

# 파이썬을 활용한 실전 웹크롤링 CAMP

2018. 04. 21  
1주차 강의안

## 금일 강의의 목적

1. 웹 크롤링의 기본적인 흐름을 이해
2. 본 강의에서 자주 사용될 단어의 정의 학습
3. 본 강의 중 실습을 진행할 환경을 조성
4. 웹크롤링을 위한 최소한의 python 코딩 학습
5. 강의시간 이외의 학습을 위한 연결고리(github / Slack)

# INDEX

## 1.이론

- 1.크롤링 / 스크래핑에 대한 이해
- 2.가상환경에 대한 이해
- 3.파이썬 코드 오류시 대처법

## 2.실습

- 1.파이썬의 기본: 변수, 자료형, 객체 이해하기
- 2.파이썬 반복문, 제어문 이용해보기
- 3.파이썬의 기본: 함수, 라이브러리 import 해보기
- 4.네이버 요일별 웹툰 페이지 크롤링
- 5.Github 방문과 Slack 계정 연결

## 1. 크롤링 / 스크래핑에 대한 이해

**컴퓨터 소프트웨어 기술을 활용하여 웹 상의 정보를  
자동적으로 수집하여 정형의 데이터 형식으로 추출하는 행위**

## 1. 크롤링 / 스크래핑에 대한 이해



컴퓨터 소프트웨어 기술을 활용하여

웹 상의 정보를

자동적으로 수집하여

정형의 데이터 형식으로 추출하는 행위

## 1. 크롤링 / 스크래핑에 대한 이해



컴퓨터 소프트웨어 기술을 활용하여



웹 상의 정보를

실제 코드를 작성하는 언어로 파이썬을 선택!



파이썬 개발환경은 모두 통일하여 진행  
— 가상환경 가이드 참고!

자동적으로 수집하여



정형의 데이터 형식으로 추출하는 행위

## 1. 크롤링 / 스크래핑에 대한 이해



컴퓨터 소프트웨어 기술을 활용하여



웹 상의 정보를

2주차에 학습할 내용인 HTML/CSS/JS

웹이라는 문서의 내용을 표현하는 언어

웹페이지에 보이는 모든것들은 위의 세가지 언어로 표현되고 있음

## 1. 크롤링 / 스크래핑에 대한 이해



서버를 대여하여 항상 정해진 시간에 자동적으로 크롤링 프로그램을 작동시킬 수 있다.

컴퓨터 소프트웨어 기술을 활용하여

웹 상의 정보를



자동적으로 수집하여



정형의 데이터 형식으로 추출하는 행위



## 1. 크롤링 / 스크래핑에 대한 이해



컴퓨터 소프트웨어 기술을 활용하여

크롤링한 데이터는 csv / xlsx와 같은 포맷의 파일로 추출하거나,  
혹은 데이터베이스에 저장시킨다.



