

세상의 속도를
따라잡고 싶다면

**Do
it!**

파이썬 생활 프로그래밍



06

Web Crawling으로 정보 모으기

06-1 Web Crawling 알아보기

06-2 Web Crawling 준비하기

06-1 Web Crawling 알아보기

Web Crawling이란 ?

- Web의 정보를 자동으로 수집해주는 Program

HTML 몰라도 Web Crawling을 할 수 있을까 ?

- Web Crawling에 필수인 HTML 요소를 찾을 수만 있다면 가능

06-2 Web Crawling 준비하기

이런 상황이라면?

- Web site에서 명언을 수집하기
- 명언 정리 사이트 Quotes to scrape.com 활용 (<https://quotes.toscrape.com>)



06-2 Web Crawling 준비하기

Quotes to scrape Page 살펴보기

- 'life' tag 선택



06-2 Web Crawling 준비하기

Quotes to scrape Page 살펴보기

- 명언 주제부터 그것을 말한 인물 정보도 탐색 가능

Quotes to Scrape

Login

Viewing tag: life

"There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle."

by Albert Einstein (about)

Tags: [inspirational](#) [life](#) [love](#) [miracle](#) [miracles](#)

"It is better to be hated for what you are than to be loved for what you are not."

by André Gide (about)

Tags: [life](#) [love](#)

"This life is what you make it. No matter what, you're going to mess up sometimes, it's a universal truth. But the good part is you get to decide how you're going to mess it up. Girls will be your friends - they'll act like it anyway. But just remember, some come, some go. The ones that stay with you through everything - they're your true best friends. Don't let go of them. Also remember, sisters make the best friends in the world. As for lovers, well, they'll come and go too. And baby, I hate to say it, most of them - actually pretty much all of them are going to break your heart, but you can't give up

Top Ten tags

[love](#)
[inspirational](#)
[life](#)
[humor](#)
[books](#)
[reading](#)
[friendship](#)
[travel](#)
[faith](#)
[quote](#)

Quotes to Scrape

Login

Albert Einstein

Born: March 14, 1879 in Ulm, Germany

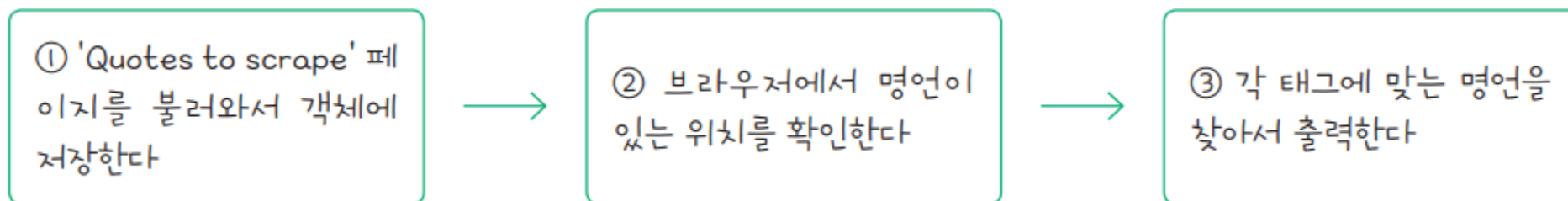
Description:

