ICT이노베이션스퀘어 AI복합교육 고급 언어과정

자연어처리를 위한
Bag of Words
Term Frequency - Inverse Document Frequency

현청천

2021.04.19

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

What is Bag of Words

나는	0
학생	1
입니다	2
당신은	3
수학	4
선생님	5

나는	학생	입니다
1	0	0
0	1	0
0	0	1
0	0	0
0	0	0
0	0	0

One-Hot

나는	학생	입니다
1	1	1
1	1	1
1	1	1
0	0	0
0	0	0
0	0	0

BoW

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

What is Bag of Words

나는	0
학생	1
입니다	2
당신은	3
수학	4
선생님	5

당신은	수학	선생님	입니다
0	0	0	0
0	0	0	0
0	0	0	1
1	0	0	0
0	1	0	0
0	0	1	0

One-Hot

당신은	수학	선생님	입니다
0			
0			
1			
1			
1			
1			

BoW

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

What is Bag of Words

나는	0
학생	1
입니다	2
당신은	3
수학	4
선생님	5

나는	학생	입니다
1	0	0
0	1	0
0	0	1
0	0	0
0	0	0
0	0	0

당신은	수학	선생님	입니다
0	0	0	0
0	0	0	0
0	0	0	1
1	0	0	0
0	1	0	0
0	0	1	0

One-Hot

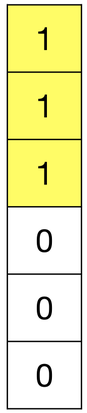
나는 학생 입니다
1
1
2
1
1
1

나는 학생 입니다
당신은 수학 선생님 입니다

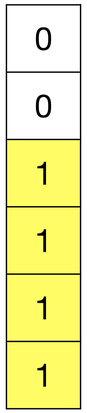
BoW

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

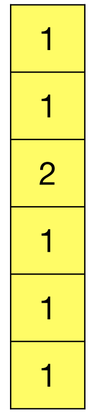
What is Bag of Words



나는 학생 입니다
입니다 학생 나는
학생 입니다 나는



당신은 수학 선생님 입니다
수학 당신은 입니다 선생님
선생님 입니다 당신은 수학



나는 학생 입니다
당신은 수학 선생님 입니다

입니다 수학 당신은
선생님 학생 입니다 나는

BoW는 단어의 순서를 알 수 없음

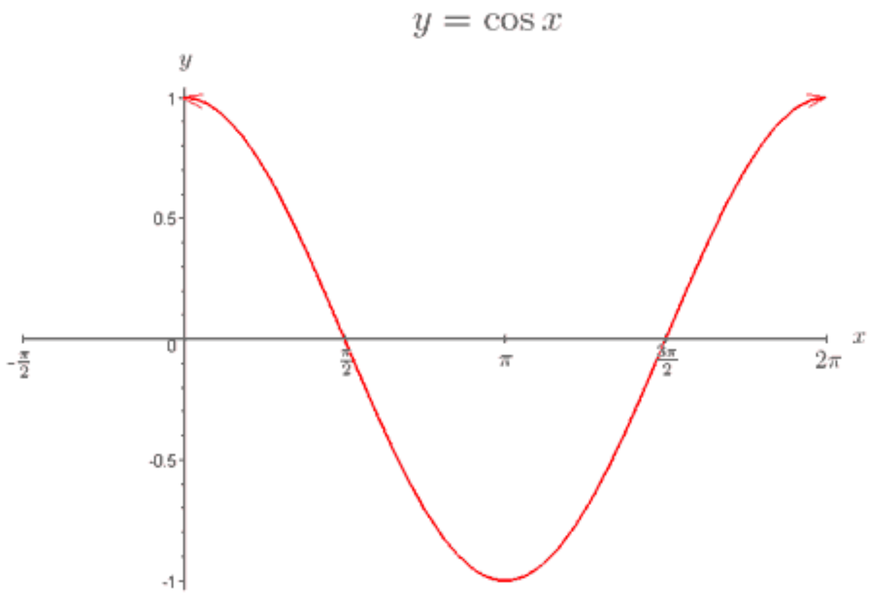
@2021 cchyun@gmail.com., Ltd All Rights Reserved.

Bag of Words Similarity

1
1
1
0
0
0

$$cos\theta = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

0
0
1
1
1
1



단어의 분포가 비슷하면 두 벡터의 각이 작아짐 - $cos\theta$ 가 커짐

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

Bag of Words Similarity

나는	0
학생	1
입니다	2
좋은	3
선생님	4
당신은	5
매우	6

$\vec{a}$	$\vec{b}$	$\vec{c}$
1	1	0
1	0	0
1	1	1
0	1	1
0	1	1
0	0	1
0	0	1

나는 학생 입니다

나는 좋은 선생님 입니다

당신은 매우 좋은 선생님 입니다

$$cos\theta = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

$\|\vec{a}\| = \sqrt{3}$

$\|\vec{b}\| = 2$

$\|\vec{c}\| = \sqrt{5}$

$\vec{a} \cdot \vec{b} = 2$

$\vec{b} \cdot \vec{c} = 3$

$\vec{c} \cdot \vec{a} = 1$

$\cos\theta_{\vec{a}\vec{b}} = \frac{1}{\sqrt{3}}$

$\cos\theta_{\vec{b}\vec{c}} = \frac{3}{2\sqrt{5}}$

$\cos\theta_{\vec{c}\vec{a}} = \frac{1}{\sqrt{15}}$

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

What is Term Frequency

영희는 매우 영어를 좋아한다
영희는 영어를 잘한다

철수는 수학을 매우 좋아한다
철수는 수학을 매우 매우 잘한다

$tf(d, w)$

	영희는	영어를	매우	좋아한다	잘한다	철수는	수학을
Doc #1	2	2	1	1	1	0	0
Doc #2	0	0	3	1	1	2	2

각 문서당 발생하는 단어수가 많은 것이 문서의 주제일 가능성이 높음

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

What is Inverse Term Frequency

$$idf(D, w) = \log \frac{N}{df(D, w)}$$

$df(d, w)$

	영희는	영어를	매우	좋아한다	잘한다	철수는	수학을
	1	1	2	2	2	1	1

$\frac{N}{df(d, w)}$

	영희는	영어를	매우	좋아한다	잘한다	철수는	수학을
	2	2	1	1	1	2	2

$\log \frac{N}{df(d, w)}$

	영희는	영어를	매우	좋아한다	잘한다	철수는	수학을
	0.6932	0.6932	0	0	0	0.6932	0.6932

특성 문서에만 발생하는 단어는 문서의 주제일 가능성이 높음

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

What is Term Frequency - Inverse Term Frequency

$$tfidf(d, w) = tf(d, w) \times idf(D, w)$$

$tfidf(d, w)$

	영희는	영어를	매우	좋아한다	잘한다	철수는	수학을
Doc #1	1.3863	1.3863	0	0	0	0	0
Doc #2	0	0	0	0	0	1.3863	1.3863

특성 문서에만 발생하는 단어는 문서의 주제일 가능성이 높음

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

감사합니다.

@2021 cchyun@gmail.com., Ltd All Rights Reserved.