자동적으로 수집하여



정형의 데이터 형식으로 추출하는 행위

## 2. 가상환경에 대한 이해

### 내 컴퓨터

PYTHON  
3.6

LIBRARY

패키지 A  
패키지 B  
패키지 C

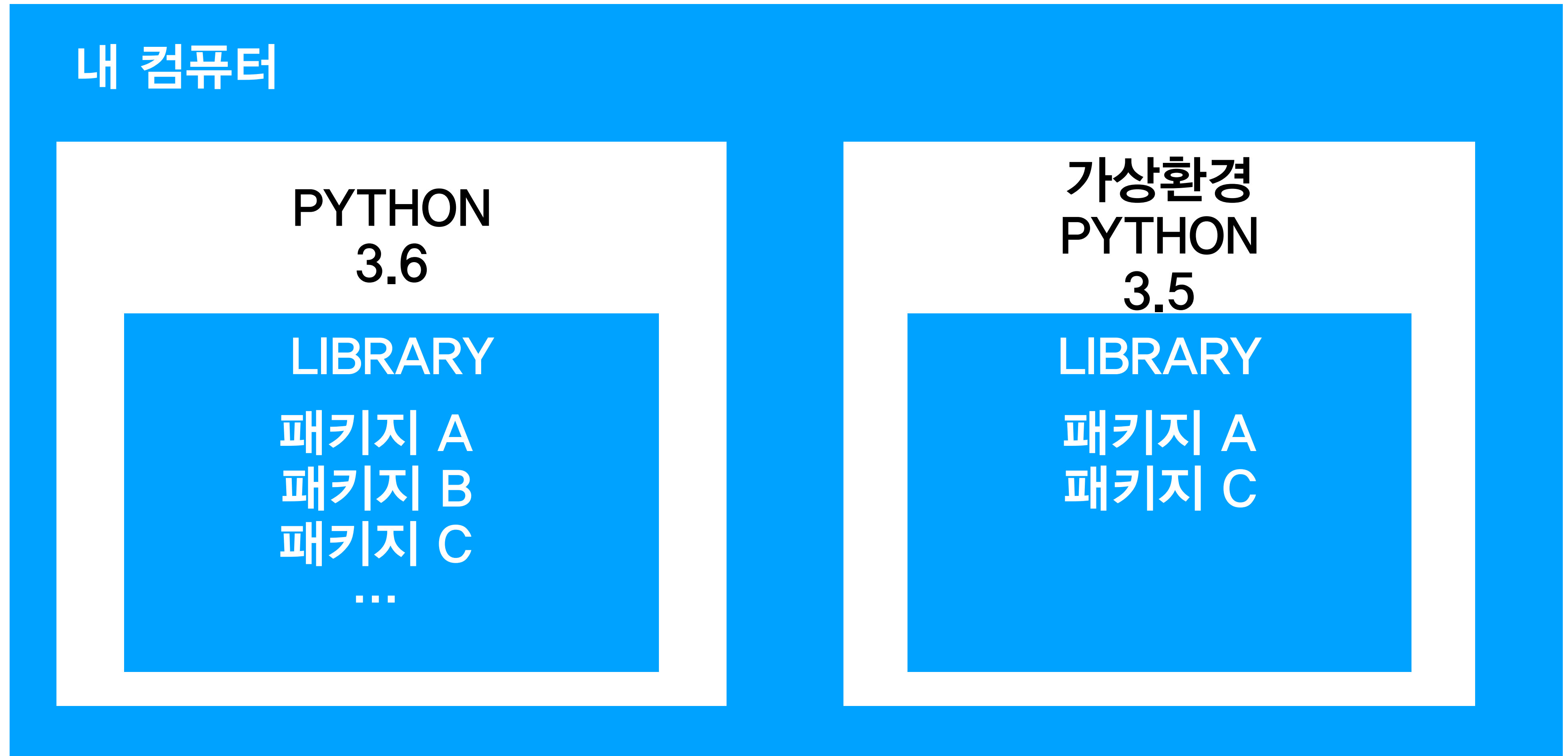
...

내 컴퓨터 안에 있는 파이썬에는  
여러가지 기능이 담겨있는 패키지들을 담은  
library가 존재합니다.

그러나 프로젝트마다 사용하는 패키지가 다르기에  
모든 프로젝트에 모든 패키지를  
연결시킬 필요가 없습니다.

그래서 우리는 가상환경을 통해 크롤링만을 위한  
개발환경을 구축하고자 합니다.

## 2. 가상환경에 대한 이해



## 2. 가상환경에 대한 이해

내 컴퓨터

가상환경  
PYTHON  
3.6.5

LIBRARY

Selenium = 3.4.3  
Bs4 = 0.0.1  
Pandas = 0.20.3  
...

왼쪽과 같이 크롤링을 위한  
패키지만 담은  
가상환경을 만들어주는 것입니다.

개발환경 가이드.pdf 파일을  
잘 따라오셨다면  
이미 완성된 개발환경으로  
이번 프로젝트를 시작할 수 있습니다.

## 3. 파이썬 코드 오류시 대처법

Traceback (most recent call last):

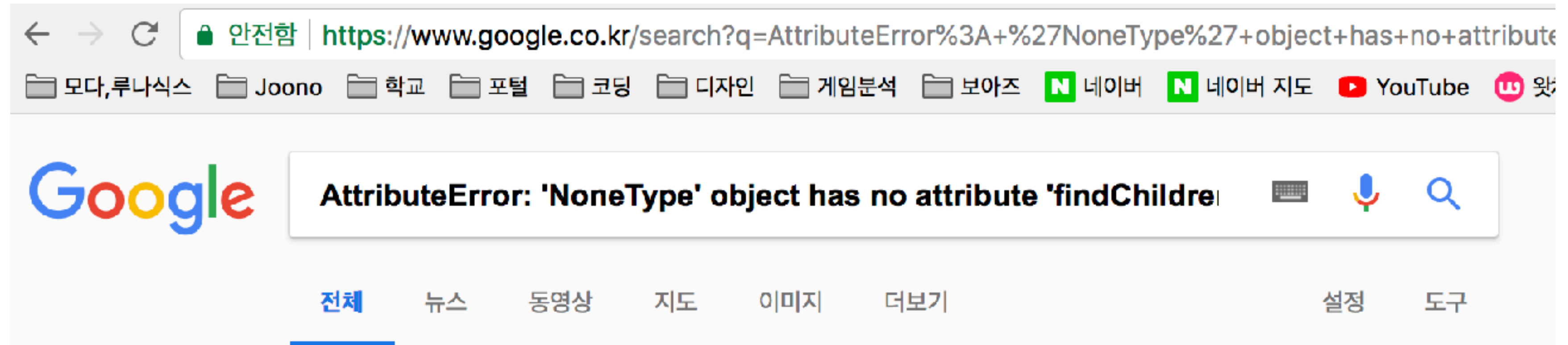
File `"/Users/joono/GoogleDrive/SKKU/BOAZ/패스트캠퍼스/fast_campus/코드/test_case/exam2.py"`, line 51, in `<module>`  
    `rel_words_chil = rel_words_parent.findChildren()`

