NLP 입문

카운트 기반의 단어표현

23.01.26 유하영

text

순서가 상관없는 문장에 대한

텍스트 수치화

KoNLPy

scikit learn

BoW(Bag of Words)

단어의 순서를 고려하지 않고
단어의 출현 빈도에만 집중하는 텍스트 수치화 방법

doc1 = "정부가 발표하는 물가상승률과
소비자가 느끼는 물가상승률은 다르다."

vocab	BoW
정부	1
가	2
발표	1
하는	1
물가상승률	2
과	1
소비자	1
느끼는	1
은	1
다르다	1

DTM(Document-Term Matrix)

다수의 문서에 대한 BoW을 하나의 행렬로 표현한 것

DTM

vocab / Document index	과일이	길고	노란	먹고	바나나	사과	싶은	저는	좋아요
0	0	0	0	1	0	1	1	0	0
1	0	0	0	1	1	0	1	0	0
2	0	1	1	0	2	0	0	0	0
3	1	0	0	0	0	0	0	1	1

한계

1. 희소표현으로 인한 공간낭비
2. 불용어로 인한 문제

TF-IDF

단어의 빈도와 역 문서 빈도를 사용하여
단어들마다 중요한 정도에 따라서 가중치를 부여하는 방법

tf(d,t) : 특정 문서 d에서의 특정 단어 t의 등장 횟수

df(t) : 특정 단어 t가 등장한 문서의 수

idf(d,t) : df(t)에 반비례하는 수

$$idf(d,t) = log(\frac{n}{1+df(t)})$$

TF-IDF

$$TF - IDF = DTM \times IDF$$

DTM

vocab / Document index	과일이	길고	노란	먹고	바나나	사과	싶은	저는	좋아요
0	0	0	0	1	0	1	1	0	0
1	0	0	0	1	1	0	1	0	0
2	0	1	1	0	2	0	0	0	0
3	1	0	0	0	0	0	0	1	1

IDF

vocab	IDF
과일이	0.693147
길고	0.693147
노란	0.693147
먹고	0.287682
바나나	0.287682
사과	0.693147
싶은	0.287682
저는	0.693147
좋아요	0.693147

doc0 – 먹고 싶은 사과

doc1 – 먹고 싶은 바나나

doc2 – 길고 노란 바나나 바나나

doc3 – 저는 과일이 좋아요

vocab / Document index	과일이	길고	노란	먹고	바나나	사과	싶은	저는	좋아요
0	0.000000	0.000000	0.000000	0.287682	0.000000	0.693147	0.287682	0.000000	0.000000
1	0.000000	0.000000	0.000000	0.287682	0.287682	0.000000	0.287682	0.000000	0.000000
2	0.000000	0.693147	0.693147	0.000000	0.575364	0.000000	0.000000	0.000000	0.000000
3	0.693147	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.693147	0.693147

감사합니다