

2022 LSK-시몬느 언어학학교 - 언어연구를 위한 Python 프로그래밍 (초급)

01

Word Cloud

윤태진 교수
성신여자대학교 영어영문학과



강의안내



01 Python 소개

02 Python Building Blocks

03 Word Cloud 만들기



01

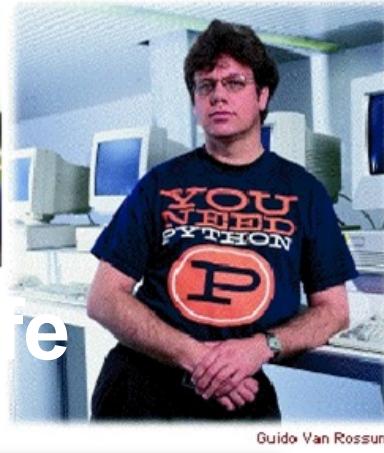
Python 소개

Guido van Rossum 귀도 반 로섬



89

크리스마스 연휴



99

DARPA

Computer Programming for Everybody



팔로우

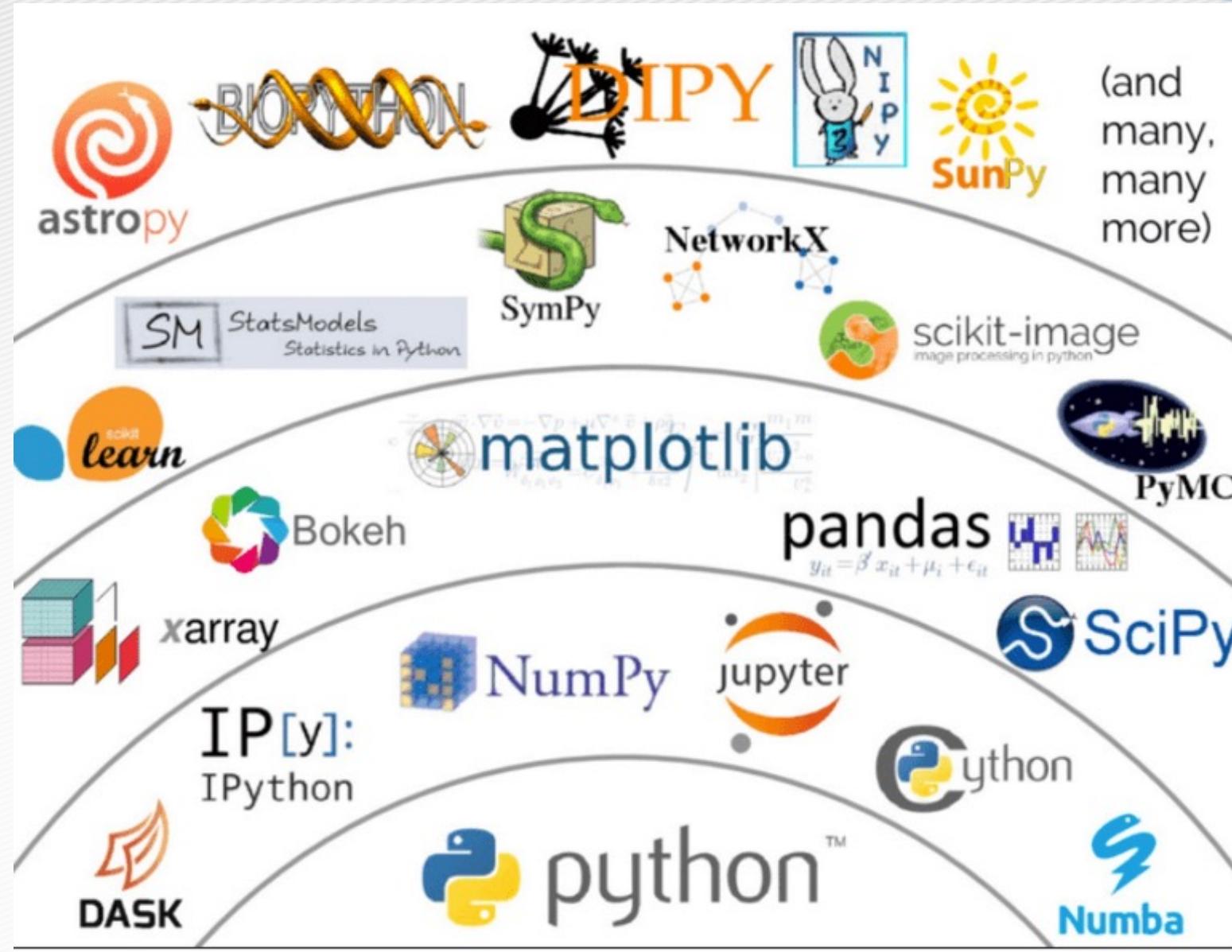
Guido van Rossum 

@gvanrossum

Python's BDFL-emeritus, Distinguished Engineer at Microsoft, Computer History Fellow. Opinions are my own. He/him.

