BOW 직접 구현

[1] 0초

```
docs = ['오늘 동물원에서 코끼리 원숭이를 보고 코끼리 원숭이에게 먹이를 줬어',
        '오늘 동물원에서 원숭이에게 사과를 줬어']

# '오늘 동물원에서 코끼리 원숭이를 보고 코끼리 원숭이에게 먹이를 줬어 오늘 동물원에서 원숭이에게 사과를 줬어']
```

0초

```
doc_ls = []
for doc in docs:
    doc_ls.append(doc.split(' '))
doc_ls
```

```
[['오늘', '동물원에서', '코끼리', '원숭이를', '보고', '코끼리', '원숭이에게', '먹이를', '줬어'],
 ['오늘', '동물원에서', '원숭이에게', '사과를', '줬어']]
```

1. 유니크한 토큰 사전 구하기

[3] 0초

```
from collections import defaultdict

word2id = defaultdict(lambda : len(word2id))
# word2id

for doc in doc_ls:
  # print(doc)
  for token in doc:
    word2id[token]
    # print(token)
word2id
```

```
[3]  defaultdict(<function __main__.<lambda>()>,
                 {'오늘': 0,
                  '동물원에서': 1,
                  '코끼리': 2,
                  '원숭이를': 3,
                  '보고': 4,
                  '원숭이에게': 5,
                  '먹이를': 6,
                  '줬어': 7,
                  '사과를': 8})
```

2. BOW 구하기

```
[4]  1 import numpy as np
     2 BoW_ls = []
     3 for i, doc in enumerate(doc_ls):
     4   bow = np.zeros(len(word2id),dtype = int)
     5   for token in doc:
     6     bow[word2id[token]] +=1
     7   BoW_ls.append(bow.tolist())
     8   # print(BoW_ls)
```

```
[5]  1 BoW_ls
```

```
[[1, 1, 2, 1, 1, 1, 1, 1, 0], [1, 1, 0, 0, 0, 1, 0, 1, 1]]
```

```python
from IPython.core import display as ICD
import pandas as pd
sorted_vocab = sorted((value, key) for key, value in word2id.items())
print('sorted_vocab',sorted_vocab)

vocab = []
for v in sorted_vocab:
    vocab.append(v[1])
print('vocab',vocab)
for i in range(len(docs)) :
    print("문서{} : {}".format(i, docs[i]))
    ICD.display(pd.DataFrame([BoW_ls[i]], columns=vocab))
    print("\n\n")
```

```
sorted_vocab [(0, '오늘'), (1, '동물원에서'), (2, '코끼리'), (3, '원숭이를'), (4, '보고'), (5, '원숭이에게'), (6, '먹이를'), (7, '줬어'), (8, '사과를')]
vocab ['오늘', '동물원에서', '코끼리', '원숭이를', '보고', '원숭이에게', '먹이를', '줬어', '사과를']
문서0 : 오늘 동물원에서 코끼리 원숭이를 보고 코끼리 원숭이에게 먹이를 줬어
```

	오늘	동물원에서	코끼리	원숭이를	보고	원숭이에게	먹이를	줬어	사과를
0	1	1	2	1	1	1	1	1	0

```
문서1 : 오늘 동물원에서 원숭이에게 사과를 줬어
```

	오늘	동물원에서	코끼리	원숭이를	보고	원숭이에게	먹이를	줬어	사과를
0	1	1	0	0	0	1	0	1	1

sklearn

[13]
```
docs = ['오늘 동물원에서 코끼리 원숭이를 보고 코끼리 원숭이에게 먹이를 줬어',
        '오늘 동물원에서 코끼리 원숭이를 보고 코끼리 원숭이에게 먹이를 줬어 오늘 동물원에서 원숭이에게 사과를 줬어']
```

[15]
```
# 토큰빈도계산 : CountVectorizer
from sklearn.feature_extraction.text import CountVectorizer
#선언
count_vect = CountVectorizer()

BoW = count_vect.fit_transform(docs)
BoW.toarray()
```

```
array([[1, 1, 1, 0, 1, 1, 1, 1, 2],
       [2, 1, 1, 1, 2, 1, 2, 2, 2]])
```