In 1879, Albert Einstein was born in Ulm, Germany. He completed his Ph.D. at the University of Zurich by 1909. His 1905 paper explaining the photoelectric effect, the basis of electronics, earned him the Nobel Prize in 1921. His first paper on Special Relativity Theory, also published in 1905, changed the world. After the rise of the Nazi party, Einstein made Princeton his permanent home, becoming a U.S. citizen in 1940. Einstein, a pacifist during World War I, stayed a firm proponent of social justice and responsibility. He chaired the Emergency Committee of Atomic Scientists, which organized to alert the public to the dangers of atomic warfare. At a symposium, he advised: "In their struggle for the ethical good, teachers of religion must have the stature to give up the doctrine of a personal God, that is, give up that source of fear and hope which in the past placed such vast power in the hands of priests. In their labors they will have to avail themselves of those forces which are capable of cultivating the Good, the True, and the Beautiful in humanity itself. This is, to be sure a more difficult but an incomparably more worthy task" ("Science, Philosophy and Religion, A Symposium," published by the Conference on Science, Philosophy and Religion in their Relation to the Democratic Way of Life, Inc., New York, 1941). In a letter to philosopher Eric Gutkind, dated Jan. 3, 1954, Einstein stated: "The word god is for me nothing more than the expression and product of human weaknesses, the Bible a collection of honorable, but still primitive legends which are nevertheless pretty childish. No interpretation no matter how subtle can (for me) change this;" (The Guardian, "Childish superstition: Einstein's letter makes view of religion relatively clear," by James Randerson, May 13, 2008). D. 1955. While best known for his mass-energy equivalence formula $E = mc^2$ (which has been dubbed "the world's most famous equation"), he received the 1921 Nobel Prize in Physics "for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect". The latter was pivotal in establishing quantum theory. Einstein thought that Newtonian mechanics was no longer enough to reconcile the laws of classical mechanics with the laws of the electromagnetic field. This led to the