◎ San Francisco Bay Area ⌂ python.org/~guido/ ☏ 가입일: 2008년 8월

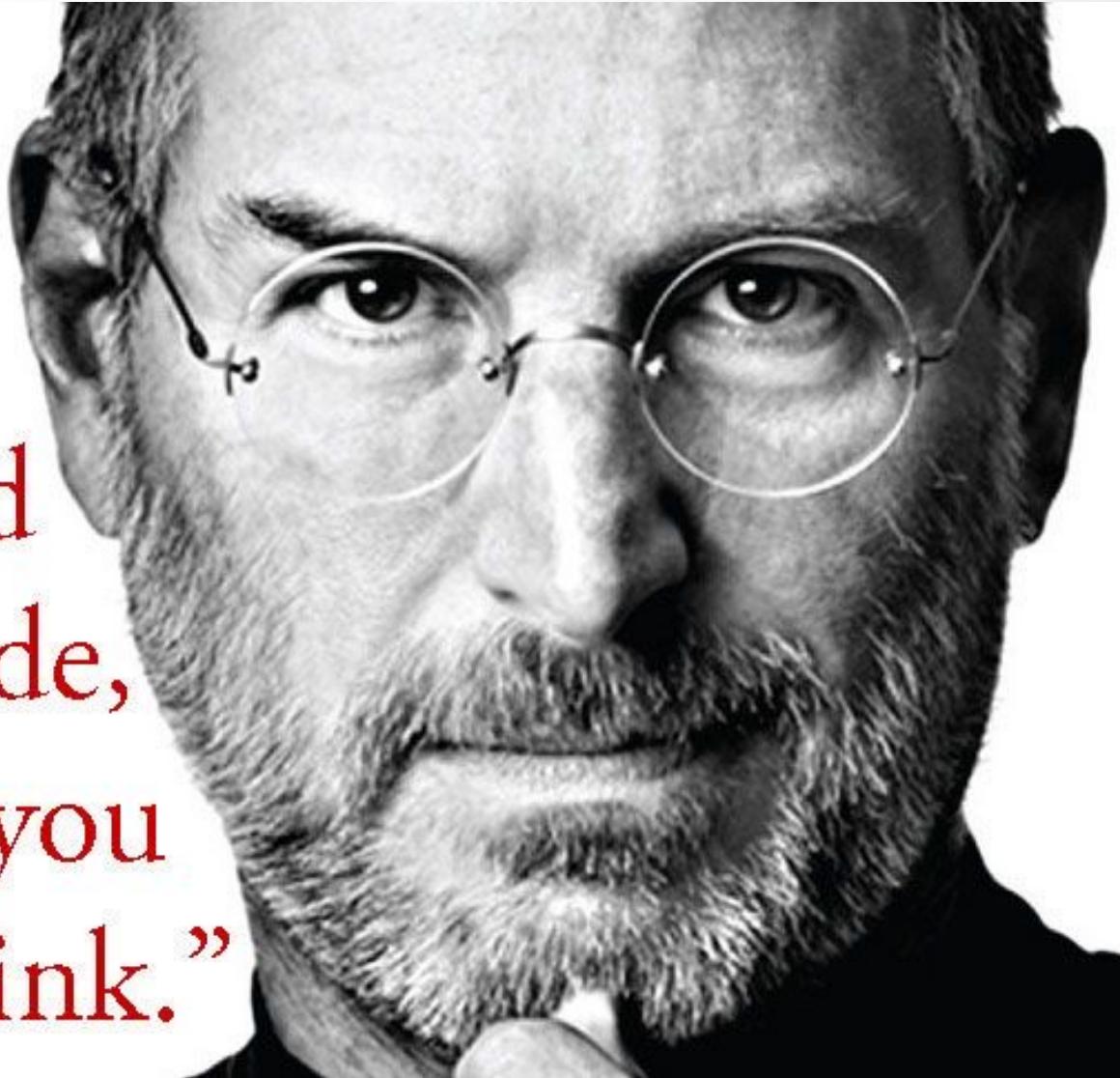
515 팔로우 중 19.9만 팔로워



Steve Jobs

1955-2011

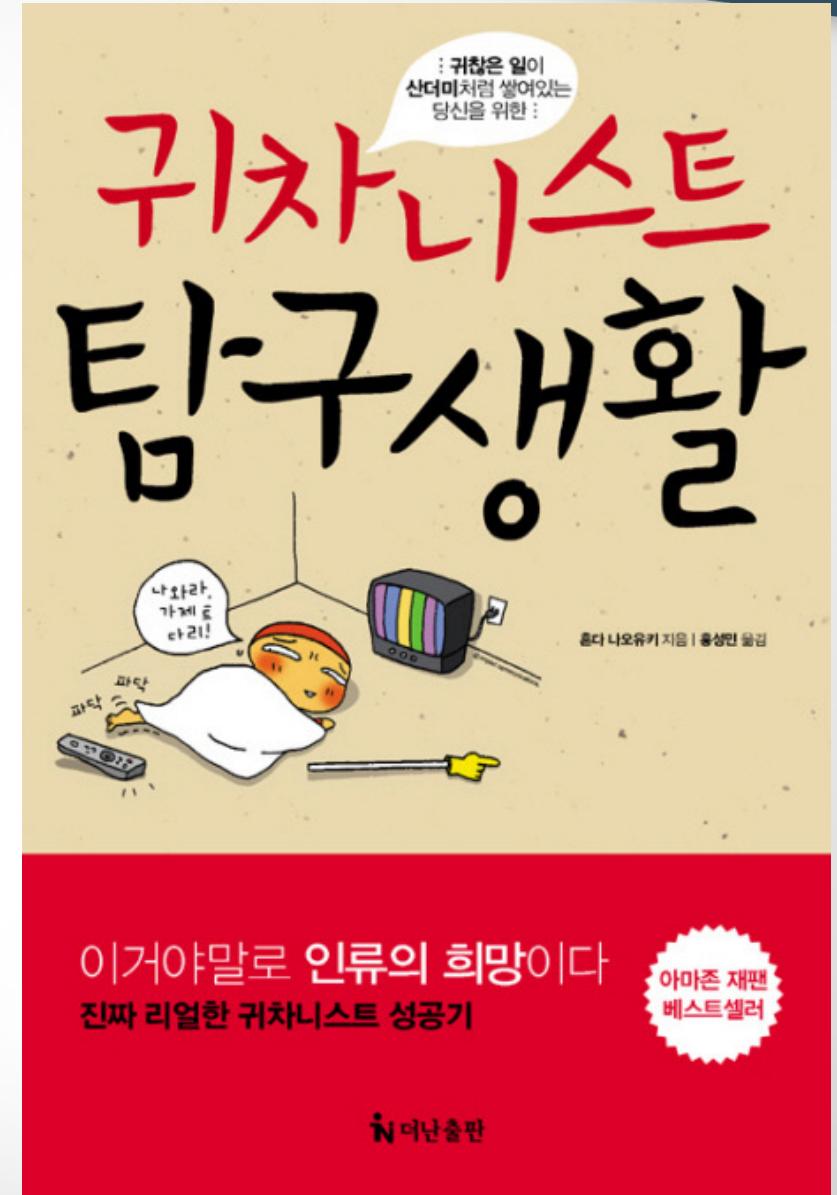
“Everyone should
learn how to code,
it teaches you
how to think.”



코딩 - What & why?



<https://www.educatecity.com/2020/06/what-are-computersr.html>





The longest English word?

