


```
import re
```

▼ 정규표현식, 문자열에서 패턴찾기

```
text = 'My id number is [G203_5A]'\ntext
```

 'My id number is [G203_5A]'

```
result = re.findall('a',text)    # re.findall(찾을 문자, 찾아낼 데이터)\nresult
```

 []


```
result = re.findall('A',text)\nresult
```

 ['A']


```
result = re.findall('i',text)\nresult
```

 ['i', 'i']


```
# 소문자 연속해서 찾기(단어가 연속되어 있으면 묶어줌) '[a-z]+'\nresult = re.findall('[a-z]+',text)\nresult
```

 ['y', 'id', 'number', 'is']

```
# 소문자 연속해서 찾기2 '[a-z]+'\nresult = re.findall('[a-z]',text)\nresult
```

 ['y', 'i', 'd', 'n', 'u', 'm', 'b', 'e', 'r', 'i', 's']

```
# 대문자 연속해서 찾기\nresult = re.findall('[A-Z]+',text)\nresult
```

 ['M', 'G', 'A']


```
# 숫자 찾기\nresult = re.findall('[0-9]',text)\nresult
```

 ['2', '0', '3', '5']


```
# 숫자 찾기2 (붙은 건 붙은 그대로)
result = re.findall('[0-9]+',text)
result
```

 ['203', '5']


```
# 영문자 및 숫자 찾기
result = re.findall('[a-zA-Z0-9]',text)
result
```

 ['M',
'y',
'i',
'd',
'n',
'u',
'm',
'b',
'e',
'r',
'i',
's',
'G',
'2',
'0',
'3',
'5',
'A']

```
# 영문자 및 숫자 연속해서 찾기
result = re.findall('[a-zA-Z0-9]+', text)
result
```

 ['My', 'id', 'number', 'is', 'G203', '5A']

```
# 영문자 및 숫자가 '아닌것' 찾기
result = re.findall('[^a-zA-Z0-9]',text)
result
```

 [' ', ' ', ' ', ' ', ' ', '[', '_', ']']

```
# 영문자 및 '_' 특수기호 찾기
result = re.findall('[Ww]',text)
result
```



```
['M',
 'y',
 'i',
 'd',
 'n',
 'u',
 'm',
 'b',
 'e',
 'r',
 'i',
 's',
 'G',
 '2',
 '0',
 '3',
 '-',
 '5',
 'A']
```

```
# 영문자 및 '-' 특수기호 연속해서 찾기
result = re.findall('[Ww]+', text)
result
```

```
 ['My', 'id', 'number', 'is', 'G203_5A']
```

```
# 영문자 및 '-' 특수기호 아닌 문자 찾기
result = re.findall('[WW]', text)
result
```

```
 [' ', ' ', ' ', ' ', ' ', '[', ' ']
```

▼ 문자열에서 특정이름 찾아내기

```
# Ww ( 1 char )    ##한 글자
# Wd ( 1 decimal )
# Ws ( 1 space )
```

```
# + ( 1, ..., N )
# ? ( 0, 1 )
# * ( 0, 1, .. N )
```

```
# Wd{N} ( 숫자가 N개 나온다. )    ##숫자가 3자리인 것
# Wd{N,M} ( 숫자가 N~M개 나온다 )    ## 숫자가 N~M자리인 것
```

```
text = """
옛날 옛적에 김길동라는 사람이 살았습니다.
그에게는 5형제가 있었는데, 김찬동, 김찬식, 김찬이, 김찬혁, 김찬민 이렇게 5명 있었습니다.
그리고 그는 결혼을 해서 김수용, 김지용, 김무용 3남매를 낳고 행복하게 잘 살았습니다.
"""
```

```
pattern = re.compile('김찬Ww')
pattern
```

```
➡ re.compile(r'김찬Ww', re.UNICODE)
```

```
brother = pattern.findall(text)
brother
```

```
➡ ['김찬동', '김찬식', '김찬이', '김찬혁', '김찬민']
```

```
pattern = re.compile('김.용')
#pattern = re.compile('김Ww영')
#둘은 동일
```

```
children = pattern.findall(text)
children
```

```
➡ ['김수용', '김지용', '김무용']
```

```
brother = set(brother)
brother
```

```
➡ {'김찬동', '김찬민', '김찬식', '김찬이', '김찬혁'}
```

```
brother = sorted(brother)
brother
```

```
➡ ['김찬동', '김찬민', '김찬식', '김찬이', '김찬혁']
```

```
text = 'A sky, a dragonfly aand a butterfly!'
```

```
pattern = re.compile('Ww+fly')
pattern.findall(text)
```

```
➡ ['dragonfly', 'butterfly']
```