06-2 Web Crawling 준비하기

Quotes to scrape Page 살펴보기

- 명언 crawling하기 흐름

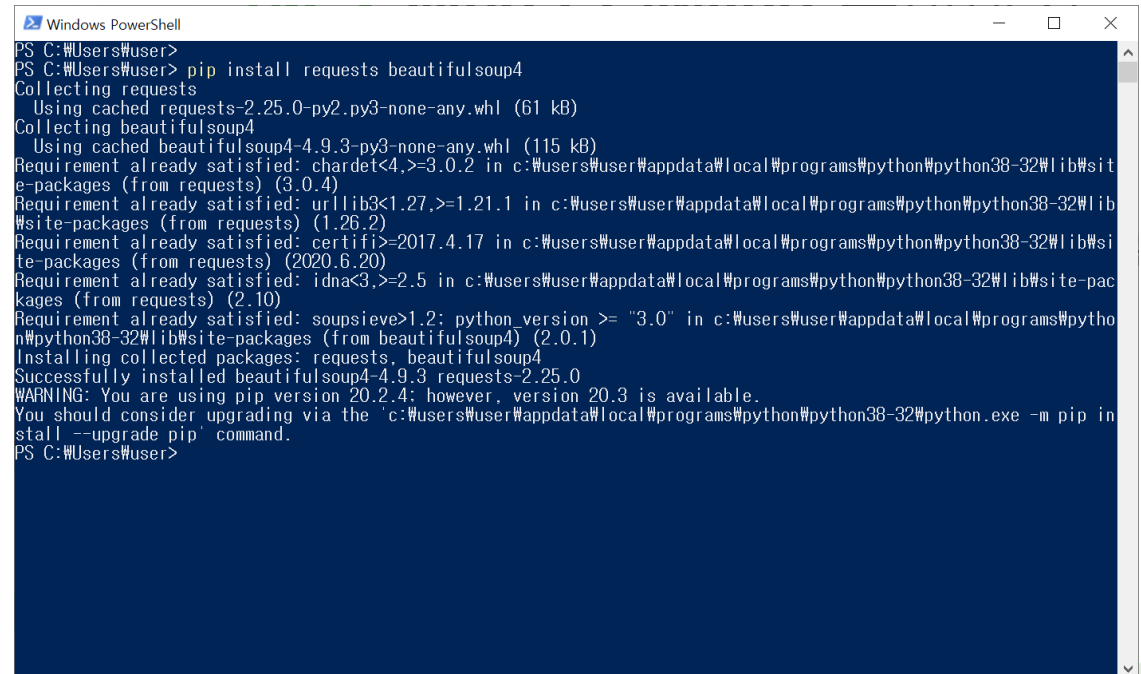


06-2 Web Crawling 준비하기

request , BeautifulSoup 설치하기

- HTML 문서와 XML 문서를 쉽게 이용할 수 있게 해주는 모듈
- pip를 통하여 설치

>> pip install requests beautifulsoup4



```
Windows PowerShell
PS C:\Users\User>
PS C:\Users\User> pip install requests beautifulsoup4
Collecting requests
  Using cached requests-2.25.0-py2.py3-none-any.whl (61 kB)
Collecting beautifulsoup4
  Using cached beautifulsoup4-4.9.3-py3-none-any.whl (115 kB)
Requirement already satisfied: chardet<4,>=3.0.2 in c:\Users\User\AppData\Local\Programs\Python\Python38-32\lib\site-packages (from requests) (3.0.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\Users\User\AppData\Local\Programs\Python\Python38-32\lib\site-packages (from requests) (1.26.2)
Requirement already satisfied: certifi>=2017.4.17 in c:\Users\User\AppData\Local\Programs\Python\Python38-32\lib\site-packages (from requests) (2020.6.20)
Requirement already satisfied: idna<3,>=2.5 in c:\Users\User\AppData\Local\Programs\Python\Python38-32\lib\site-packages (from requests) (2.10)
Requirement already satisfied: soupsieve>1.2; python_version >= "3.0" in c:\Users\User\AppData\Local\Programs\Python\Python38-32\lib\site-packages (from beautifulsoup4) (2.0.1)
Installing collected packages: requests, beautifulsoup4
Successfully installed beautifulsoup4-4.9.3 requests-2.25.0
WARNING: You are using pip version 20.2.4; however, version 20.3 is available.
You should consider upgrading via the 'c:\Users\User\AppData\Local\Programs\Python\Python38-32\python.exe -m pip install --upgrade pip' command.
PS C:\Users\User>
```


06-2 Web Crawling 준비하기

기본 모듈 import 하기

- os, re, usecsv import

```
>>> import os, re, usecsv
```

- request import Web에 요청 보낼 때 사용

```
>>> import requests
```

06-2 Web Crawling 준비하기

기본 모듈 import 하기

- urllib.request import

```
>>> import urllib.request as ur
```

- BeautifulSoup import

```
>>> from bs4 import BeautifulSoup as bs
```

06-2 Web Crawling 준비하기

Web 문서 자료를 가져와 가공하기

- urlopen으로 Web사이트 정보 가져오기

```
>>> url = 'http://quotes.toscrape.com/'
```

- url 주소에 정보 요청

```
>>> html = ur.urlopen(url)
```

06-2 Web Crawling 준비하기

Web 문서 자료를 가져와 가공하기

- read() 를 통하여 받은 data 확인

```
>>> html.read()[:100]
b'<!DOCTYPE html>\n<html lang="en">\n<head>\n\t<meta charset="UTF-8">\n\t<title>Quotes
to Scrape</title>\n'
```

06-2 Web Crawling 준비하기

Web 문서 자료를 가져와 가공하기

- BeautifulSoup 로 parsing하기 쉬운 형태로 변환

뷰티풀수프 사용법

```
bs(html.read(), 'html.parser')
```

06-2 Web Crawling 준비하기

Web 문서 자료를 가져와 가공하기

- BeautifulSoup 로 parsing하기 쉬운 형태로 변환

```
>>> html = ur.urlopen(url)
>>> soup = bs(html.read(), 'html.parser')
```

06-2 Web Crawling 준비하기

한 줄로 모두 실행하기

- 이전에 진행했던 것들을 한 줄로 실행

```
>>> soup = bs(ur.urlopen('http://quotes.toscrape.com/').read(), 'html.parser')
```

06-2 Web Crawling 준비하기

특정 Tag에서 텍스트만 추출하기

- HTML의 구조 살펴보기

```
<HTML>
<head>
    <title> 페이지 제목 </title>
</head>
<body>
    <h1> 글 제목 </h1>
    <p> 글 본문 </p>
</body>
</HTML>
```