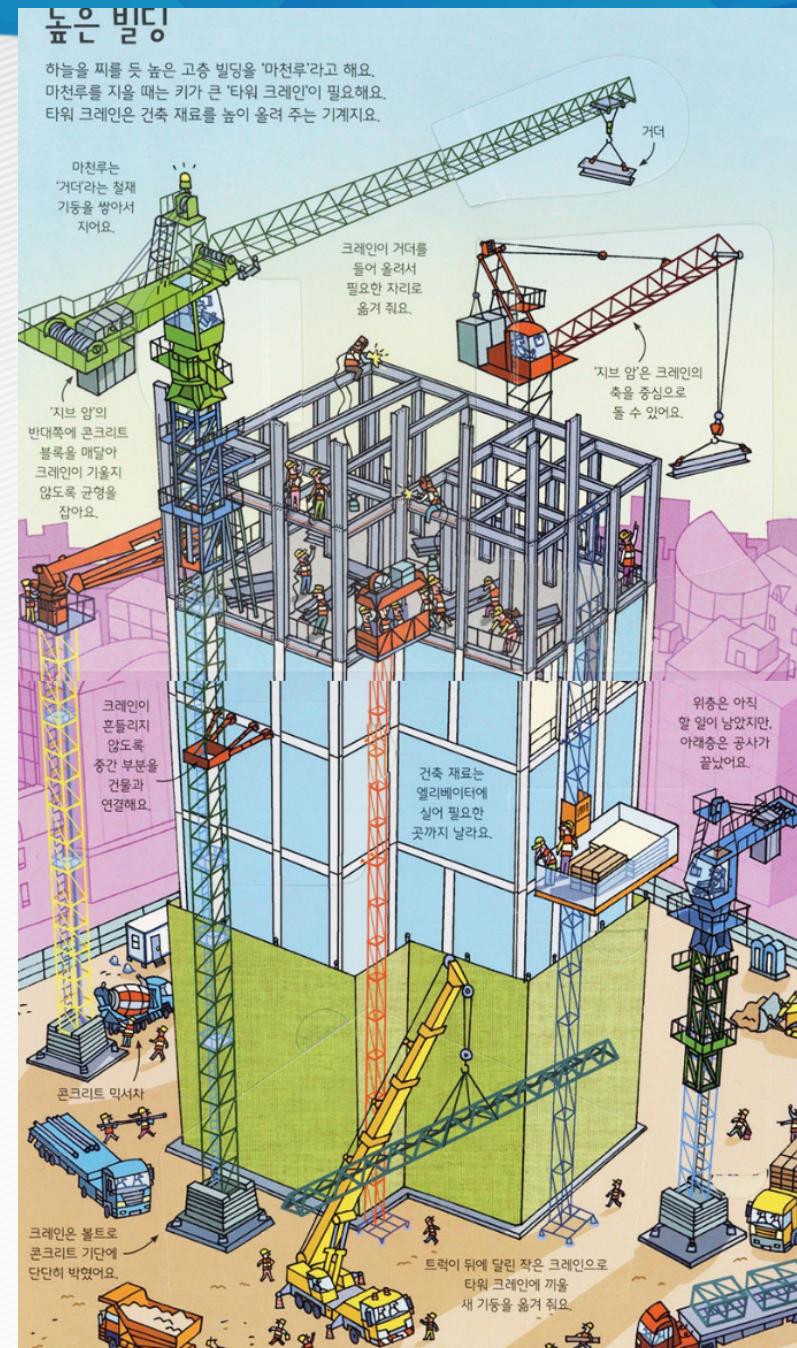
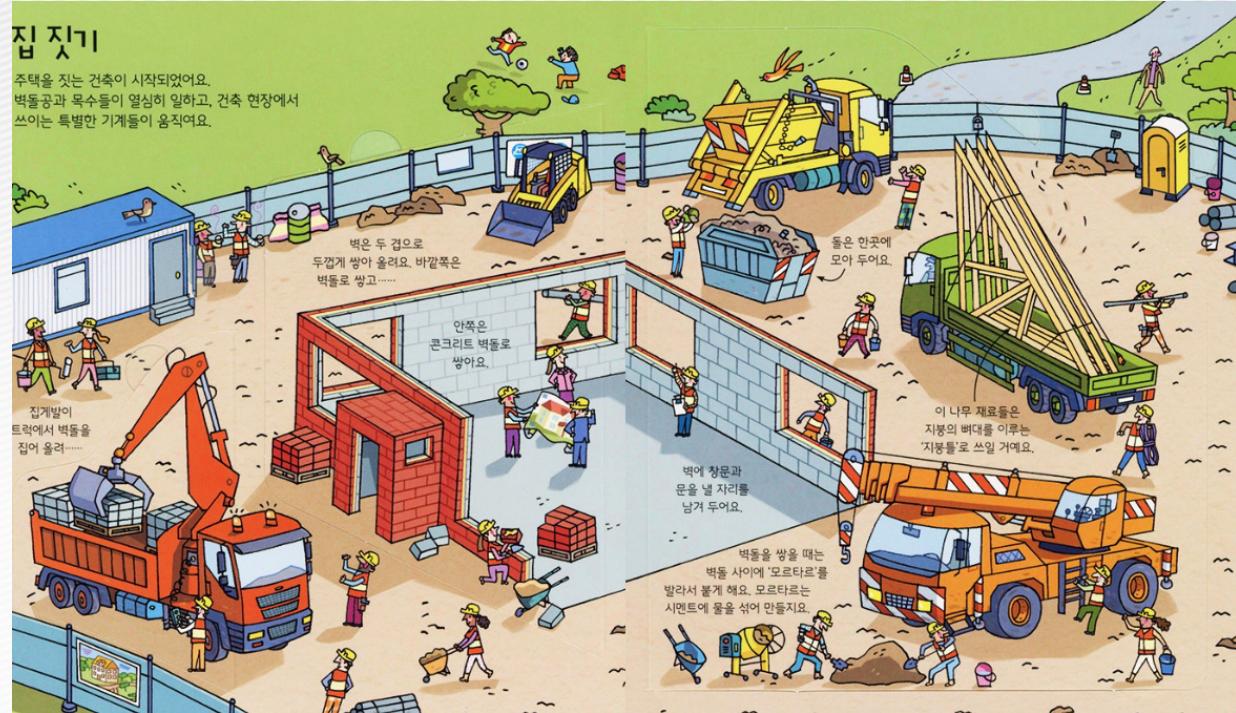
Pneumonoultramicroscopicsilico volcanoconiosis





