5장 벡터의 유사도

Vector Similarity

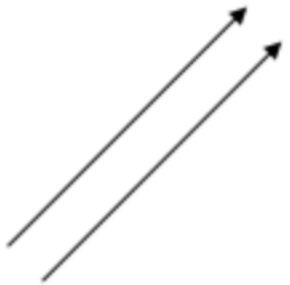
엄호용

목차 Table of Contents

01 코사인 유사도 Cosine Similarity

코사인 유사도 : -1

코사인 유사도 : 0



코사인 유사도 : 1

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

01 코사인 유사도 Cosine Similarity

	바나나	사과	저는	좋아요
문서1	0	1	1	1
문서2	1	0	1	1
문서3	2	0	2	2

- 문서1과 문서2의 cosine 유사도

$$similarity = \frac{0 \times 1 + 1 \times 0 + 1 \times 1 + 1 \times 1}{\sqrt{0^2 + 1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 0^2 + 1^2 + 1^2}} = \frac{2}{\sqrt{9}} \approx 0.67$$

- 문서2과 문서3의 cosine 유사도

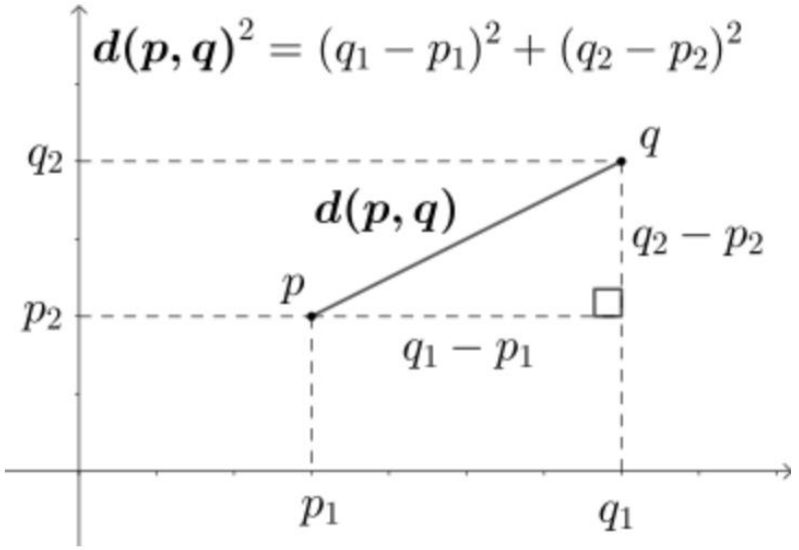
$$similarity = \frac{1 \times 2 + 0 \times 0 + 1 \times 2 + 1 \times 2}{\sqrt{1^2 + 0^2 + 1^2 + 1^2} \times \sqrt{2^2 + 0^2 + 2^2 + 2^2}} = \frac{6}{\sqrt{36}} = 1$$

```
문서 1과 문서2의 유사도 : 0.67
문서 1과 문서3의 유사도 : 0.67
문서 2과 문서3의 유사도 : 1.00
```

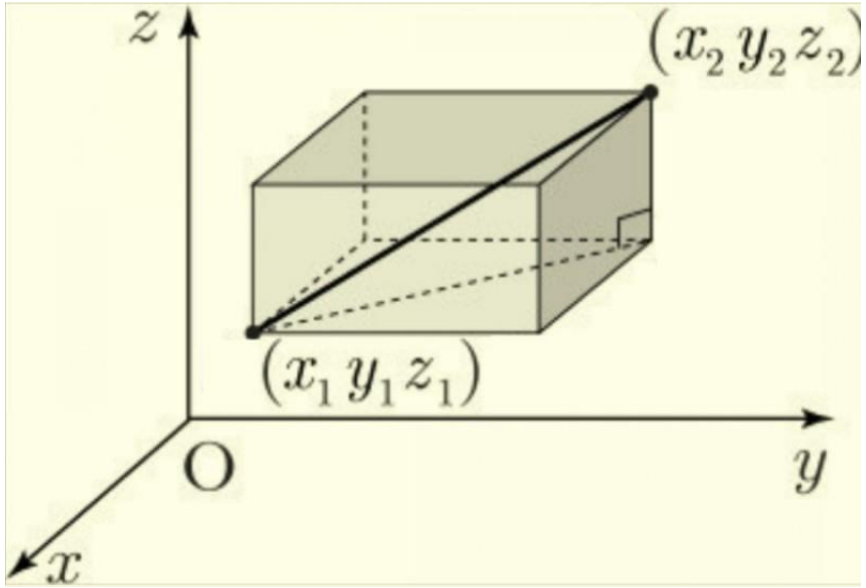
$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

02 유클리드 거리 Euclidean Distance

2차원의 경우



3차원의 경우



$$\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ \ldots \ + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