06-2 Web Crawling 준비하기

특정 Tag에서 텍스트만 추출하기

- find_all로 원하는 Tag만 모으기
 - 텍스트는 Tag로 둘러싸여 있음

```
...(생략)...  
<div class="quote" itemscope="" itemtype="http://schema.org/CreativeWork">  
  <span class="text" itemprop="text">"The world as we have created it is a process of our  
  thinking. It cannot be changed without changing our thinking."</span>  
  <span>by <small class="author" itemprop="author">Albert Einstein</small>  
  <a href="/author/Albert-Einstein">(about)</a>  
</span>  
...(생략)...
```

06-2 Web Crawling 준비하기

특정 Tag에서 텍스트만 추출하기

- find_all로 원하는 Tag만 모으기
 - find_all로 span Tag만 출력

```
>>> soup.find_all('span')
```

첫 번째 원소입니다

```
[<span class="text" itemprop="text">"The world as we have created it is a process of  
our thinking. It cannot be changed without changing our thinking."</span>, <span>by  
<small class="author" itemprop="author">Albert Einstein</small>  
<a href="/author/Albert-Einstein">(about)</a>  
</span>, <span class="text" itemprop="text">"It is our choices, Harry, that show what
```

06-2 Web Crawling 준비하기

특정 Tag에서 텍스트만 추출하기

- Tag에서 텍스트만 출력하기
 - span tag만 모아 quote에 저장

```
>>> quote = soup.find_all('span')
```

06-2 Web Crawling 준비하기

특정 Tag에서 텍스트만 추출하기

- Tag에서 텍스트만 출력하기
 - .text를 붙여 텍스트만 출력

```
>>> quote[0].text  
'"The world as we have created it is a process of our thinking. It cannot be changed  
without changing our thinking.'"
```

06-2 Web Crawling 준비하기

특정 Tag에서 텍스트만 추출하기

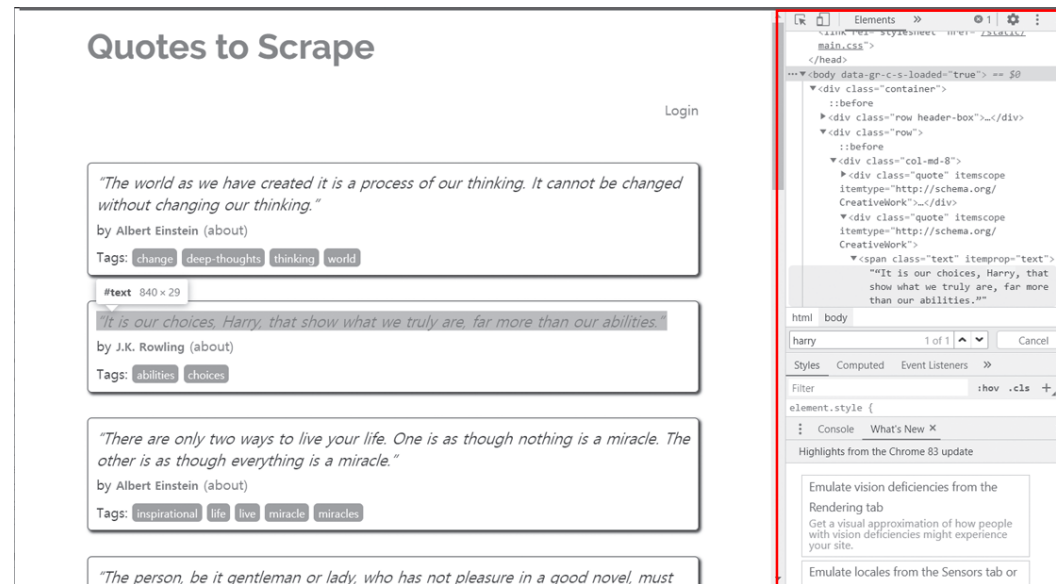
- Tag에서 텍스트만 출력하기
 - 모든 Tag의 모든 텍스트 추출

```
>>> for i in quote:
    i.text
# quote 리스트에서 텍스트만 추출합니다
'"The world as we have created it is a process of our thinking. It cannot be changed
without changing our thinking."'
'by Albert Einstein\n(about)\n'
'"It is our choices, Harry, that show what we truly are, far more than our abilities."'
'by J.K. Rowling\n(about)\n'
'"There are only two ways to live your life. One is as though nothing is a miracle.
The other is as though everything is a miracle."
... (생략)...
'\ntruth\n'
'\nsmile\n'
' r'
```