▼ 핸드폰 번호에 대한 파싱

```
\d{2,3}[-]? \d{3,4}[-]? \d{4}
```

```
text = """
010-5670-3847    # space, -, . => []
010 5670 3847
010.5670 3847
"""
```

```
pattern = re.compile('Wd{3}[-W.]?Wd{4}[-W.]?Wd{4}') #.이나 공백이나 -
```

```
pattern.findall(text)
```

```
↳ ['010-5670-3847', '010 5670 3847', '010.5670 3847']
```

```
text = """
010-5670-3847    # space, -, . => []
옛날에는 011-1052-3847 이었는데..
010 5670 3847
010.-5670 3847
사는 동네가 금영아파트 103동112호
그리고, 사무실번호는 02-3450-4012이고
우편번호는 100-711, 청파로 887번지
"""

pattern = re.compile("Wd{2,3}[-W.]{1,2}?Wd{3,4}[-W.]?Wd{4}")
# ? : 숫자가 {3}거나 {3,4} 이거나 {4}이거나
```

```
pattern.findall(text)
```

```
↳ ['010-5670-3847',
    '011-1052-3847',
    '010 5670 3847',
    '010.-5670 3847',
    '02-3450-4012']
```

▼ 주민번호에 대한 파싱

- 뒷자리를 암호화 시키기
- 숫자 6자리 - 숫자 7자리

```
text = """
김부영 990811-3097532
김오영 011108-3123441
김지영 111009-4085122
"""
```

```
pattern = re.compile("Wd{6}-?Wd{7}") # ?Wd
# -가 있을 수도 있고...볼을 수도 있고...
```

```
pattern.findall(text)
```

```
↳ ['990811-3097532', '011108-3123441', '111009-4085122']
```

```
pattern = re.compile("Wd{6}-Wd{7}")
# 정규표현식 GROUP
# 1. 생년월일 그룹 <birth>
# 2. 주민등록번호 뒷자리 그룹 <secret>
```

```
pattern = re.compile("(?P<birth>Wd{6})-(?P<secret>Wd{7})")
# ?P<>로 패턴 파라미터
```

```
pattern.findall(text)
```

```
↳ [('990811', '3097532'), ('011108', '3123441'), ('111009', '4085122')]
```

Wg로 불러오기

```
result = pattern.sub("Wg<birth>-*****",text)
print(result)
```

```
↳
김부영 990811-*****
김오영 011108-*****
김지영 111009-*****
```

```
pattern = re.compile('(?P<name>Ww{3}) (?P<birth>Wd{6})-(?P<secret>Wd{7})')
```

```
pattern.findall(text)
```

```
↳ [('김부영', '990811', '3097532'),
    ('김오영', '011108', '3123441'),
    ('김지영', '111009', '4085122')]
```

```
print(text)
```

```
↳
김부영 990811-3097532
김오영 011108-3123441
김지영 111009-4085122
```

암호화

```
result = pattern.sub("Wg<name>(Wg<birth>-*****)",text)
result
```

```
↳ 'Wn    김부영(990811-*****)Wn    김오영(011108-*****)Wn    김지영(111009-*****)Wn'
```

```
result = result.split('Wn')
result
```

```
↳ ['',
    '    김부영(990811-*****)',
    '    김오영(011108-*****)',
    '    김지영(111009-*****)',
    '']
```

```
result.pop(0)
result
```

```
↳
```

```
[ '    김부영(990811-*****) ',  
  '    김오영(011108-*****) ',  
  '    김지영(111009-*****) ',  
  '' ]
```

```
result.pop(-1)  
result
```

```
↳ [ '    김부영(990811-*****) ',  
    '    김오영(011108-*****) ',  
    '    김지영(111009-*****) ' ]
```

```
for idx, val in enumerate(result) :  
    val = val.replace(" ", "")  
    print(idx, val)
```

```
↳ 0 김부영(990811-*****)  
   1 김오영(011108-*****)  
   2 김지영(111009-*****)
```