**AttributeError: 'NoneType' object has no attribute 'findChildren'**

파이썬에서 오류가 발생되면  
오류가 발생한 코드와  
어떠한 에러가 발생했는지를 보여줍니다.

이때, 위의 빨간 부분을 구글에 검색해 봅니다.

### 3. 파이썬 코드 오류시 대처법



python - AttributeError: 'NoneType' object has no attribute 'findChildren ...

<https://stackoverflow.com/q/31238260> ▼ 이 페이지 번역하기

soup.find('ul',{'class': 'tags'}) is returning None . If you want to use this in a list comprehension you need to filter out values which are None before using them. There's a trick where you put the value in a list so you can filter it: tags\_dict['tags'] = [child.text for tag in [soup.find('ul',{'class': 'tags'})] if tag for child in tag.

python - AttributeError: 'NoneType' object has no attribute 'hide ...

<https://stackoverflow.com/.../attributeerror-nonetype-object-has-n...> ▼ 이 페이지 번역하기

2017. 4. 18. - You call hide() on the returned node from self.findChild(...) . The problem is that self.findChild(...) returned None (it didn't find the tag you thought it would), so you actually try to call hide() on None .

python - Can't find children leafs using elementtree.find ...

2018년 3월 2일

python

2017년 8월 19일

python - Strange error with AttributeError: 'NoneType' object has ...

2017년 5월 24일

python - PyQt5: Got AttributeError while using QObject and QThread ...

2013년 12월 18일

stackoverflow.com 검색결과 더보기



## 3. 파이썬 코드 오류시 대처법

← → ↻ 안전함 | <https://stackoverflow.com/questions/31238260/attributeerror-nonetype-object-has-no-attribute-findchildren-beautifu>

모다,루나식스 Joono 학교 포털 코딩 디자인 게임분석 보아즈 N 네이버 N 네이버 지도 YouTube 왓챠 :: 오늘 영화도 부탁...

stackoverflow Questions Developer Jobs Tags Users Search...

1 Answer active oldest votes

▲ 0 ▼

`soup.find('ul',{'class': "tags"})` is returning `None` .


If you want to use this in a list comprehension you need to filter out values which are `None` before using them.

There's a trick where you put the value in a list so you can filter it:

✓

```
tags_dict['tags'] = [child.text
                     for tag in [soup.find('ul',{'class': "tags"})]
                     if tag
                     for child in tag.findChildren('a')]
```

share improve this answer edited Jul 6 '15 at 20:42 answered Jul 6 '15 at 5:47

 Peter Wood 14.7k ● 3 ● 28 ● 62

Thanks Peter! I made that modification and was getting: `UnboundLocalError: local variable 'tag' referenced before assignment` , so I changed course a bit, ditching the list comprehension for an if statement that checks for the presence of 'ul'. `if soup.find('ul',{'class': "tags"}) != None:`  
`for tag in soup.find('ul',{'class': "tags"}).findChildren('a'): tags_dict['tags']`  
`= tag.text` I can run error-free, *but* the resulting dict contains an entry only for the *last* id in my postids list. I think I am close so will be sure to update once I solve. – [ispoorbutsexy](#) Jul 6 '15 at 18:39

Ah, I got the loop the wrong way around, fixed. – Peter Wood Jul 6 '15 at 20:42

Success! Thanks Peter, working as expected for me. Updated the initial post with the code I ended up using. – [ispoorbutsexy](#) Jul 7 '15 at 0:54

## 3. 파이썬 코드 오류시 대처법

대부분 오류를 검색하면 제일 상단에 Stackoverflow 사이트의 질문과 답변이 나오게 됩니다.

