Chapter01_ PyTorch_NLP_Basics_1.ipynb

pytorch를 이용한 자연어입문 · 2024. 9. 18. 14:53

TF-IDF 표현

Term-Frequency-Inverse-Document-Frequency

TF = Term-Frequency 단어의 등장횟수

= the, a -> 자주등장해도 의미가 없어서

IDF = Inverse-Document-Frequency

그 수치값을 깎자는 의미

여러 문서에서 공통적으로 나오는 단어들을

수치를 낮추기 위해서 쓰는것

N : 전체 문서 개수

Nw: 단어 w를 포함한 문서의 개수

```
N_the = 10
IDF(the) = log(11/11) + 1
N_apple = 2
IDF(apple) = log(11/3) + 1
```

$$IDF(w) = \log\left(\frac{N+1}{N_w+1}\right)+1$$

```
모든 문서에 등장 (즉,  Nw=N )이면  IDF(w)=0
반대로 한 문서에만 등장하면  IDF(w)=logN
```

아래그림

출처: https://wikidocs.net/31698

04-04 TF-IDF(Term Frequency-I...

이번에는 DTM 내에 있는 각 단어에 대한 중요도를 계산할 수 있는 TF-IDF 가중치에 대해서 알아보겠습니다. TF-IDF를 ...

wikidocs.net

아래그림)

문서에 따른 나온 단어의 빈도수

	과일이	길고	노란	먹고	바나나	사과	싶은	저는	좋아요
문서1	0	0	0	1	0	1	1	0	0
문서2	0	0	0	1	1	0	1	0	0
문서3	0	1	1	0	2	0	0	0	0
문서4	1	0	0	0	0	0	0	1	1

아래그림)

단어가 많이 나올수록 IDF값이 적어짐

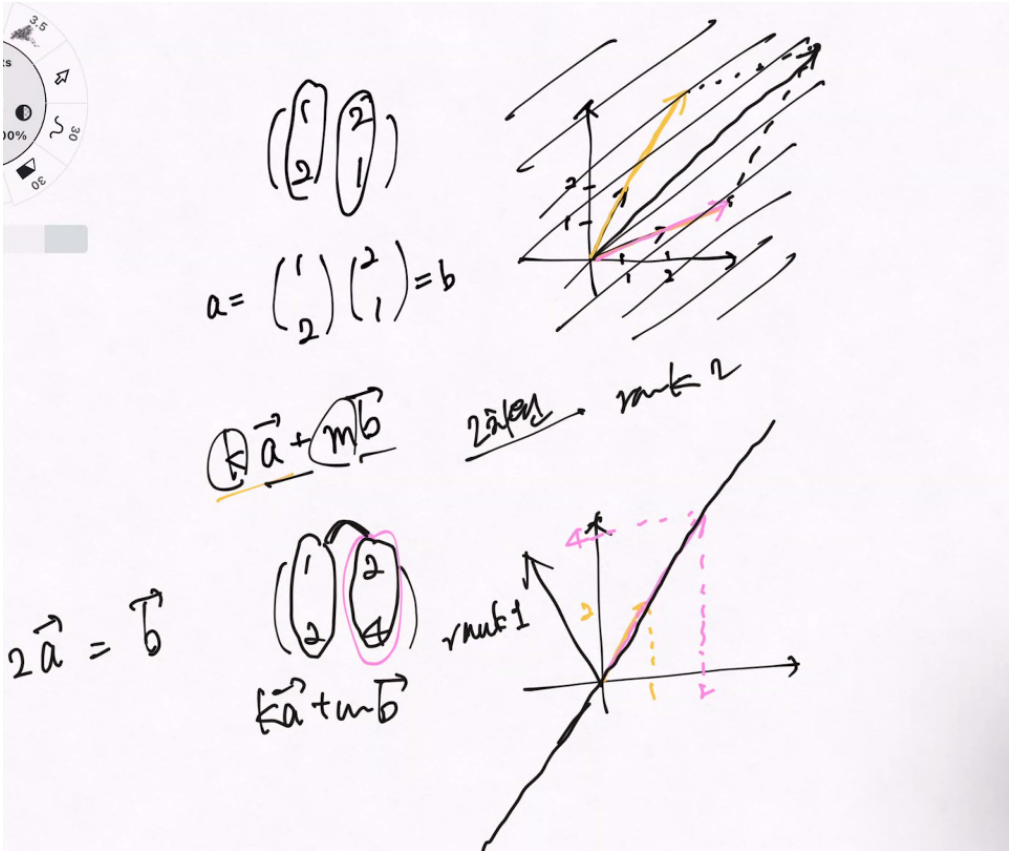
단어	IDF(역 문서 빈도)
과일이	ln(4/(1+1)) = 0.693147
길고	ln(4/(1+1)) = 0.693147
노란	ln(4/(1+1)) = 0.693147
먹고	ln(4/(2+1)) = 0.287682
바나나	ln(4/(2+1)) = 0.287682
사과	ln(4/(1+1)) = 0.693147
싶은	ln(4/(2+1)) = 0.287682
저는	ln(4/(1+1)) = 0.693147
좋아요	ln(4/(1+1)) = 0.693147