06-2 Web Crawling 준비하기

Web browser에서 특정 Tag 찾아 명언 출력하기

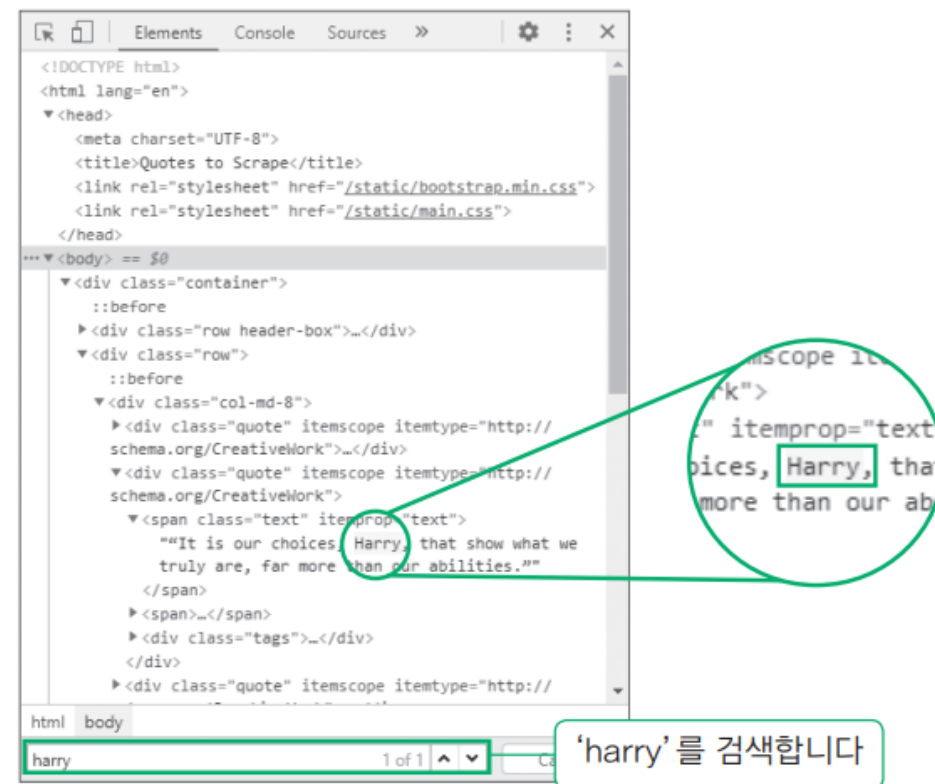
- 정보 위치 확인하기 위해 browser 콘솔 사용



06-2 Web Crawling 준비하기

Web browser에서 특정 Tag 찾아 명언 출력하기

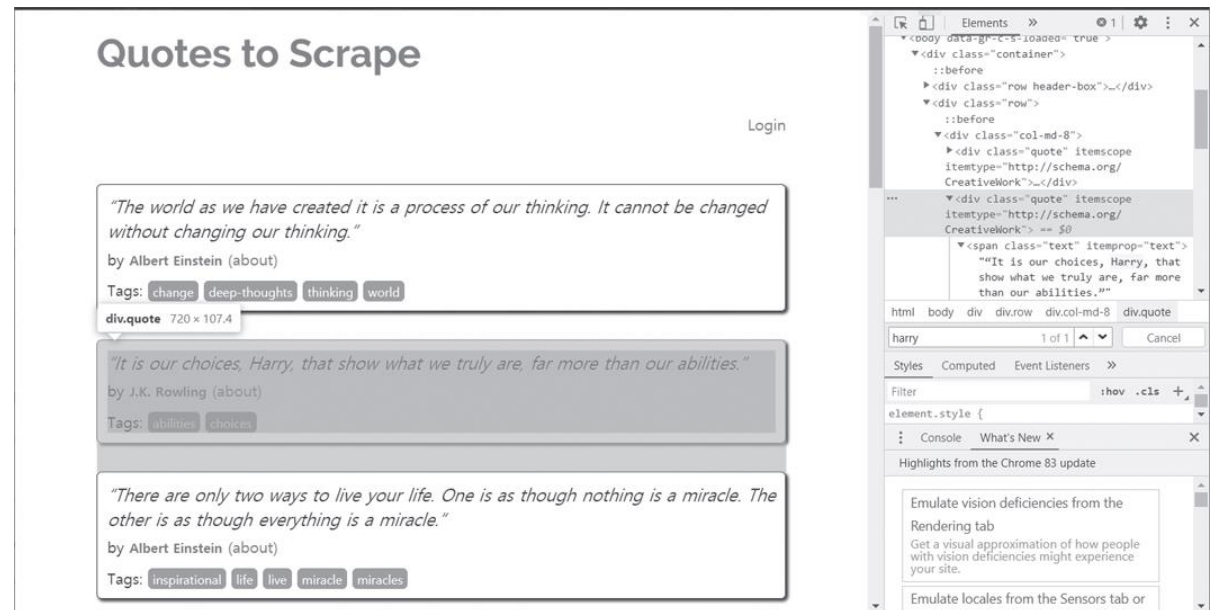
- 명언 일부를 입력하여 검색 가능



06-2 Web Crawling 준비하기

Web browser에서 특정 Tag 찾아 명언 출력하기

- 검색된 Tag의 렌더링 위치 표시



06-2 Web Crawling 준비하기

Web browser에서 특정 Tag 찾아 명언 출력하기

- div Tag 안에 정의된 특정 클래스 찾아가기

```
soup.find_all('div', {"class" : "quote"})
```

06-2 Web Crawling 준비하기

Web browser에서 특정 Tag 찾아 명언 출력하기

- div Tag 안에 정의된 특정 클래스 찾아가기

```
>>> soup.find_all('div', {"class": "quote"})[0]
<div class="quote" itemscope="" itemtype="http://schema.org/CreativeWork">
  <span class="text" itemprop="text">"The world as we have created it is a process of our
  thinking. It cannot be changed without changing our thinking."</span>
  <span>by <small class="author" itemprop="author">Albert Einstein</small>
  <a href="/author/Albert-Einstein">(about)</a>
</span>
<div class="tags">
  Tags:
    <meta class="keywords" content="change,deep-thoughts,thinking,world"
  itemprop="keywords"/>
  <a class="tag" href="/tag/change/page/1/">change</a>
  <a class="tag" href="/tag/deep-thoughts/page/1/">deep-thoughts</a>
  <a class="tag" href="/tag/thinking/page/1/">thinking</a>
  <a class="tag" href="/tag/world/page/1/">world</a>
</div>
</div>
```

06-2 Web Crawling 준비하기

Web browser에서 특정 Tag 찾아 명언 출력하기

- Tag에서 텍스트만 추출

```
>>> soup.find_all('div', {"class" : "quote"})[0].text
'\n"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."\nby Albert Einstein\n(about)\n\n\nTags:\n\n\nchange\ndeep-thoughts\nthinking\nworld\n\n'
```