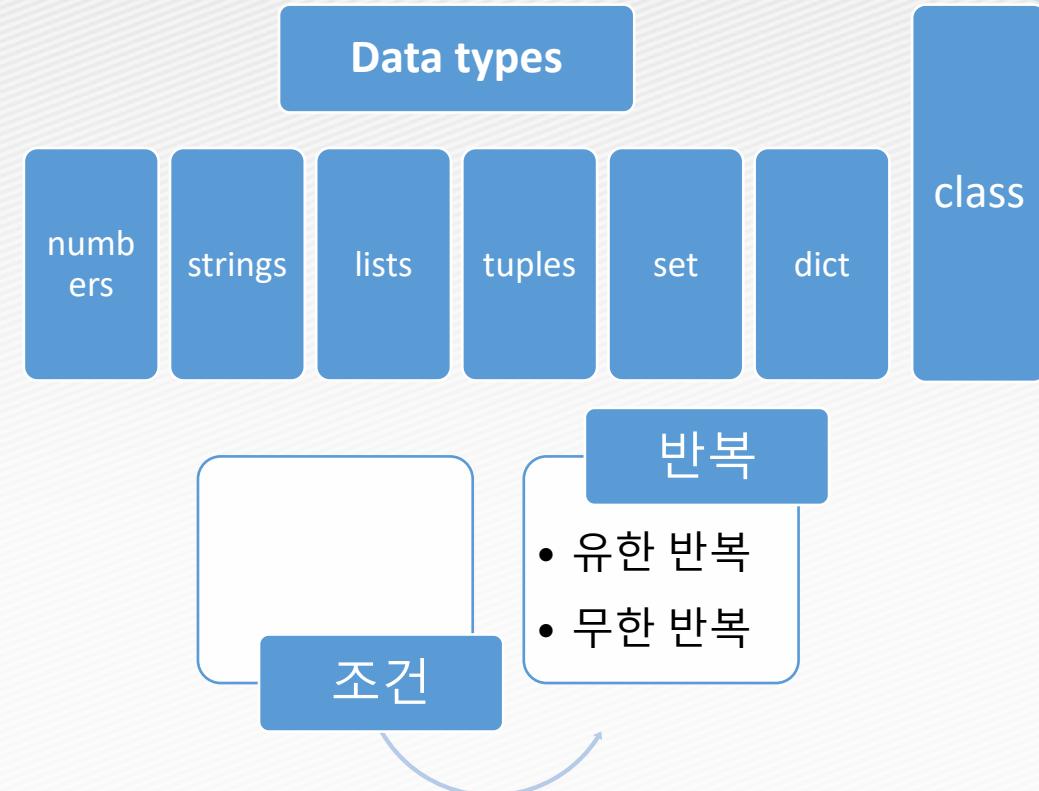
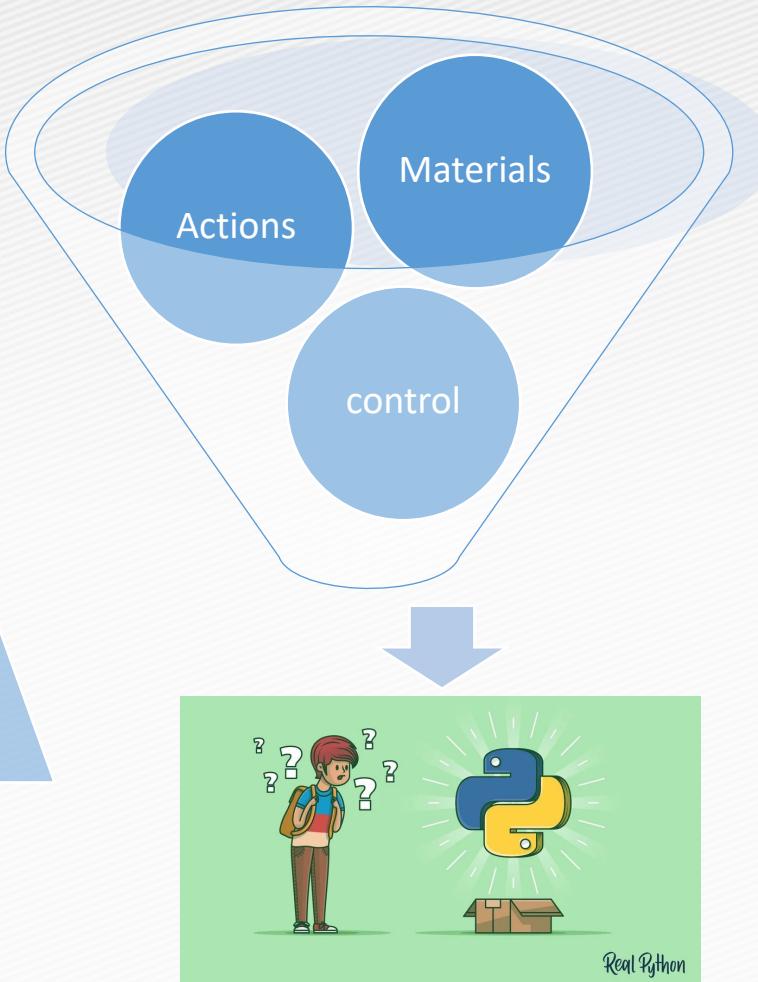
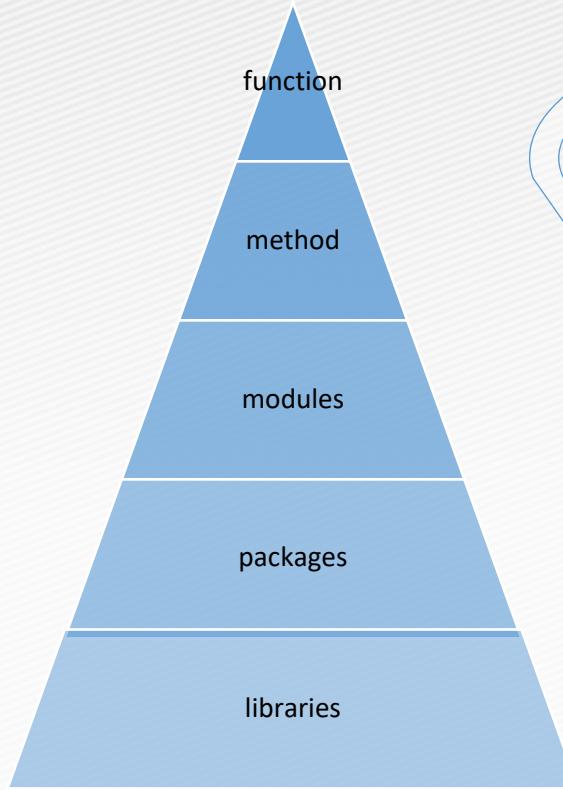
02

