* **문서 군집화**

**문서 군집화(Document Clustering)** : 비슷한 텍스트 구성의 문서를 군집화(Clustering)하는 것

동일한 군집에 속하는 문서를 같은 카테고리 소속으로 분류할 수 있으므로 텍스트 분류 기반의 문서 분류와 유사함.

**텍스트분류의 문서 분류**는 **사전에 결정 카테고리 값을 가진 학습 데이터 세트**가 필요한 데 반해, **문서 군집화는** 학습 데이터 세트가 필요 없는 **비지도 학습 기반**으로 동작함

**문서 군집화**는 **문서를 피처 벡터화 한 데이터 세트**에 군집화 알고리즘을 적용하여 수행할 수 있으며, 일반적으로 **K-Means군집화가 적용**될 수 있음.

* **문서 유사도 측정 지표**

**Cosine Similarity ( 코사인 유사도 )**

**Jaccard Similarity**

**Manhattan Distance**

**Euclidean Distance**

* **코사인 유사도**

**텍스트, 안테나이(가) 표시된 사진

자동 생성된 설명**

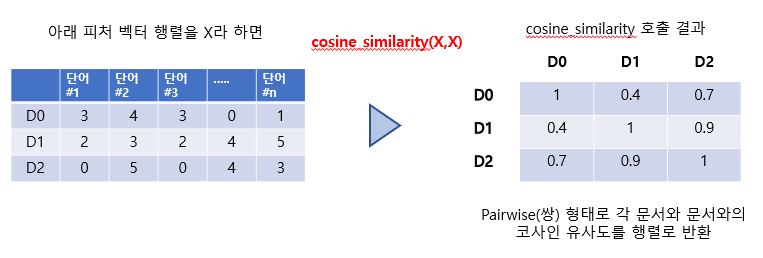
**텍스트이(가) 표시된 사진

자동 생성된 설명**

코사인 유사도는 벡터 크기의 비교가 아닌 **벡터 방향성의 비교에 중점**

* **사이킷런 cosine\_similarity()**

**sklearn.metrics.pairwise.cosine\_similarity( X, Y=None, dense\_output = True )**

****

* **한글 NLP처리의 어려움**

띄어쓰기 / 다양한 조사 / 주어/목적어가 생략되어도 의미전달 가능 / 의성어,의태어,높임말 등

* **한글 형태소 분석**

**형태소 :** 단어로서 의미를 가지는 최소 단위

머릿결 🡪 머리 + 결 , 너를 🡪 너 + 를(을) , 첫째 🡪 첫 + 째

**형태소 분석(Morphological analysis) :** 말뭉치를 이러한 형태소 어근 단위로 쪼개고

각 형태소에 품사 태깅(POS tagging)을 부착하는 작업

* KoNLPy 소개

**KoNLPy :** 기존 C/C++ Java 개발된 한글 형태소 엔진을 파이썬 래퍼(wrapper)로 재작성한 패키지

* **텍스트 분석 Summary**

**텍스트 분석 머신러닝 프로세스 :** 텍스트 전처리 🡪 피처벡터화 🡪 ML 학습/예측 평가

**텍스트 전처리 :** 클린징, 토큰화, 필터링/ 스톱워드 제거, 어근화 (Stemming, Lemmatiation)

**피처 벡터화 :** Bag of Words : 단순 카운트 기반 벡터화 , TF-IDF 벡터화

**텍스트 분석 주요 영역 :**

* 텍스트 분류 : 지도 학습 기반의 텍스트 카테고리 분류
* 감성 분석 : 지도학습 기반, 감성사전 기반
* 텍스트 요약 : 토픽 모델링
* 텍스트 군집화와 유사도 측정
* 한글 NLP
* 비정형 데이터와 정형 데이터 결합을 통한 예측 분석