executed in 28.7s, finished 19:35:49 2021-03-14

In [ ]:

1