Building Blocks



요리조리 열어 보는 집 짓는 기계들

- http://picturebook-museum.com/user/book_detail.asp?idx=17363



How to do things with Python

Data Types



things



01

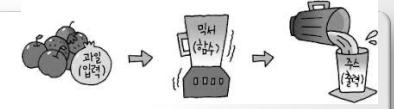
doing

02

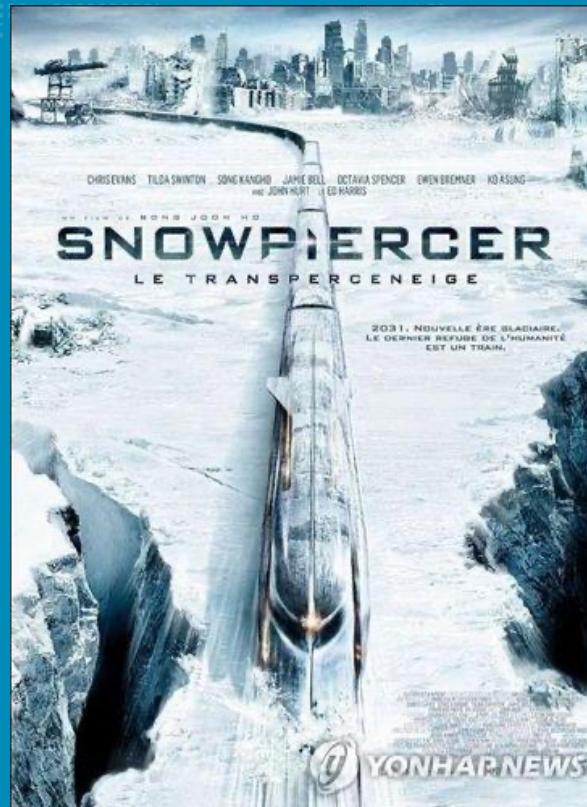
Function

Method

클래스에서 정해진 함수



len(string) string.split()



- len(): 문자열, 리스트 등의 길이를 알려주는 함수
- split(): 문자열 양쪽에 불필요한 문자를 빼어 내어 리스트(list)를 만드는 메소드

공백(space), \t, \n, \r

len(' Pneumonoultramicroscopicsilicovolcanoconiosis')
결과

'사과 배 포도 오렌지 감'.split()
결과 ['사과', '배', '포도', '오렌지', '감']

'사과, 배, 포도, 오렌지, 감'.split()
결과 ['사과,', '배,', '포도,', '오렌지,', '감']

'사과, 배, 포도, 오렌지, 감'.split(',')
결과 ['사과', ' 배', ' 포도', ' 오렌지', ' 감']

반복문

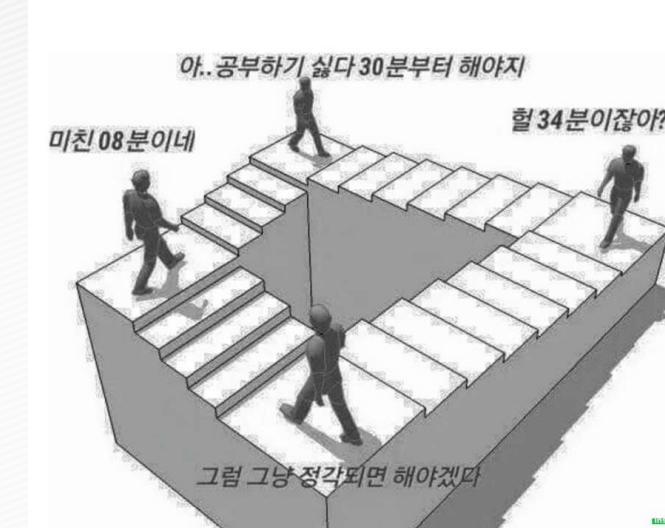
- for loop
- while loop

유한 loop



Angry Birds

무한 loop



Penrose Stairs

<https://www.linkedin.com/pulse/gamification-lessons-from-dr-angry-birds-gurpreet-bajaj>

http://www.koreapas.com/bbs/view.php?id=gofun&page=1447&sn1=&divpage=1&sn=off&ss=on&sc=on&select_arrange=headnum&desc=asc&no=234553

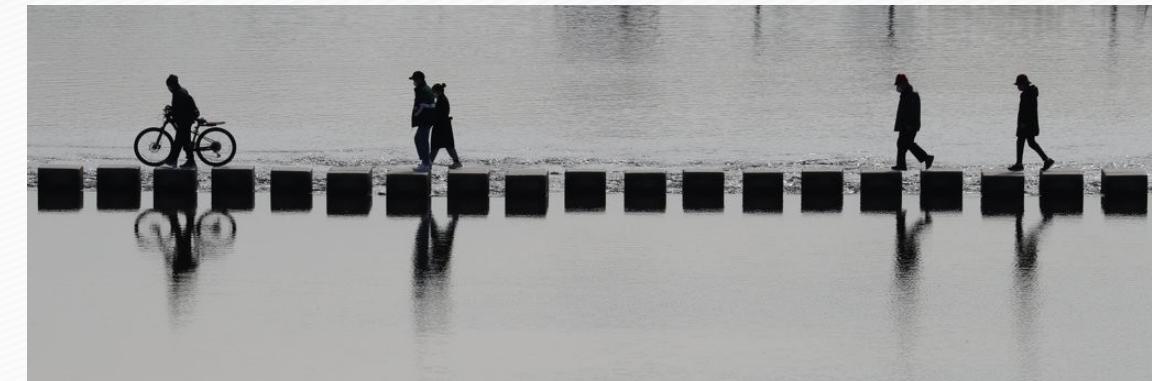
리스트에서 가장 긴 단어 찾기



```
words = ['hello', 'computer', 'bookshelves', 'chair', 'bicycle']
```

```
longest = len(words[0])
longest_word = words[0]
```

```
for word in words[1:]:
    if len(word) > longest:
        longest_word = word
    longest = len(word)
print("Longest word:", longest_word)
print("Length of the word:", longest)
```



List Comprehension

- 리스트 내에 for 반복문

$$A = \{x^2 : x \in \{0, 1, 2, \dots, 9\}\}$$

$$B = \{1, 2, 2^2, 2^3, \dots, 2^{20}\}$$

$$C = \{x | x \in A \text{ and } x \text{ is even}\}$$



[표현식 for x in 데이터집합 (if 조건)]

[표현식 for x in 데이터집합 (if 조건)]

```
a = [ x**2 for x in range(10) ]  
print(a)
```

```
for x in range(10):  
    print(x**2, end=" ")
```

