[딥러닝 파이토치 교과서]9장 자연어 전처리

작성일시	@2024년 12월 23일 오후 11:59
주제	

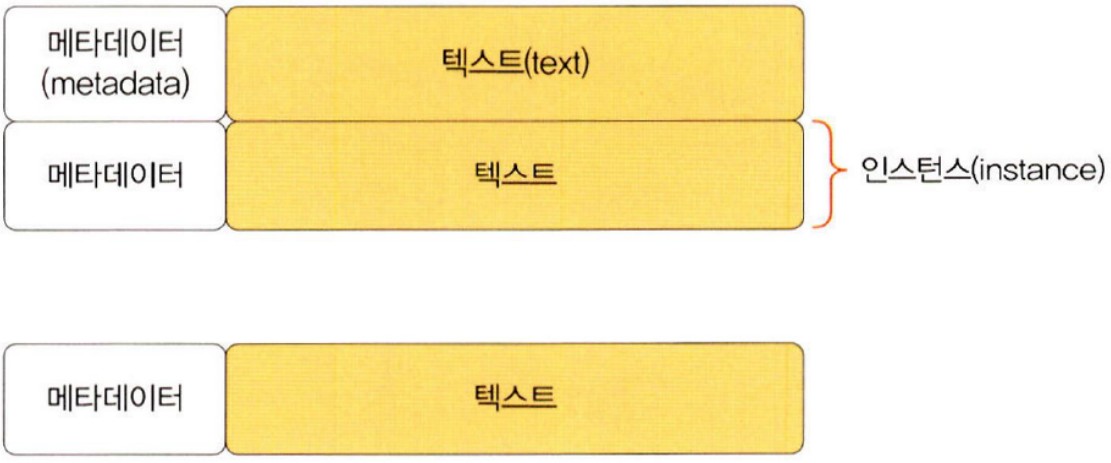
9.1 자연어 처리란

9.1.1 자연어 처리 용어 및 과정

자연어 처리 관련 용어

- 말뭉치(corpus): 자연어 처리에서 모델을 학습시키기 위한 데이터이며, 자연어 연구를 위해 특정한 목적에서 표본을 추출한 집합이다.

그림 9-2 말뭉치(corpus)



- 토큰(token): 문서를 나누는 단위.
  - 토큰 생성(tokenizing): 문자열을 토큰으로 나누는 작업
  - 토큰 생성 함수: 문자열을 토큰으로 분리하는 함수
- 토큰화(tokenization): 텍스트를 문장이나 단어로 분리하는 것. 토큰화 단계를 마치면 텍스트가 단어 단위로 분리된다.

- 불용어(stop words): 문장 내에서 많이 등장하는 단어. 분석과 관계없으며, 자주 등장하는 빈도 때문에 성능에 영향을 미치므로 사전에 제거해 주어야 한다.
- 어간 추출(stemming): 단어를 기본 형태로 만드는 작업.
- 품사 태깅: 주어진 문장에서 품사를 식별하기 위해 붙여 주는 태그(식별 정보)

자연어 처리 과정

그림 9-6 자연어 처리 과정