사실 나의 코드와 질문자의 코드가 완벽하게 같지는 않지만 적어도 오류가 나는 원인을 답변자가 상세히 설명하기 때문에 웬만한 오류는 구글링으로 해결이 가능합니다.





# 1. 파이썬의 기본: 변수, 자료형, 객체 이해하기

## 변수

글로 생각한다면 대명사에 해당하는 개념!

**이것 = 사과**라고 한다면 우리는 이러한 선언 뒤에서 **‘이것’**을 사용한다면 사과가 이것에 대입되어 사용된다는 것을 알수 있다.

컴퓨터 언어에서도  $a = 3$ 이라고 선언한 후에 뒤에서  $a$ 를 입력한다면 컴퓨터는 자동적으로 3으로 인식하게 된다.

또한  $a = 3$  후에 다시  $a = 5$ 라고 선언한다면 가장 최근의 선언인 5로 인식하게 된다.



# 1. 파이썬의 기본: 변수, 자료형, 객체 이해하기

## 자료형

자료형이란 변수의 성격을 나타내는 것이다.

파이썬에서는 여러가지 자료형이 있다.

이 강의에서는 기본적인 int(정수형) / float(실수형) / string(문자형)을 주로 사용할 것이다.

**정수형** : 소수점이 없는 숫자

Ex) 5

**실수형** : 소수점을 포함한 숫자

Ex) 5.7

**문자형** : 말 그대로 문자를 표현하는 자료형

Ex) '안녕하세요' - 문자형은 따옴표와 쌍따옴표를 사용하여 구분한다.



# 1. 파이썬의 기본: 변수, 자료형, 객체 이해하기

## 객체

객체란 변수, 함수, 자료구조 등등 여러가지의 컴퓨터 프로그래밍 안의 개념들을 실제 메모리에 할당하여 식별자를 가지고 구분이 가능한 것을 말한다.

실제 변수를 선언하게 되면 우리가 흔히 말하는 ram의 일부분에 변수가 할당되어 저장되게 된다.



# 1. 파이썬의 기본: 변수, 자료형, 객체 이해하기

# 1. 파이썬의 기본: 변수, 자료형, 객체 이해하기

```
variable_int = 3
```

# print(x)는 x를 출력하는 함수입니다.

```
print(variable_int)
```

```
variable_float = 3.6
```

```
print(variable_float)
```

# str(x) / int(x) / float(x)는 괄호안의 x의 자료형을 변경해줍니다.

# 그러나 x가 문자열인 경우 int나 float로 변환은 불가능합니다.

```
variable_string = 'hello'
```

```
print(variable_string)
```

```
print(type(variable_int))
```

```
print(type(variable_string))
```

# 아래와 같이 " 혹은 ""로 감싼 대상은 문자열로 취급합니다.

```
variable_int_to_string = str(variable_int)
```

```
print(variable_int_to_string)
```



# 1. 파이썬의 기본: 변수, 자료형, 객체 이해하기

```
# 파이썬의 기본 배열중 리스트와 딕셔너리
# 리스트 []를 사용하여 선언
list_ = ['a', 3, 5]
# list는 index를 사용하여 내용에 접근
# 파이썬은 0부터 시작하는 index를 가지고 있음
print(list_[0])
# 리스트에 새로운 변수를 삽입하는 함수 append
list_.append(7)
print(list_[3])

# 딕셔너리
# 딕셔너리는 {key : value}로 이루어진다
# key는 중복이 불가능하지만 value는 중복이 가능
dic = {'a' : 1, 'b' : 2}
# 딕셔너리는 key값으로 내용에 접근이 가능하다
print(dic['a'])
# 딕셔너리에 새로운 변수를 삽입하는 함수 update
dic.update({'c' : 3})
print(dic['c'])
```



## 2. 파이썬 반복문, 제어문 이용해보기

### 반복문

반복문이란 원하는 코드를 원하는 만큼 자동적으로 반복하여 실행하게 하는 문장이다. Python에서는 for문과 while문을 사용한다.

for문은 정해진 크기의 배열의 길이만큼 반복하거나, 혹은 내가 정해주는 횟수만큼 반복하는 문장이다.

그러나 while문은 정해진 길이가 없이 어떠한 조건을 만족할 때까지 무한히 반복이 가능한 문장이다.

또한 반복문에서 continue와 pass를 사용할 수 있습니다. 제어문을 학습하면서 설명 드리겠습니다.



