

유사성 측정 방법

유클리드 거리

Euclidean Distance

코사인 유사도

Cosine Similarity

유클리드 거리

Euclidean Distance

두 점 사이의 거리 계산 공식

유클리드 거리

Euclidean Distance

인공지능 분야

두 가지 개체의 속성 값들이 여러 개일 경우,
이들 속성 값들에 의한 두 개체 사이의 유사도 측정

유클리드 거리

Euclidean Distance

인공지능 분야

두 벡터 간의 유사성 측정

유클리드 거리

Euclidean Distance

두 점 A, B

$A=(A_1, A_2, \dots, A_n), B=(B_1, B_2, \dots, B_n)$

$$D(A, B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \dots + (A_n - B_n)^2} = \sqrt{\sum_{i=1}^n (A_i - B_i)^2}$$

유클리드 거리

Euclidean Distance

문서1 : 컴퓨터가 좋아요
문서2 : 컴퓨터가 좋아요, 좋아요
문서3 : 컴퓨터가 싫어요

Bags of Words
(BoW) 표현

	컴퓨터	좋아요	싫어요
문서1	1	1	0
문서2	1	1	0
문서3	1	0	1

Euclidean
Distance

$D(\text{문서1}, \text{문서2}) < D(\text{문서1}, \text{문서3})$

유클리드 거리

Euclidean Distance

한계

문서의 방향성(의미) 뿐 아니라 문서의 길이도 유사도에 영향을 끼침

문서1 : 컴퓨터가 좋아요

문서2 : 컴퓨터가 좋아요, 컴퓨터가 좋아요

문서3 : 컴퓨터가 싫어요

Bags of Words
(BoW) 표현

	컴퓨터	좋아요	싫어요
문서1	1	1	0
문서2	2	2	0
문서3	1	0	1

Euclidean
Distance

$D(\text{문서1}, \text{문서2}) > D(\text{문서1}, \text{문서3})$

코사인 유사도

Cosine Similarity

벡터의 크기는 고려하지 않고
두 벡터 간 각도만 고려

코사인 유사도

Cosine Similarity

방향이 얼마나 유사한지를 $[-1, 1]$ 의 값으로 표현

1 : 0° \rightarrow 방향 일치

0 : 90°

-1 : 180° \rightarrow 방향 반대

코사인 유사도

Cosine Similarity

$$\text{Cosine Similarity} = \cos(\theta) = \frac{A \circ B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

코사인 유사도

Cosine Similarity

한계 극복

문서의 길이로부터 독립적인 유사도 계산 가능

문서1 : 컴퓨터가 좋아요

문서2 : 컴퓨터가 좋아요, 컴퓨터가 좋아요

문서3 : 컴퓨터가 싫어요

Bags of Words
(BoW) 표현

	컴퓨터	좋아요	싫어요
문서1	1	1	0
문서2	2	2	0
문서3	1	0	1

Cosine
Similarity

$C(\text{문서1}, \text{문서2}) > C(\text{문서1}, \text{문서3})$
(1에 가까울수록 유사)

참조

유튜브 동빈나 : 컴퓨터가 두 데이터(이미지 혹은 자연어)의 유사성을 측정하는 방법: 유클리드 거리, 코사인 유사도

<https://www.youtube.com/watch?v=EGEQutnxjDU>