▼ Quiz

- Email 패턴 : ID + @ + URL

```
# 작업 폴더 확인  
%ls data
```

```
with open('./data/emails.txt', encoding='utf-8') as fp :  
    data = fp.read()  
    fp.close()
```

```
data
```



```
'IdWtDocNumberWtMetadataSubjectWtMetadataToWtMetadataFromWtSenderPersonIdWtMetadataDateSent'
```

```
pattern = re.compile("Ww[a-zA-Z0-9]{0,12}[ -W.]?Ww@.?Ww{0,12}?.?Ww{0,12}")
```

```
result = pattern.findall(data)
```

```
result
```



['Sullivan11@state.gov',
'Sullivan11@state.gov',
'_Brose@armed-servic',
'MillsCD@state.gov',
'MillsCD@state.gov',
'hdr22@clintonernail.com',
'hdr22@clintonemail .corn',
'MillsCD@state.gov',
'MillsCD@state.gov',
'snipead@state.gov',
'hrod17@clintonemail.com',
'hrod17@clintonemail.com',
'Russorv@state.gov',
'hrod17@clintonernailcom',
'Russorv@state.gov',
'hrod17@clintonernailcom',
'Russorv@state.gov',
'MillsCD@state.gov',
'MillsCD@state.gov',
'hrod17@clintonemail.corn',
'hrod17@clintonemail.corn',
'Sullivanli@stategov>',
'Sullivanli@stategov>',
'SullivanJJ@state.gov',
'SullivanJJ@state.gov',
'Reinesp@stategov;',
'Reinesp@stategov;',
'MillsCD@state.gov',
'MillsCD@state.gov',
'snipead@state.gov',
'KohlifF@state.gov',
'KohlifF@state.gov',
'ShermanWR@state.gov',
'ShermanWR@state.gov',
'ShermanWR@state.gov',
'Sullivanii@state.gov',
'Sullivanii@state.gov',
'Sullivanii@state.gov',
'Sullivanii@state.gov',
'ldr22@dintonemail.com',
'sullivanj.j@state.gov',
'hrod17@clintonernail.com',
'esullivanj.j@state.gov',
'hrod17@clintonernail.com',
'esullivanj.j@state.gov',
'Russorv@state.gov',
'hrod17@clintor iernail',
'Russont@state.gov',
'hrod17@clintor iernail',
'Russont@state.gov',
'Russorv@state.gov',
"Russorv@stategov",
"Russorv@stategov",
'sullivanj.j@state.gov',
'sullivanj.j@state.gov',
'hrod17@clintonemail.com',
'hrod17@clintonemail.com',
'sullivanii@state.gov'.

sullivanj.j@state.gov,
'sullivanj.j@state.gov',
'sullivanj.j@state.gov',
'HDR22@clintonemail.com',
'HDR22@clintonernallycorn',
'SullivanD@state.gov',
'Sullivanit@state.gov',
'Sullivanit@state.gov',
'HDR22@clintonemail.com',
'MillsCD@state.gov',
'MillsCD@state.gov',
'AbedinH@state.gov',
'AbedinH@state.gov',
'BitterR@state.gov',
'SullivanJJ@state.gov',
'MillsCD@state.gov',
'MillsCD@state.gov',
'Russorv@state.gov',
'Russorv@state.gov',
'hrod17@clintonemail.com',
'SullivanJJ@state.gov',
'hrod17@clintonemail.com',
'Russorv@state.gov',
'SullivanJJ@state.gov',
'Sullivarill@state.gov',
'Sullivarill@state.gov',
'MillsCD@state.gov',
'MillsCD@state.gov',
'MillsCD@state.gov',
'MillsCD@state.gov',
'AbedinH@state.gov',
'AbedinH@state.gov',
'MillsCD@state.gov',
'MarshallCP@state.gov',
'HOR22@clintonemail.com',
'hrod17@clintonemail.com',
'hrod17@clintonemail.com',
'Sullivan13@state.gov',
'sullivanj.j@state.gov',
'sullivanj.j@state.gov',
'hrod17@clintonemail.com',
'hrod17@clintonemail.com',
'sullivanj.j@state.gov',
'11@state.gov',
'MillsCD@state.gov',
'MillsCD@state.gov',
'sullivanj.j@state.gov',
'sullivanj.j@state.gov',
'hrod17@clintonemail.com',
'hrod17@clintonemail.com',
'sullivanj.j@state.gov',
'SullivanAt@state.gov',
'sullivanj.j@state.gov',
'gl@state.gov',
'hrod17@clintonemail.com',
'hrod17@clintonemail.com',
'gl@state.gov',
'MarshallCP@state.gov',
'MarshallCP@state.gov',

