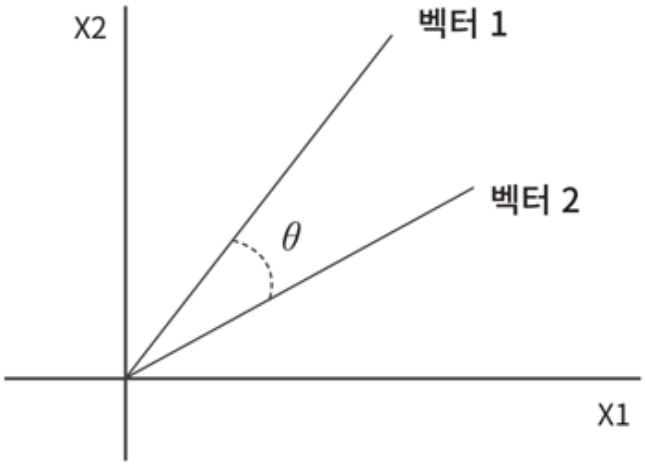
문서 유사도

- 문서와 문서 간의 유사도 비교
- 유사도 측정 방법
  - 코사인 유사도
  - 유클리드 거리
  - 자카드 유사도

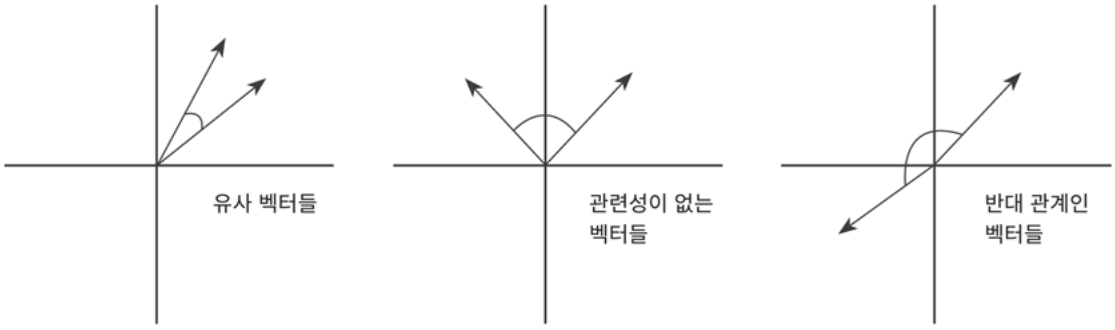
1. 코사인 유사도(Cosine Similarity)

- 문서 사이의 유사도 측정 방법
- 두 벡터 사이의 사잇각을 구해서 얼나마 유사한지 수치로 적용
  - 벡터와 벡터 간의 유사도를 비교할 때 벡터의 크기 보다는 벡터의 상호 방향성이 얼마나 유사한지에 기반



두 벡터 사잇각

- 두 벡터 사잇각에 따라 상호 관계가 유사하거나 관련이 없음을 나타냄



두 벡터의 코사인 유사도

- 벡터 A와 B의 내적 값 : 두 벡터의 크기를 곱한 값의 코사인 각도 값을 곱한 것

$$ A*B=||A||||B||cos θ $$

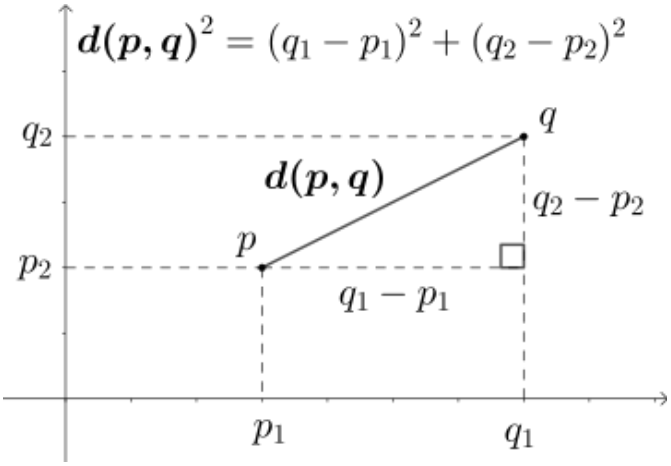
- 유사도 $ cos θ$: 두 벡터의 내적을 총 벡터의 크기의 합으로 나눈 것

$$ similarity = cos θ = { A*B ₩over {||A||||B||} } = { {₩sum_{i=1}^n A_i B_i } ₩over {₩sqrt{₩sum_{i=1}^n A_i^2 }} {₩sqrt{₩sum_{i=1}^n B_i^2 }} }$$

- 문서 유사도 비교를 위해 코사인 유사도를 많이 사용하는 이유
  - 문서를 피처 벡터화하면 차원이 매우 큰 희소행렬이 되므로, 유클리드 거리 기반 지표는 정확도가 떨어짐
  - 문서가 매우 긴 경우 단어의 빈도수도 더 많을 것이기 때문에 이러한 빈도수에만 기반해서 공정한 비교를 할 수 없음

2. 유클리드 거리(Euclidean distance)

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



3. 자카드 유사도(Jaccard Similarity)

- 두 집합이 있는 경우, 두 집합의 합집합에서 교집합의 비율을 계산
- 0과 1사이의 값을 가짐
- 두 집합이 동일하면 자카드 유사도는 1
- 두 집합의 공통원소가 없다면 유사도는 0

$$ J(A, B) = {|A ₩cap B| ₩over |A ₩cup B|} $$

$$ J(doc_1, doc_2) = {|doc_1 ₩cap doc_2| ₩over |doc_1 ₩cup doc_2|} $$