02 유클리드 거리 Euclidean Distance

	바나나	사과	저는	좋아요
문서1	2	3	0	1
문서2	1	2	3	1
문서3	2	1	2	2

	바나나	사과	저는	좋아요
문서Q	1	1	0	1

```
문서1과 문서Q의 거리 : 2.23606797749979
문서2과 문서Q의 거리 : 3.1622776601683795
문서3과 문서Q의 거리 : 2.449489742783178
```

$$\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ \dots \ + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

- 문서1과 Q의 유클리드 거리

$$\sqrt{(2-1)^2 + (3-1)^2 + (0-0)^2 + (1-1)^2} = \sqrt{5} \approx 2.236$$

- 문서2과 Q의 유클리드 거리

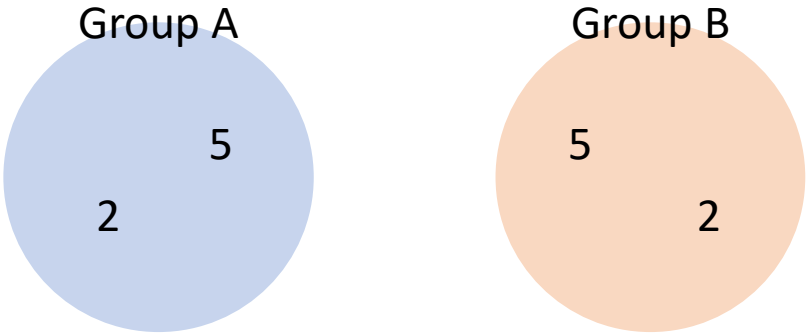
$$\sqrt{(1-1)^2 + (2-1)^2 + (3-0)^2 + (1-1)^2} = \sqrt{10} \approx 3.162$$

03 자카드 유사도 Jaccard Similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

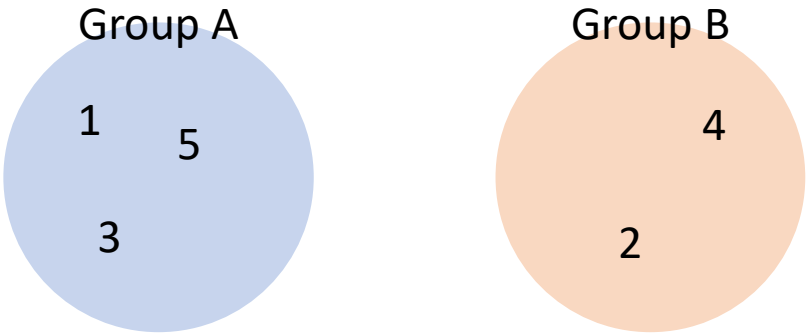
$$J(doc_1, doc_2) = \frac{doc_1 \cap doc_2}{doc_1 \cup doc_2}$$

#Case 1: 두 집합이 일치



$$J(A,B) = \frac{2}{2} = 1.00$$

#Case 2: 두 집합이 상호배타적



$$J(A,B) = \frac{0}{5} = 0.00$$

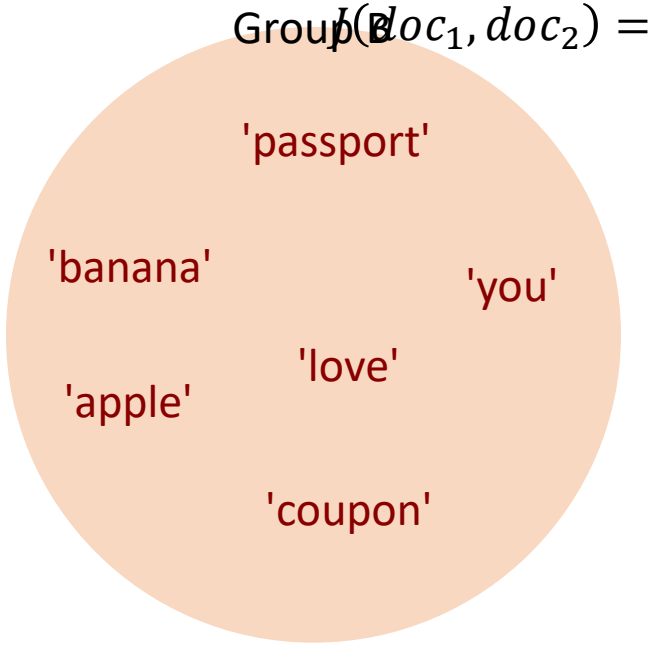
03 자카드 유사도 Jaccard Similarity

교재 예시

```
doc1 = "apple banana everyone like likey watch card holder"
doc2 = "apple banana coupon passport love you"
자카드 유사도 : 0.16666666666666666
```

$$J(doc_1, doc_2) = \frac{doc_1 \cap doc_2}{doc_1 \cup doc_2}$$

$$J(doc_1, doc_2) = \frac{2}{12} \approx 0.167$$



감사합니다!
Q&A