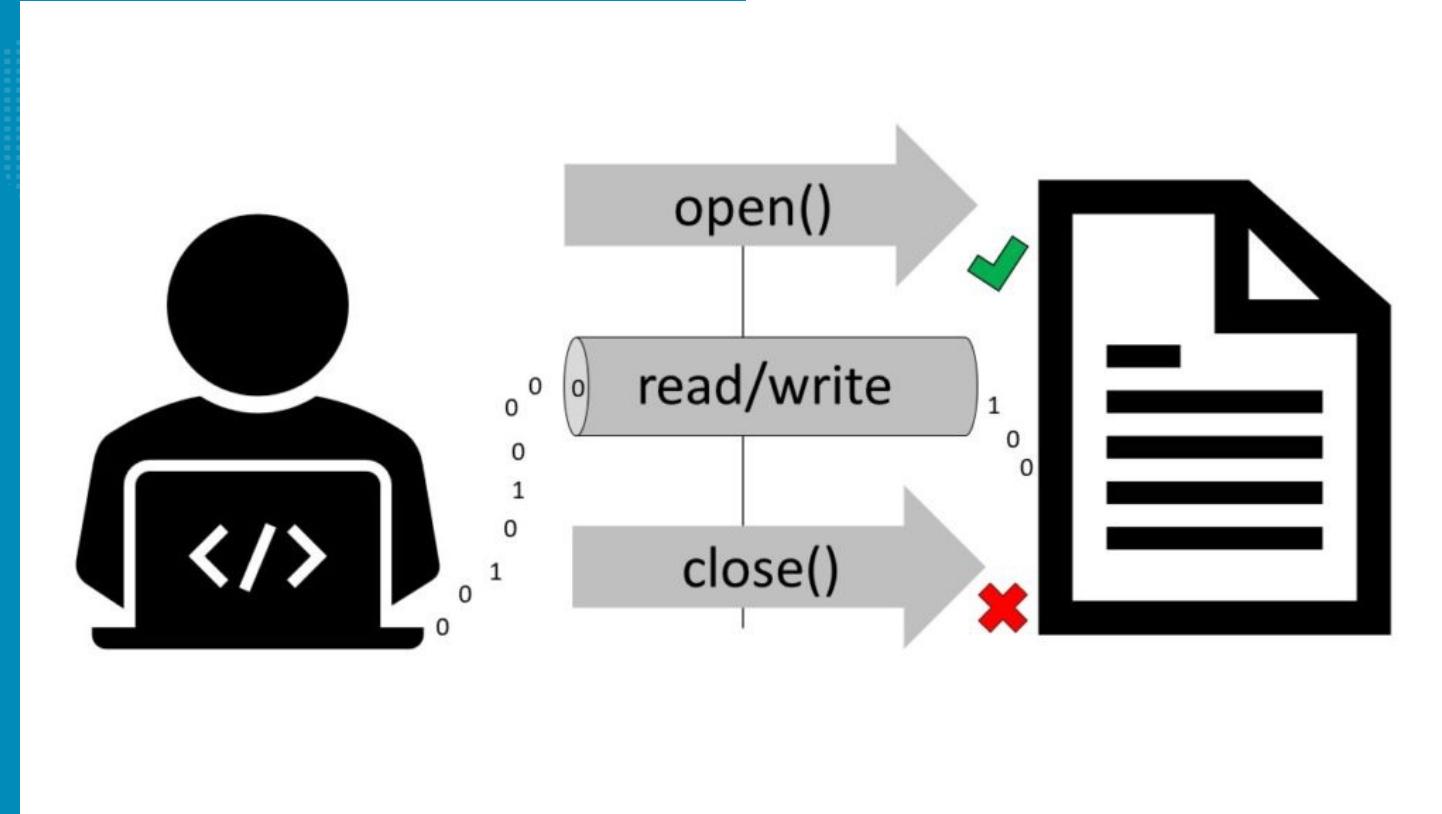
```
b = [ 2**b for b in range(21) ]  
print(b)
```

```
for b in range(21):  
    print(2**b, end=' ')
```



Alice's Adventures in Wonderland에서 가장 긴 단어는?

파일 입출력 과정



```
f = open('data/alice.txt')  
text = f.read()  
f.close()  
print(text)
```

```
f = open('alice.txt')
text = f.read()
f.close()
words = text.split()

longest = len(words[0])
longest_word = words[0]

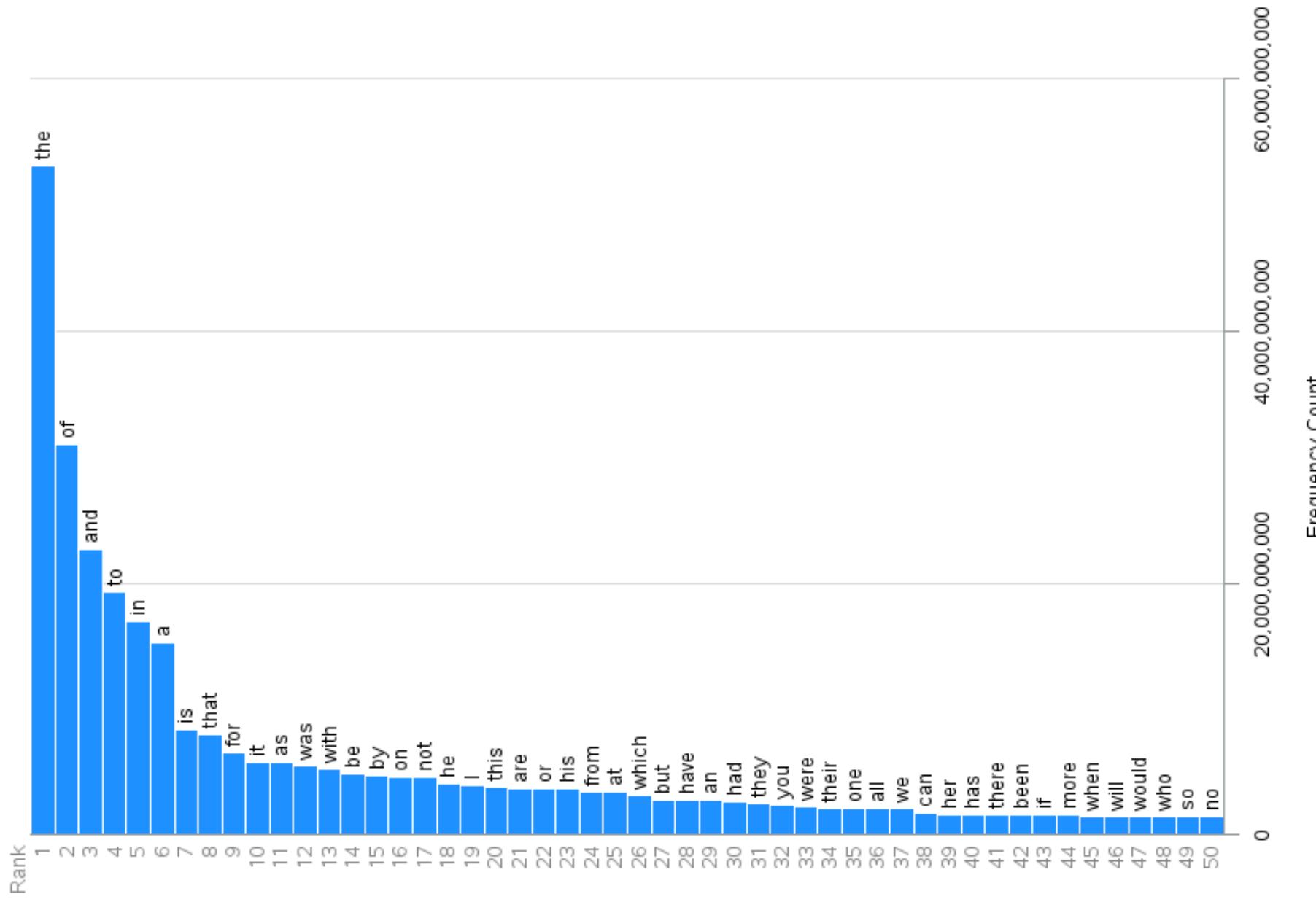
for word in words[1:]:
    if len(word) > longest:
        longest_word = word
        longest = len(word)
print("Longest word:", longest_word)
print("Length of the word:", longest)
```



