BERTopic 개요

1. BERTopic

BERTopic은 BERT embeddings과 클래스 기반(class-based) TF-IDF를 활용하여 주제 설명에서 중요한 단어를 유지하면서도 쉽게 해석할 수 있는 조밀한 클러스터를 만드는 토픽 모델링 기술입니다. BERTopic 알고리즘은 크게 세 가지 과정을 거칩니다.

1) 텍스트 데이터를 SBERT로 임베딩합니다.

SBERT를 사용하여 문서를 임베딩합니다. 이때, BERTopic은 기본적으로 아래의 BERT들을 사용합니다.

- "paraphrase-MiniLM-L6-v2" : 영어 데이터로 학습된 SBERT
- "paraphrase-multilingual-MiniLM-L12-v2" : 50개 이상의 언어로 학습된 다국어 SBERT

2) 문서를 군집화합니다.

UMAP을 사용하여 임베딩의 차원을 줄이고 HDBSCAN 기술을 사용하여 차원 축소된 임베딩을 클러스터링하고 의미적으로 유사한 문서 클러스터를 생성합니다.

3) 토픽 표현을 생성

마지막 단계는 클래스 기반 TF-IDF로 토픽을 추출합니다.