5. 문서 유사도

5장 내용

- **문서 유사도(Document similarity):** 두 문서가 얼마나 유사한지 비교하는 방식
- **Cosine 유사도:** 두 벡터간의 각도를 구하여 유사성을 조사
- **유클리드 거리**: 벡터간의 거리를 측정하여 유사성을 조사
- **Jaccard 유사도:** 공통 단어의 비율을 이용하여 유사성을 조사

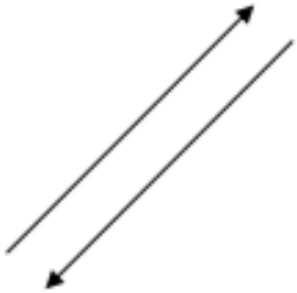
문서 유사도(Document similarity)

- 문서들간에 동일한 단어를 얼마나 포함하는지를 통해 유사성을 검사
- 단어들을 수치화하는 방식(DTM, Word2Vec 등)과 거리를 측정하는 방식에 의해 성능이 결정됨

Cosine 유사도

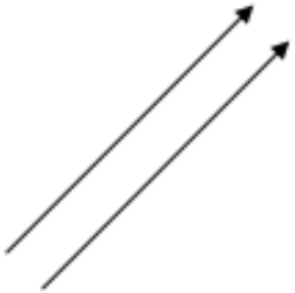
- 문서 단어의 표현 벡터들이 이루는 각도를 통해 문서의 유사성을 판단하는 방식
- 벡터의 크기와 내적을 이용하여 각도를 계산

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



코사인 유사도 : -1

코사인 유사도 : 0



코사인 유사도 : 1

DTM을 이용한 유사도 계산 사례

- 문서 사례

  문서1 : 저는 사과 좋아요

  문서2 : 저는 바나나 좋아요

  문서3 : 저는 바나나 좋아요 저는 바나나 좋아요

- 문서 단어 행렬

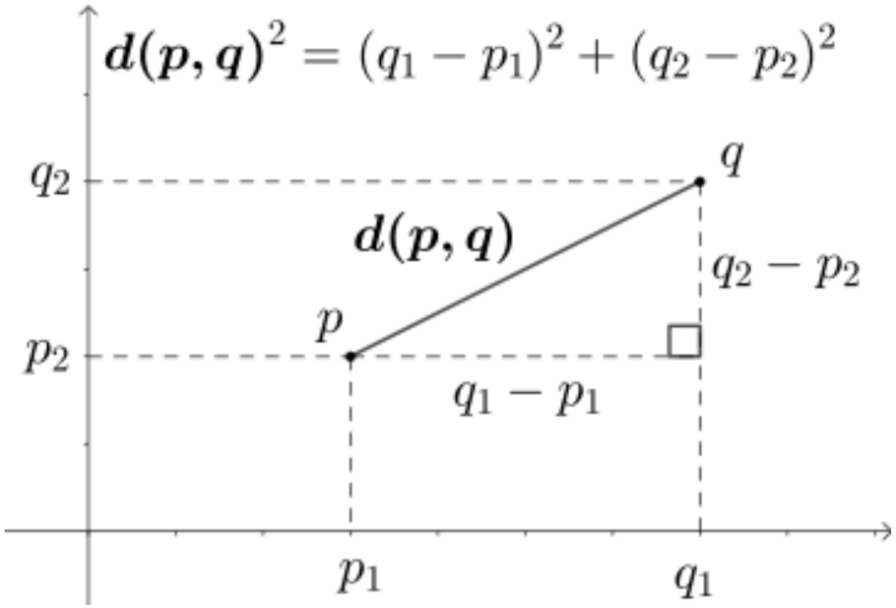
-	바나나	사과	저는	좋아요	문서 단어 벡터
문서1	0	1	1	1	[0,1,1,1]
문서2	1	0	1	1	[1,0,1,1]
문서3	2	0	2	2	[2,0,2,2]

- 코사인 유사도

문서 조합	유사도
1-2	0.67
1-3	0.67
2-3	1.00

유클리드 거리(Euclidean distance)

- 벡터 공간에서 벡터간의 거리에 의해 유사성을 조사
- 2차원 공간에서는 다음과 같이 정의됨



- 코사인 유사도에 비해 성능이 낮음

Jaccard 유사도

- A와 B 두 집합의 유사도를 교집합의 비율에 의해 계산함

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

- 문서의 경우 교집합과 합집합은 두 문서의 단어들을 비교하여 구함

Jaccard 유사도 계산 사례

- 문서

  doc1 = "apple banana everyone like likey watch card holder“

  doc2 = "apple banana coupon passport love you"

- 포함된 단어의 공유 여부에 의해 교집합과 합집합을 구함

  교집합: {‘apple’, ‘banana’}

  합집합: {'card', 'holder', 'passport', 'banana', 'apple', 'love',
  'you', 'likey', 'coupon', 'like', 'watch', 'everyone’}

- **유사도:**

$$J(doc1, doc2) = \frac{|doc1 \cap doc2|}{|doc1 \cup doc2|} = \frac{2}{12} = 0.167$$