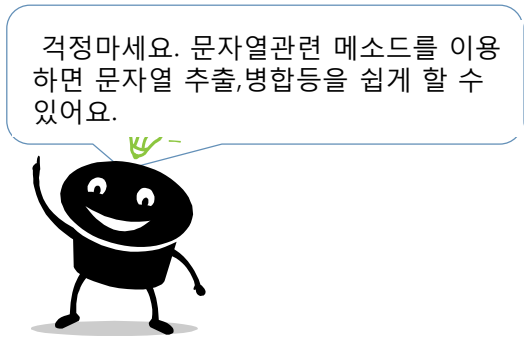
8. pandas –문자열처리

이번시간에 학습할 내용은..

문자열 처리를 위한 기본 메소드 ,정규표현식을 이용한 문자열처리, pandas의 문자열 메소드등을 학습.

1. 텍스트 처리관련 기본메소드



* 주요 내장 문자열 메소드

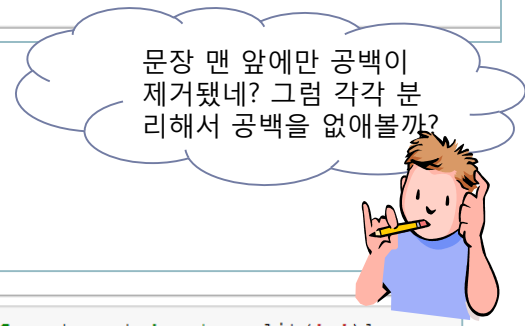
메소드	설명
Upper/lower	문자열 대문자로 변경/소문자로 변경
title	각 단어의 첫 문자를 대문자로 변경
count	문자열 출현 횟수 리턴
find	찾고자하는 단어의 첫글자 위치를 리턴(* 찾고자하는 단어가 없을 경우 -1리턴)
replace	문자열을 다른 문자열로 치환
Strip / rstrip / lstrip	문자열 좌우공백을 제거 / 오른쪽 문자열 공백 제거 / 왼쪽 문자열 공백 제거
split	구분자 기준으로 단어리스트로 분리
Ljust / rjust	지정된 문자열 길이에서 Ljust – 문자열 왼쪽정렬하고 남은 길이만큼 공백처리하여 리턴 Rjust – 문자열 오른쪽 정렬하고 남은 길이만큼 공백처리하여 리턴

```
In [3]: str=' i , like,cold brew coffee'
        str.split(',')

Out[3]: [' i ', ' like', 'cold brew coffee']
```

```
In [4]: a=str.strip()
        a

Out[4]: 'i , like,cold brew coffee'
```




```
In [5]: result=[str_out.strip() for str_out in str.split(',')]
        result

Out[5]: ['i', 'like', 'cold brew coffee']
```

str에서 coffee 문자열이 시작되는 위치를 알려줌

```
In [6]: a=str.find('coffee')
        a

Out[6]: 20
```

```
In [8]: a=str.index('coffee')
        a

Out[8]: 20
```

'in' 키워드를 통해서도 해당 문자열 존재여부를 알 수 있어요.

```
In [11]: a='coffee' in str
         a

Out[11]: True
```

str 에서 "coffee"를 "americano"로 변경

```
In [12]: str.replace("coffee","americano")

Out[12]: ' i , like,cold brew americano'
```

새로운 문자열객체가 생성

```
In [13]: print(str)
          i , like,cold brew coffee
```

str결과를 확인해보세요. str 내용이 변경되지는 않죠?

