15. 집합(set)

집합에 관련된 것들을 쉽게 처리하기 위한 자료형이다.

집합 자료형은 다음과 같이 set 키워드 혹은 { } 를 이용해 만들 수 있다.

중복을 허용하지 않는다.

순서가 없다 (unordered) --> indexing 을 허용 않는다.

교집합 (&), 합집합(|), 차집합(-)

In [1]:

```
1 s1 = set([1,2,3])
2 print(s1)
```

```
{1, 2, 3}
```

In [2]:

```
1 type(s1)
```

Out[2]:

```
set
```

In [3]:

```
1 s2 = {'a', 'b', 'c'}
```

In [4]:

```
1 s2[1]
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-4-8755b941c10d> in <module>
----> 1 s2[1]

TypeError: 'set' object is not subscriptable
```

In [5]:

```
1 s1 = set([1, 2, 3, 4, 5, 6])
2 s2 = set([4, 5, 6, 7, 8, 9])
```

교집합

In [6]:

```
1 s1 & s2
```

Out[6]:

{4, 5, 6}

In [7]:

```
1 s1.intersection(s2)
```

Out[7]:

{4, 5, 6}

합집합

In [8]:

```
1 s1 | s2
```

Out[8]:

{1, 2, 3, 4, 5, 6, 7, 8, 9}

In [9]:

```
1 s1.union(s2)
```

Out[9]:

{1, 2, 3, 4, 5, 6, 7, 8, 9}

차집합

In [10]:

```
1 s1 - s2
```

Out[10]:

{1, 2, 3}

In [11]:

```
1 s1.difference(s2)
```

Out[11]:

{1, 2, 3}

element 제거, 추가

In [12]:

```
s1.remove(2)
```

In [13]:

```
s1
```

Out[13]:

{1, 3, 4, 5, 6}

In [14]:

```
s1.add('added')
```

In [15]:

```
s1
```

Out[15]:

{1, 3, 4, 5, 6, 'added'}

여집합

In [16]:

```
A = {1,2}
B = {3, 4, 5}
Union = A | B

AC = Union.difference(A)

print(AC)
```

{3, 4, 5}

연습문제

자카드유사도 (Jaccard Similarity)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

두개의 문서간의 유사도

$$J(doc1, doc2) = \frac{doc1 \cap doc2}{doc1 \cup doc2}$$

In [9]:

```
doc1 = "apple banana everyone like likely watch card holder"
doc2 = "apple banana coupon passport love you"

tokenized_doc1 = doc1.split()
tokenized_doc2 = doc2.split()

print(tokenized_doc1)
print(tokenized_doc2)
```

['apple', 'banana', 'everyone', 'like', 'likely', 'watch', 'card', 'holder']
['apple', 'banana', 'coupon', 'passport', 'love', 'you']

두 문서의 전체 단어 집합

In [3]:

```
union = # CODE HERE
print(union)
print(len(union))
```

{'holder', 'like', 'coupon', 'everyone', 'likely', 'passport', 'apple', 'banana', 'you', 'wat
ch', 'card', 'love'}
12

두 문서에 공통으로 나오는 단어

In [4]:

```
inter = # CODE HERE
print(inter)
```

{'banana', 'apple'}

In [8]:

```
print("두 문서간의 Jaccard 유사도 : {:.2f} ".format(len(inter)/len(union)))
```

두 문서간의 Jaccard 유사도 : 0.17