* **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