2. 정규표현식 이용한 문자열 처리

- 정규표현식 – 문자열 처리를 유연하게 하기위한 기법
- 파이썬에서는 정규표현식 이용하여 문자열을 처리하는 re모듈을 제공

* 참고 – 주요 정규푠식 정리

표현식	설명	\w	알파벳이나 숫자
^	문자열 시작	\W	알파벳이나 숫자를 제외한 문자
$	문자열 종료	\d	숫자[0~9]와 동일
*	앞 문자가 없을 수도 무한정 많을 수도 있음.	\D	숫자를 제외한 모든 문자
+	앞문자가 하나 이상	\특수문자	해당 특수문자를 의미
?	앞 문자 없거나 하나 있음	(?!)	대/소문자 구분하지 않음
[]	문자집합이나 범위를 나타내며, 두 문자 사이에는 – 기호로 범위표시		
{}	횟수 또는 범위를 나타냄		
()	소괄호 안의 문자를 하나의 문자로 인식		
|	패턴 안에서 or 연산을 수행할 때 사용		
\s	공백문자		
\S	공백문자가 아닌 나머지 문자		

```
In [1]: import re

In [8]: souce_a="dodo/mimi/solsol/lala,sisi"
        mat_1=re.compile('\/+')
        mat_1.split(souce_a)

Out[8]: ['dodo', 'mimi', 'solsol', 'lala,sisi']
```

\(백슬래쉬) 뒤에 특수문자 ➔\, 쉼표, \ / 슬래쉬 의미
+ ➔해당 규칙을 반복하라는 의미
➔정규표현식을 컴파일

정규표현식에 의해 문자열 분할

```
In [5]: souce_a="dodo/mimi/solsol/lala,sisi"
        re.split('\/+',souce_a)

Out[5]: ['dodo', 'mimi', 'solsol', 'lala,sisi']
```

이렇게 split 메소드안에 정규표현식을 함께 입력할 수 있어요.
그러면, 컴파일이 먼저 수행된 후 split이 실행돼요.

3. pandas에서 text분석을 위한 문자열 함수

➢ 데이터 분석을 위해 문자열 정제작업은 꼭 필요. 이를 간결하게 처리하기 위한 문자열 메소드제공

* 문자열 메소드

메소드	설명
cat	선택적인 구분자와 함께 요소별 문자열 이어붙임
contains	문자열이 패턴이나 정규표현식을 포함하는지 나타내는 불리언 배열 반환
count	일치하는 패턴의 개수를 반환
findall	각 문자열에 대해 일치하는 정규표현식의 전체 목록을 구함
join	Series외 각 요소를 주어진 구분자로 연결
match	주어진 정규표현식으로 각 요소에 대한 re.match를 수행하여 일치하는 그룹의 리스트로 반환
get	I번째 요소를 반환

Series 자료구조형 emailinfo 객체를 그림처럼 생성해요.

```
In [1]: import pandas as pd
        import numpy as np
        import re

In [2]: people_info={'홍길동':'mrhong@gmail.com','장길산':'roadmount3@naver.com',
                     '홍길순':'hks0900@naver.com','임꺽정':'tmplim1@gmail.com',
                     '무명순':np.nan}
        emailinfo=pd.Series(people_info)
        emailinfo

Out[2]: 무명순                      NaN
        임꺽정        tmplim1@gmail.com
        장길산     roadmount3@naver.com
        홍길동         mrhong@gmail.com
        홍길순        hks0900@naver.com
        dtype: object
```

```
In [9]: emailinfo.str.contains('naver')

Out[9]: 무명순       NaN
        임꺽정     False
        장길산      True
        홍길동     False
        홍길순      True
        dtype: object
```

'naver'문자열을 포함하는지 여부를 체크하여 논리형으로 보여줌

```
In [15]: p=r'[a-z0-9._%+-]+@[a-z0-9]+\.[a-z]{2,4}'
```

정규표현식 규칙을 생성

```
In [16]: emailinfo.str.findall(p,flags=re.IGNORECASE)
```

해당 규칙에 맞는 모든 목록을 표시해줌.

```
Out[16]: 무명순                          NaN
         임꺽정        [tmplim1@gmail.com]
         장길산     [roadmount3@naver.com]
         한영자                        []
         홍길동         [mrhong@gmail.com]
         홍길순        [hks0900@naver.com]
         dtype: object
```

요소별로 구분하고자 할 경우 그림처럼 괄호로 묶어주세요.

```
In [22]: p=r'([a-z0-9._%+-]+)@([a-z0-9]+)\.([a-z]{2,4})'
```

```
In [24]: emailinfo.str.findall(p,flags=re.IGNORECASE)
```

```
Out[24]: 무명순                            NaN
         임꺽정        [(tmplim1, gmail, com)]
         장길산     [(roadmount3, naver, com)]
         한영자                            []
         홍길동         [(mrhong, gmail, com)]
         홍길순        [(hks0900, naver, com)]
         dtype: object
```