영어 단어 중 빈도수가 가장 높은 것?

50 Most Frequent Words in English Writing

Based on Google books data



**“You shall know
a word by the
company it keeps”**

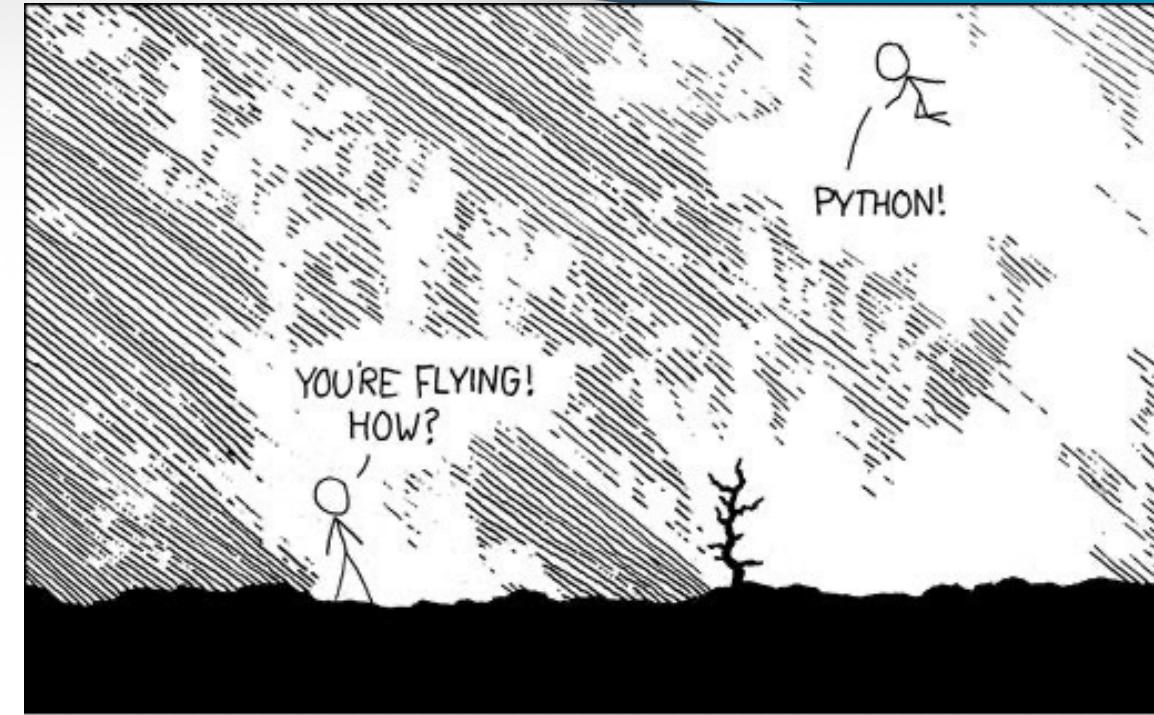
Firth, J. R. 1957





Library & Class

import 사용하기





pip 사용법

파이썬은 파이썬 표준 라이브러리(Python Standard Library, PSL) 이외에도 파이썬 패키지 인덱스(Python Package Index, PyPI)를 통해 다양한 패키지를 사용할 수 있습니다.

실습

pip3 install wordcloud

```
text = "Sungshin Sungshin Sungshin Sungshin University University"
```

```
import wordcloud  
wc_obj = wordcloud.WordCloud()  
library class
```

```
wc = wc_obj.generate(text)
```

method

```
wc.words_
```

attribute

pip3 install 패키지

python3 -m pip install 패키지

python3 -m pip install -U 패키지

-U: upgrade





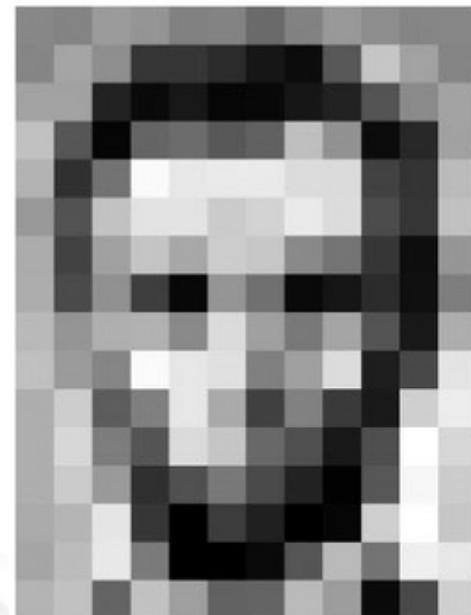
03

Word cloud 만들기

Image as numbers



- 흑백 그림은 픽셀의 밝기 값을 0-255 사이의 값으로 표현한 이미지



157	153	174	168	150	152	129	151	172	161	155	156
156	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	191	111	120	204	166	15	56	180
194	68	137	251	257	239	239	228	227	87	71	201
172	106	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	52	22	148
199	168	191	193	158	227	178	143	182	105	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	85	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	98	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