'hdr22@clintonemail.corn',
'hdr22@clintonemail.corn',
'HDR22@clintonemail.com',
'SullivanJJ@state.gov',
'SullivanJJ@state.gov',
'HDR22@clintonemail.com',
'BurnsW.1@state.gov',
'BurnsW.1@state.gov',
'sullivanjj@state.gov',
'sullivahu@state.gov',
'hrod17@clintonernail.com',
'hrod17@clintonernail.com',
'sullivahu@state.gov',
'HDR22@clintonemail.com',
'MillsCD@state.gov',
'Sullivan@state.gov',
'Sullivan@state.gov',
'HDR22@clintonemail.com',
'MillsCD@state.gov',
'MillsCD@state.gov',
'HDR22@clintonemail.com',
'AbedinH@state.gov',
'Russorv@state.gov',
'hrod17@clintonemail.com',
'Russorv@state.gov',
'hrod17@clintonemail.com',
'Russorv@state.gov',
'MillsCD@state.gov',
'MillsCD@state.gov',
'Sullivan@state.gov',
'Sullivan@state.gov',
'MillsCD@state.gov',
'mailto:Taclips@state.gov',
'MillsCD@state.gov',
'mailto:Taclips@state.gov',
'sullivanjj@state.gov',
'sullivanjj@state.gov',
'hrod17@clintonemail.com',
'hrod17@clintonemail.com',
'sullivanjj@state.gov',
'Sullivan11@state.gov',
'HOR22@clintonemail.com',
'Hurna@clintonemaii.com',
'Sullivanii@state.gov',
'HDR22@clintonemail.corn',
'Huma@clintonernail.com',
'SullivanJJ@state.gov',
'Sullivanii@stategov>',
'Sullivanii@stategov>',
'Sullivanii@state.govj',
'HDR22@clintonemail.coml',
'Hurna@clintonernail.corn',
'Sullivanii@state.gov',
'HOR22@clintonemail.corni',
'SullivanJJ@state.govj',
'Russorv@state.gov',
'hrod17@clintonemail.com',
'Russorv@state.gov',


'hrod17@clintonemail.com' ,
'Russorv@state.gov' ,
'Russorv@state.gov' ,
'Russoiv@state.gov' ,
'hrod17@clintonemail.com' ,
'hrod17@clintonemail.com' ,
'Russoiv@state.gov' ,
'HanleyMR@state.gov' ,
'HanleyMR@state.gov' ,
'HDR22@clintonemail.comj' ,
'HanleyiVIR@state.govj' ,
'HDR22@clintonemail.corn' ,
'HDR22@clintonernail.corn' ,
'huma@clintonemail.com' ,
'huma@clintonemail.com' ,
'MillsCD@state.gov' ,
'MillsCD@state.gov' ,
'sullivanjj@state.gov' ,
'sullivanjj@state.gov' ,
'sullivanjj@state.gov' ,
'MillsCD@state.gov' ,
'MillsCD@state.gov' ,
'Sullivanil@state.gov' ,
'Sullivanil@state.gov' ,
'SullivanJJ@state.gov' ,
'SullivanJJ@state.gov' ,
'HDR22@clintonernail.com' ,
'abedinh@state.gov' ,
'abedinh@state.gov' ,
'abedinh@state.gov' ,
'Abedinfi@state.gov' ,
'Sullivan33@state.gov' ,
'OpsNewsTicker@state.gov' ,
'MillsCD@state.gov' ,
'MillsCD@state.gov' ,
'thecable@foreignpolicy.com' ,
'millscd@state.gov' ,
'milliscd@stategov"' ,
'hrod17@clintonemail.com' ,
'hrod17@clintonemail.com' ,
'milliscd@stategov"' ,
'MillsCD@state.gov' ,
'thecable@foreignpolicy.comj' ,
'MillsCD@state.gov' ,
'MillsCD@state.gov' ,
'MillsCD@state.gov' ,
'MillsCD@state.gov' ,
'paclips@state.gov']

```
result = list(set(result))
```

```
for val in result :  
    val = val.replace("ㅈ", "")  
    val = val.replace("ㅊ", "")  
    result_.append(val)
```

```
with open('./data/emails2.txt', 'w') as fp:  
    fp.write()
```

val

 {'.', '@', 'a', 'c', 'e', 'g', 'i', 'l', 'o', 'p', 's', 't', 'v'}