06-2 Web Crawling 준비하기

Web browser에서 특정 Tag 찾아 명언 출력하기

- 반복문을 통해 모든 명언 출력

```
>>> for i in soup.find_all('div',{'class':"quote"}):
    print(i.text)

"The world as we have created it is a process of our thinking. It cannot be changed
without changing our thinking."
by Albert Einstein
(about)

...(생략)...

"A day without sunshine is like, you know, night."
by Steve Martin
(about)

Tags:

humor
obvious
simile
```

```
web_crawling_1.py
File Edit Format Run Options Window Help
1 #
2 # 참조 : http://hleecaster.com/python-web-crawling-with-beautifulsoup/
3 # "11가지 프로젝트로 시작하는 생활프로그래밍", 이창현 저, 이지스퍼블리싱, 2020
4 #
5
6 D = True
7 #D = False
8
9 import os, re
10 import usecsv
11 import requests
12 import urllib.request as ur
13
14 from bs4 import BeautifulSoup as bs
15
16 try:
17     ...
18     # urlopen으로 Web사이트 정보 가져오기
19     url = "https://quotes.toscrape.com"
20
21     html = ur.urlopen(url)
22     if D:
23         print(">> html : ", html)
24
25     # read() 를 통하여 받은 데이터 확인
26     if D:
27         #line = html.read()[:100]
28         line = html.read()
29         print(">> line : ",line)
30
31     # BeautifulSoup로 parsing하기 쉬운 형태로 변환
32     soup = bs(html.read(), 'html.parser')
33     ...
34
35     url = "https://quotes.toscrape.com"
36     webpage = requests.get(url)
37     if D:
38         print("1) >> webpage : ", webpage)
39
40     soup = bs(webpage.content, 'html.parser')
41
42     if D:
43         print("2) >> soup : ", soup)
44
45     # 특정 tag에서 텍스트만 추출하기
46     # tag에서 텍스트만 출력하기
47     # span tag만 모아 quote에 저장
48
49     quote = soup.find_all('span')
50
```

```
51 '''
52 if D:
53     print("3) >> span tag의 첫번째 text를 출력")
54     print(quote[0].text)
55 '''
56 if D:
57     print("3) >> span tag의 text를 모두 출력")
58
59 for i in quote:
60     print(i.text)
61
62 quote_1 = soup.find_all('div', {"class" : "quote"})
63 if D:
64     print("4) >> div tag 내 class = quote 인 tag를 모두 리턴")
65     print("5) >> class = quote 인 tag의 text 값 출력")
66
67 for i in quote_1:
68     print(i.text)
69
70 except Exception as e:
71     print("Warning : ", e)
72
```

```
Python 3.8.5 Shell
File Edit Shell Debug Options Window Help
Python 3.8.5 (tags/v3.8.5:580fbb0, Jul 20 2020, 15:43:08) [MSC v.1926 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: C:\W과소사\W과소사-강의예제\Wweb_crawling\Wweb_crawling_1.py ==
=====
1) >> webpage : <Response [200]>

2) >> soup : Squeezed text (234 lines).

3) >> span tag의 text를 모두 출력
"The world as we have created it is a process of our thinking. It cannot be changed without chan
ging our thinking."
by Albert Einstein
(about)

"It is our choices, Harry, that show what we truly are, far more than our abilities."
by J.K. Rowling
(about)

"There are only two ways to live your life. One is as though nothing is a miracle. The other is as t
hough everything is a miracle."
by Albert Einstein
(about)

"The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably s
tupid."
by Jane Austen
(about)

"Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolu
tely boring."
by Marilyn Monroe
(about)

"Try not to become a man of success. Rather become a man of value."
by Albert Einstein
(about)

"It is better to be hated for what you are than to be loved for what you are not."
by André Gide
(about)

"I have not failed. I've just found 10,000 ways that won't work."
by Thomas A. Edison
(about)

"A woman is like a tea bag; you never know how strong it is until it's in hot water."
by Eleanor Roosevelt
(about)

"A day without sunshine is like, you know, night."
by Steve Martin
(about)
Ln: 19 Col: 18
```

감사합니다

