Regular Expression

Index

Regex

정규표현식

개요

- 일정한 규칙(패턴)을 가진 문자열을 추출, 변경할 시 사용하는 식

- html 소스에서 링크주소만 가져오는 패턴 예제
- https://regexr.com/4c44n

```
<ul class="type06_headline">
<li>
    <dl>
        <dt>
            <a href="https://news.naver.com/main/read.nhn?
mode=LS2D&amp;mid=shm&amp;sid1=101&amp;sid2=259&amp;oid=015&amp;aid=0004123580">
                '취준생 선호 1위'에도 아쉬워 한 국민은행
            </a>
        </dt>
        <dd>
            <span class="lede">뉴스카페 복지제도·워라밸 이유로 꼽자 "편한 직장으로만 보일라" 고
</span>
            <span class="writing">한국경제</span>
            <span class="date is_outdated">3시간전 </span>
        </dd>
    </dl>
```

re

- 파이썬에서 정규표현식을 사용할 수 있게 하는 라이브러리
- import re

re

re module

검색

- match(), 문자열 처음부터 매치여부를 조사, 객체 리턴
- search(), 문자열 전체를 조사, 처음 검색된 최초 문자열 객체 리턴
- findall(), 매치되는 모든 문자열 리스트로 리턴
- finditer(), 매치되는 모든 문자열의 반복가능한 객체로 리턴

```
import re

text = "I like orange! I love orange!"
result = re.match("orange", text)
print(result)
```

결과 : None

정규표현식 리턴 객체의 메서드

- group(), 매치된 문자열의 리턴
- start(), 매치된 문자열의 시작위치 리턴
- end(), 매치된 문자열의 끝 위치 리턴
- span(), 매치된 문자열의 (시작, 끝)에 해당하는 튜플 리턴

```
import re

text = "orange! I love orange!"
result = re.match("orange", text)
print(result)
```

결과 : <re.Match object; span=(0, 6), match='orange'>

```
print(result.group())
print(result.start())
print(result.end())
print(result.span())
```

re.search()

- 문자열 전체를 조사, 처음 검색된 최초 문자열 객체 리턴

```
import re

text = “I like orange! I love orange!”
result = re.search(“orange”, text)
print(result)
```

결과 : <re.Match object; span=(7, 13), match='orange'>

re.findall()

- 매치되는 모든 문자열 리스트로 리턴

```
result = re.findall(“orange”, text)
print(result)
```

결과 : ['orange', 'orange']

re.finditer()

- 매치되는 모든 문자열의 반복가능한 객체로 리턴

```
result = re.finditer(“orange”, text)
print(result)
```

결과 : <callable_iterator object at 0x000002A7585A54B0>

실습

- 해당 기사에서 네이버가 총 몇 번 나오는지 정규표현식을 이용 파이썬에서 확인
- https://regexr.com/4gntf

re

정규표현식

문자

- 표현식 : orange https://regexr.com/4cfc5

- 표현식 : like orange https://regexr.com/4cfce

^, $

- ^ 문자열의 시작 https://regexr.com/4cfct
- 표현식 : ^I like

- $ 문자열의 끝 https://regexr.com/4cfd3
- 표현식 : orange!$

\ 특수문자

- 표현식 : \$ https://regexr.com/4cfdu

- \ ^ $ * + ? . [ ] ( ) | : , - 등

. 모든 문자

- 표현식 : . https://regexr.com/4cfea
- 문자가 하나씩 추출됨

- 표현식 : https://regexr.com/4cfe7
- 4개씩 추출

[] 범위 판단

- 표현식 : [orn] https://regexr.com/4cfih
- 문자 하나씩 추출

- 표현식:[orn][orn]. https://regexr.com/4cfik

[] 범위 판단 응용

- 표현식 : [0-9] https://regexr.com/4cfiq
- 표현식:[A-Za-z] https://regexr.com/6t03b
- 표현식:[가-힣] https://regexr.com/4cfj6
- 표현식:[^A-Za-z] https://regexr.com/4cfjf

() 그룹

- 표현식 : (orange) https://regexr.com/4cfkv
- 표현식:(orange|like) https://regexr.com/4cfl2

? - 없거나 한 개, * - 0개 이상, + - 1개 이상

- 표현식 : a.c https://regexr.com/4cfl8
- 표현식 : a.?c https://regexr.com/4cflk
- 표현식 : ab*c https://regexr.com/4cflq
- 표현식 : ab+c https://regexr.com/4cflt

[]* []+

- 표현식 : [^]+ https://regexr.com/4cfmf
- 표현식 : a[bd]*c https://regexr.com/4cfmi

{} 개수

- 표현식 : .{5} https://regexr.com/6p3dt
- 표현식 : [abc]{3} https://regexr.com/6t03h

\d, \D, \w, \W, \s, \S

- \d = [0-9]
- \D = [^0-9]
- \w = [a-zA-Z0-9_]
- \W = [^a-zA-Z0-9_]
- \s = [\t\n\r\f\v] 공백, 탭, 라인피드, 캐리지리턴, 폼피드, 수직탭
- \S = [^ \t\n\r\f\v]

?<= – 전방탐색, ?= – 후방탐색

- 표현식 : oran(?=ge!)

https://regexr.com/4chn2

- 표현식 : (?<=ora)nge!

https://regexr.com/4chnn

?! 부정형

- 표현식 : ((?!
).)*<\/span>

https://regexr.com/6p3ef

파이썬에서 전화번호 찾기

- https://regexr.com/4chor

- 표현식 : [0-9]{3}-[0-9]{3,4}-[0-9]{4}

```
import re

numbers = """
010-2334-3234
02-302-3033
010-1321-4043
02-01-32
33-3303-3033
016-444-3042
"""
```

re

정규표현식

re.sub(정규표현식, 치환할 문자, 대상 문자)

- 대상 문자 내에서 정규표현식에 일치하는 문자를 치환할 문자로 변경

```
import re

text = """
수강해야하는 과목
(machine learning, deep learning)
수강한 과목
[python, django, web design]
"""
re.sub("\[.+\]", "", text)
```

실습 01

- 정상적인 이메일만 추출해주세요

https://regexr.com/4chri

결과 :
jkilee@gmail.com
kttredef@naver.com
adekik@best.kr
adefgree@korea.co.kr

실습 02

- 텍스트중에 <내용> 괄호로 묶여진 텍스트를 괄호 포함 모두 제거해주세요

https://regexr.com/4rdvb

결과:
안녕하세요 저는 홍길동입니다. 나이는 24살 세계 최고의 데이터 분석가가 되고싶습니다.

실습 03

1. 정규표현식을 이용 내용 을 각각 추출 https://regexr.com/4rdve
2. 추출된 항목에서 과 태그를 모두제거
3. 각각 총 3개의 항목을 리스트에 넣기

결과
[“네이버가 뉴스 서비스에 인공지능(AI)을 도입해 페이지 뷰(PV)를 늘리고 이용자를 끌어 모으고 있다. “,
“네이버는 5일 오전 서울 강남구 그랜드 인터컨티넨털호텔에서 AI 콜로키움 2019를 열고 이 같은 AI 성과와 전략을 소개했다.”,
“이날 기조연설에서김광현 네이버 서치앤클로바리더는 "AI 뉴스 추천 시스템인 에어스(AiRS)를 도입하면서뉴스 소비량이 확대되고 있다" 고 말했다.”]

심화 : 위의 1, 2 과정을 하나의 정규식으로 해결해보세요