pixel

157	153	174	168	150	152	129	151	172	161	155	156
156	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	191	111	120	204	166	15	56	180
194	68	137	251	257	239	239	228	227	87	71	201
172	106	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	52	22	148
199	168	191	193	158	227	178	143	182	105	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	85	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	98	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

```
import matplotlib.pyplot as plt

plt.figure()
plt.imshow(wc)
plt.axis("off")
plt.show()

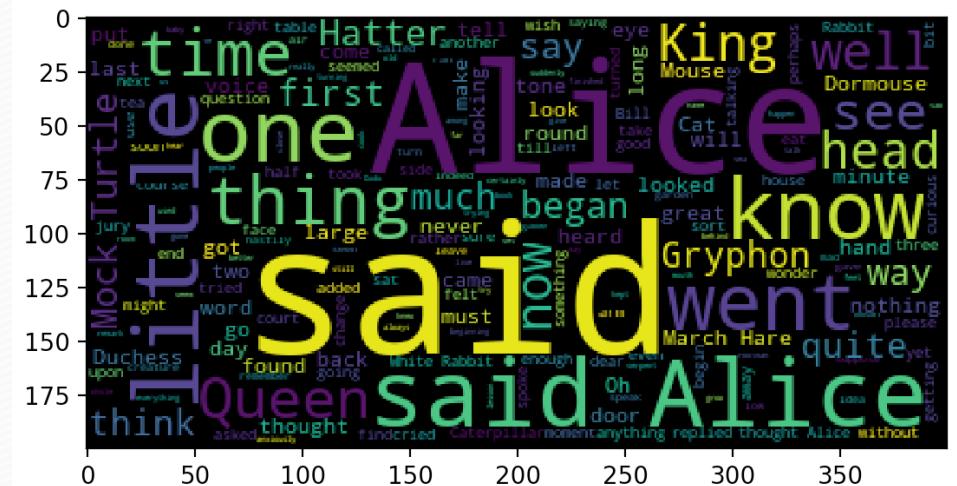
plt.savefig("wc_sungshin.png")
```

Alice's Adventures in Wonderland

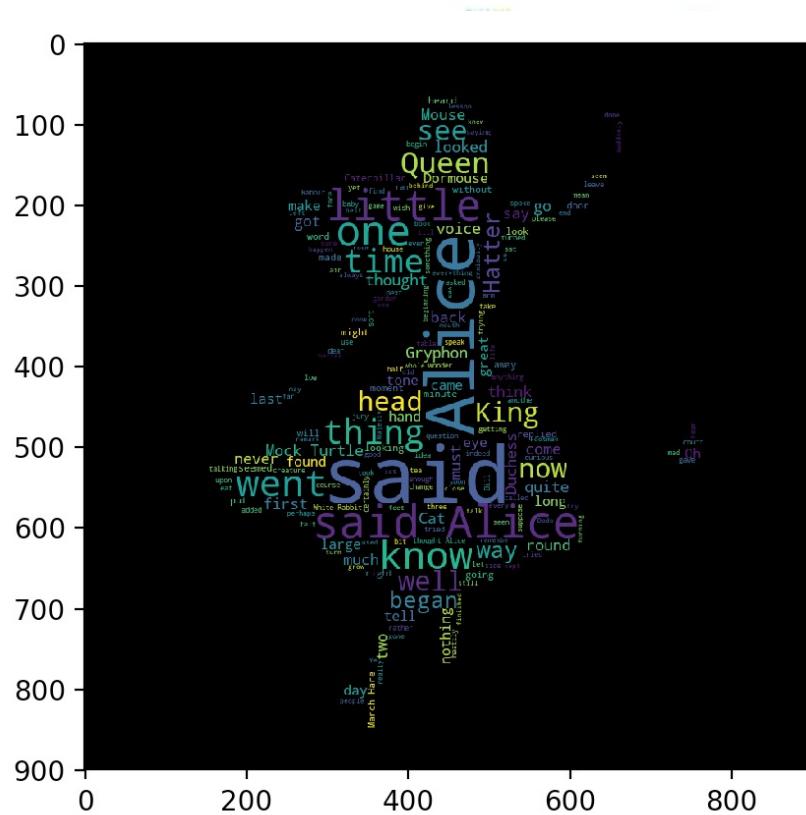
```
import wordcloud
text = open('data/alice.txt').read()
wc_obj = wordcloud.WordCloud()
wc = wc_obj.generate(text)
```

```
import matplotlib.pyplot as plt
```

```
plt.figure()
plt.imshow(wc)
plt.axis("off")
plt.show()
plt.savefig("wc_alice.png")
```



Masked Word Cloud 만들기



준비물: alice_mask.png

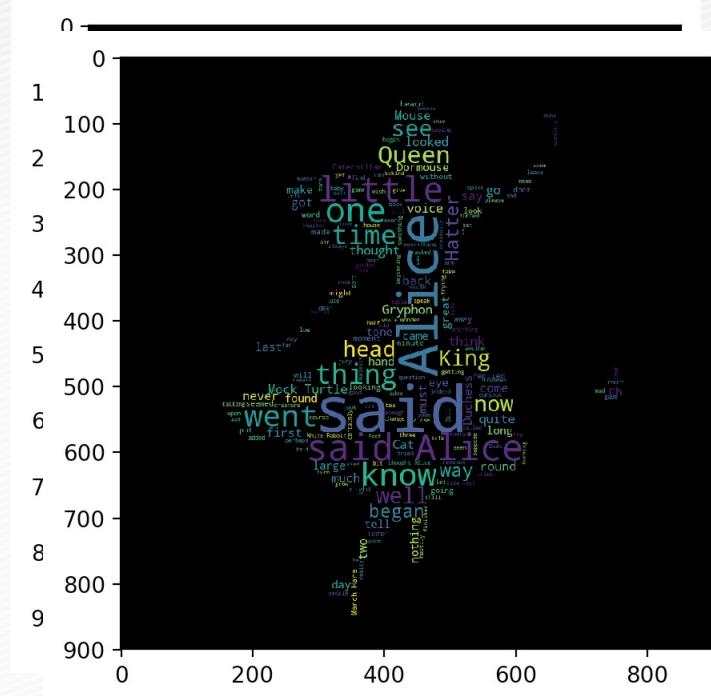
```
from PIL import Image
import numpy as np
alice_mask = np.array(Image.open('data/alice_mask.png'))
plt.figure()
plt.imshow(alice_mask)
plt.axis("off")
```

```
array([[255, 255, 255, ..., 255, 255, 255],
       [255, 255, 255, ..., 255, 255, 255],
       [255, 255, 255, ..., 255, 255, 255],
       ...,
       [255, 255, 255, ..., 255, 255, 255],
       [255, 255, 255, ..., 255, 255, 255],
       [255, 255, 255, ..., 255, 255, 255]], dtype=uint8)
```



```
import wordcloud
wc_obj = wordcloud.WordCloud(mask=alice_mask)
wc = wc_obj.generate(text)
```

```
plt.figure()
plt.imshow(alice_mask)
plt.imshow(wc)
plt.axis('off')
plt.show()
```



THANK YOU

