정규식 활용

2019년 6월 3일 월요일 오후 2:58

이메일 호스트를 추출하는 다양한 방법

이메일 호스트를 추출하는 다양한 방법에 대해 다시 한 번 살펴보겠습니다.

먼저 문자열 메소드를 사용하는 방법입니다. find 메소드와 리스트 슬라이싱을 활용해 다음과 같이 찾을 수 있었습니다.

```
data = 'From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008'
atpos = data.find('@')
print(atpos)
# 21
sppos = data.find(' ',atpos)
print(sppos)
# 31
host = data[atpos+1 : sppos]
print(host)
# uct.ac.za
```

다음은 split 메소드를 활용하는 방법입니다. 공백 문자를 기준으로 1차적으로 문자열을 나누고, '@'이 포함되어있는 문자열을 '@'을 기준으로 나누었습니다.(중복 split)

```
line = 'From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008'
words = line.split()
email = words[1]
pieces = email.split('@')
print(pieces[1])
# 'uct.ac.za'
```

이번엔 정규식을 사용한 방법입니다. 여기에서 '^ '는 공백문자가 아닌 문자를 의미하며, '^'가 중간에 들어갈 경우 뒤에 오는 문자를 제외한 패턴을 의미합니다.

```
import re
lin = 'From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008'
y = re.findall('@([^ ]*)',lin)
print(y)
# ['uct.ac.za']
```

여기에서 조금 더 정교하게 패턴을 추출하려면 다음과 같이 코드를 작성할 수도 있습니다.

```
import re
lin = 'From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008'
y = re.findall('^From .*@([^ ]*)',lin)
print(y)
# ['uct.ac.za']
```

종합 예제 : 패턴 추출 및 최댓값 찾기

지금까지 배운 내용들을 종합하면 텍스트 파일에서 특정 패턴을 찾고, 그 패턴들 중 가장 큰 값이 어떤 것인지 찾을 수 있습니다.

```
import re
hand = open('mbox-short.txt')
numlist = list()
for line in hand:
```

```
    line = line.rstrip()
    stuff = re.findall('^X-DSPAM-Confidence: ([0-9.]+)', line)
    if len(stuff) != 1 :  continue
    num = float(stuff[0])
    numlist.append(num)
print('Maximum:', max(numlist))
```

예외 문자(Escape Character)

지금까지 다양한 특수 문자를 배웠습니다. 그런데 만약 그런 특수 문자로 이루어진 패턴을 찾으려면 어떻게 해야 할까요?

그럴 때는 역슬래시(\)를 사용하면 됩니다.

예를 들어, '$' 문자가 포함된 패턴을 찾고 싶을 때는 다음과 같이 코드를 작성할 수 있습니다.

```
import re
x = 'We just received $10.00 for cookies.'
y = re.findall('\$[0-9.]+',x)
print(y)
# ['$10.00']
```

출처: <https://www.edwith.org/python-network-data/lecture/24453/>