텐서

고차원 배열(하지만 1,2차원도 포함)
배열: 나열한것(1x4, 2x4... 고차원배열)

스칼라는 하나의 숫자입니다.
벡터는 숫자의 배열입니다.
행렬은 숫자의 2-D 배열입니다.
텐서는 숫자의 N-D 배열입니다.

행렬 A의 열들 중에서 선형 독립인 열들의 개수(2이면 2차원, 3이면 3차원)
[출처] [기초 선형대수] 행렬에서 Rank (랭크) 란?|작성자 PN



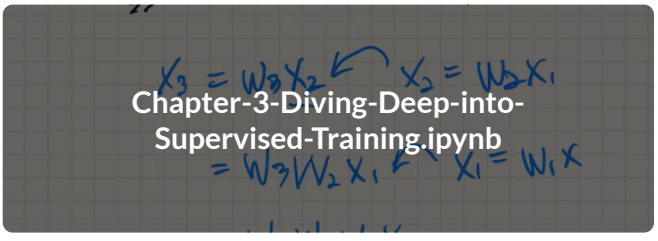

```
x = torch.arange(6).view(2,3)
describe(x)
describe(torch.cat([x, x], dim=0))  # concatenate
describe(torch.cat([x, x], dim=1))
describe(torch.stack([x, x]))

>>>
 torch.LongTensor
크기: torch.Size([2, 3])
값:
tensor([[0, 1, 2],
        [3, 4, 5]])

타입: torch.LongTensor
크기: torch.Size([4, 3])
값:
tensor([[0, 1, 2],
        [3, 4, 5],
        [0, 1, 2],
        [3, 4, 5]])
```

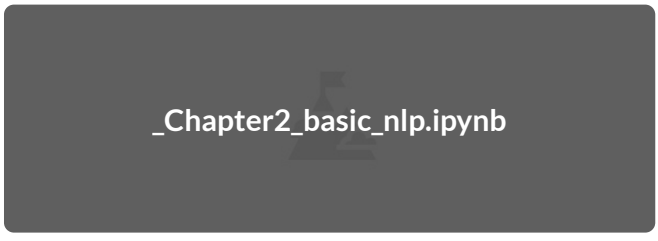
```
타입: torch.LongTensor
크기: torch.Size([2, 6])
값:
tensor([[0, 1, 2, 0, 1, 2],
        [3, 4, 5, 3, 4, 5]])

타입: torch.LongTensor
크기: torch.Size([2, 2, 3])
값:
tensor([[[0, 1, 2],
         [3, 4, 5]],

        [[0, 1, 2],
         [3, 4, 5]]])
```

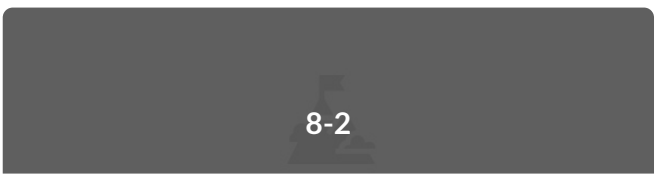

'pytorch를 이용한 자연어입문' 카테고리의 다른 글

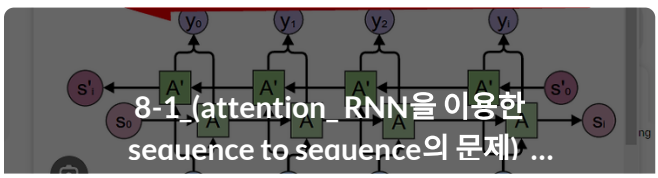
Chapter-3-Diving-Deep-into-Supervised-Training.ipynb (0)	2024.09.18
_Chapter2_basic_nlp.ipynb (1)	2024.09.18
8-2 (0)	2024.06.28
8-1_(attention_ RNN을 이용한 sequence to sequence의 문제)_ NMT_No_Sampling.ipynb (1)	2024.06.21
8-0 (0)	2024.06.14

관련글 관련글 더보기









자연어(NLP)

네이쳐2024 님의 블로그입니다.

구독하기 +

댓글 0

이름

비밀번호

내용을 입력하세요.

등록