## 2. 파이썬 반복문, 제어문 이용해보기

```
list_ = ['a', 1, 3, 5, 7]
# 단순한 숫자 범위에서의 for문
for number in range(0, 5):
    print(number)
# 리스트 안의 내용을 하나씩 사용하는 for문
for item in list_:
    print(item)
# enumerate를 사용하여 index와 내용물 둘 다를 사용하는 for문
for index, item in enumerate(list_):
    print(str(index+1) + 'th item is ' + str(item))

# n이 10보다 작은 경우에만 반복하고 n이 10 이상이 되는 순간 멈추게 되는 while문
n = 0
while n < 10:
    print(n)
    n = n + 1
```

## 2. 파이썬 반복문, 제어문 이용해보기

### 제어문

제어문이란 조건에 따라 각기 다른 코드를 실행하고자 할 때 사용합니다.

조건은 여러개로 설정이 가능합니다.

또한 반복문 안에 조건문을 넣어서 조건에 맞는 경우 다음 반복으로 강제로 이동시키는 (continue)와, 조건에 맞는 경우 그대로 코드를 진행시키는 명시만 해주는 (pass) 를 사용 가능합니다.





## 2. 파이썬 반복문, 제어문 이용해보기

# 7이라는 숫자를 n이라는 변수에 할당해 놓습니다.

```
n = 7
```

# n이 5보다 큰지 작거나 같은지 비교하는 조건문입니다.

```
if n > 5:
```

```
    print('n is bigger than 5')
```

```
else:
```

```
    print('n is smaller than or equal to 5')
```

# for문과 if문을 함께 쓴 코드입니다. 들여쓰기로 구분하는 부분을 유심히 봐주시기 바랍니다.

```
for number in range(0,10):
```

```
    if (number >= 0) and (number < 4):
```

```
        print(number, ' ', '0 <= number < 4')
```

```
    elif (number >= 4) and (number < 7):
```

```
        print(number, ' ', '4 <= number < 7')
```

```
    else:
```

```
        print(number, ' ', '7 <= number')
```

# for문과 if문을 함께 써서, 조건을 만족한다면 바로 다음 루프로 넘어가게 한 코드입니다.

```
for number in range(10,20):
```

```
    if number > 15:
```

```
        continue
```

```
    print(number, ' ', 'number <= 15')
```



### 3. 파이썬의 기본: 함수, 라이브러리 import 해보기

## Import문

웹크롤링 강의를 진행하려면 기본 파이썬에서 제공하는 함수 뿐만 아니라 다른 패키지들의 함수도 사용해야 하기 때문에 패키지를 불러오는 import문을 사용하게 됩니다.

패키지는 함수들을 모아 놓은 집합으로 생각하시면 됩니다.  
함수들을 모아 놓았기 때문에 패키지를 불러오면  
그 안의 함수를 사용할 수 있게 됩니다.

실습에는 무조건 패키지 불러오기가 필요합니다.



### 3. 파이썬의 기본: 함수, 라이브러리 import 해보기

# 패키지를 불러옵니다.

```
import package
```

# 패키지를 불러와서 pk라는 이름으로 사용합니다.

```
import package as pk
```

# package에서 module을 불러옵니다.

```
from package import module
```

# module에서 function이나 variable을 불러옵니다.

```
from module import function or variable
```

# package나 module에서 모든것을 불러옵니다.

```
from package or module import *
```



## 4. 네이버 요일별 웹툰 페이지 크롤링

아래의 코드는 앞으로 실습할 환경과 강의 진행을 위한 맛보기 코드입니다.

앞으로 모두 배울 내용이니 이해가 되지 않으시더라도

한번 코드를 읽어보시기 바랍니다.



## 4. 네이버 요일별 웹툰 페이지 크롤링

```
import requests
from bs4 import BeautifulSoup
import re

days = ['mon', 'tue', 'wed', 'thu', 'fri', 'sat', 'sun']
day = days[0]

url = 'http://comic.naver.com/webtoon/weekdayList.nhn?
week=wed'+day+'&view=list&order=ViewCount'

response = requests.get(url)

html = response.text

soup = BeautifulSoup(html, 'html.parser')
title_pattern = re.compile('^/webtoon/list.nhn\?titleId=')
titles = soup.find_all('a', attrs = {'href' : title_pattern})
title_text = []
for title in titles:
    title_text.append(title.text.strip())

while '전체보기' in title_text: title_text.remove('전체보기')
while '' in title_text: title_text.remove('')
while 'NEW' in title_text: title_text.remove('NEW')

print(title_text)
```



## 5. Github 방문과 Slack 계정 연결

Q & A