Doc2Vec란 무엇인가?

Doc2Vec은 문서(Document) 단위의 임베딩을 생성하는 대표적인 딥러닝 기반 임베딩 기법입니다.
Word2Vec이 단어를 벡터로 변환하는 것과 달리, Doc2Vec은 **문장, 문단, 전체 문서**와 같은 더 큰 텍스트 단위를 고정 길이의 벡터로 변환합니다. 2014년 Le & Mikolov에 의해 제안되었습니다.

Word2Vec와 Doc2Vec의 차이

구분	Word2Vec	Doc2Vec
임베딩 대상	단어(Word)	문서(문장, 문단, 전체 문서)
입력	단어 시퀀스	문서(문장, 문단 등)
출력	각 단어의 벡터	각 문서의 벡터
활용 예시	유사 단어 찾기, 단어 군집화 등	문서 분류, 유사 문서 검색 등
한계	문맥 정보 제한적	문맥+문서 전체 의미 반영 가능

- **Word2Vec**: 각 단어를 고정된 벡터로 변환. 문장이나 문서의 의미를 파악하려면 여러 단어 벡터의 평균 등 추가적인 처리가 필요.
- **Doc2Vec**: 문서 전체의 의미를 하나의 벡터로 직접 표현. 문서 간 유사도 계산, 분류 등에 바로 활용 가능.

Doc2Vec의 주요 원리

Doc2Vec은 Word2Vec의 구조를 확장하여, **문서 고유의 벡터(문서 태그, document vector)**를 추가로 학습합니다. 대표적으로 두 가지 방식이 있습니다.

1. **Distributed Memory (DM) 모델**
   - 문맥 단어 벡터 + 문서 벡터를 합쳐서 다음 단어를 예측
   - 문서 벡터가 문맥 정보를 기억하는 역할
2. **Distributed Bag of Words (DBOW) 모델**
   - 문서 벡터만으로 문서 내 임의의 단어를 예측
   - Word2Vec의 Skip-gram과 유사

Doc2Vec의 활용 예시

- 문서 분류(뉴스, 이메일, 리뷰 등)
- 유사 문서 검색(질문-답변 매칭, 추천 시스템 등)
- 문서 군집화 및 시각화

요약

- **Word2Vec**: 단어를 벡터로 변환, 문맥 정보는 제한적
- **Doc2Vec**: 문서 전체를 벡터로 변환, 문서 의미를 직접적으로 반영
- Doc2Vec은 Word2Vec의 한계를 극복하여, 문서 단위의 자연어처리 작업에 효과적으로 사용됨