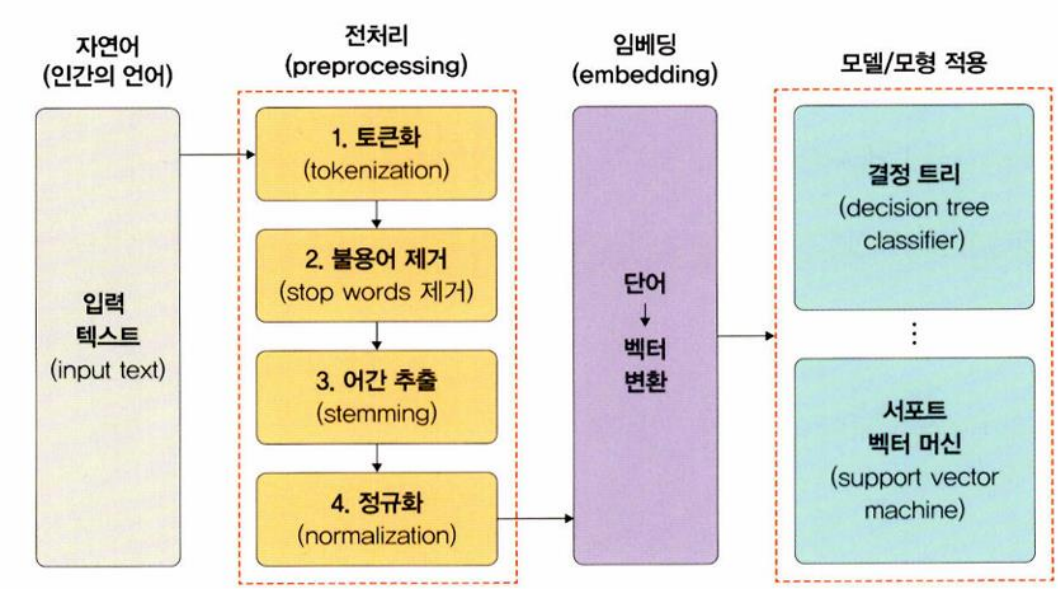
9장. 자연어 전처리

1. 자연어 처리
   1) 관련 용어
      말뭉치(코퍼스): 모델 학습을 위한 데이터
      토큰: 문서를 나누는 단위
      토큰화: 텍스트를 문장이나 단어로 분리하는 것
      불용어: 문장 내에서 자주 등장하는 단어
      어간 추출: 단어를 기본 형태로 만드는 작업
      품사 태깅: 주어진 문장에서 품사를 식별하기 위해 붙여주는 태그
      Det(한정사) Noun(명사) Verb(동사) Prep(전치사)
   2) 자연어 처리 과정



2. 자연어 처리를 위한 라이브러리
   1) NLTK
      교육용으로 개발된 자연어 처리 및 문서 분석용 파이썬 라이브러리
      말뭉치, 토큰 생성, 형태소 분석, 품사 태깅 등의 기능 제공

   2) KoNLPy
      한국어 처리용 라이브러리
   3) Gensim
      워드투벡터 라이브러리
   4) 사이킷런

문서 전처리용 라이브러리

3. 전처리
   1) 결측치 check
      Isnull()매서드 사용
   2) 결측치 처리
      모든 행 NaN이면 삭제

4. 토큰화
   1) 문장 토큰화
   2) 단어 토큰화
5. 불용어 제거
6. 어간 추출
   단어 그 자체만 고려
7. 표제어 추출
   의미도 고려
8. 정규화
   데이터셋이 가진 특성의 모든 데이터가 동일한 정도의 범위 갖도록 하는 것