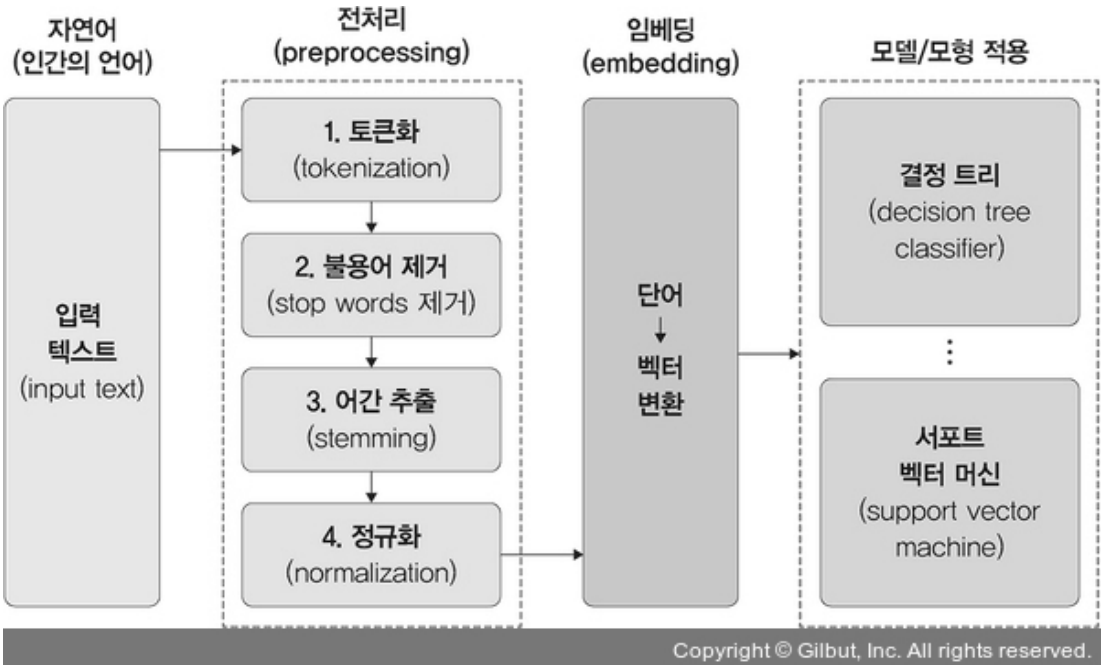
15주차 예습과제

9장 자연어 전처리

▼ 9.1 자연어 처리란

자연어 처리 관련 용어

- 말뭉치(코퍼스) : 자연어 처리에서 모델을 학습시키기 위한 데이터
- 토큰 : 문서를 나누는 단위 (토큰 생성 : 문자열을 토큰으로 나누는 작업, 토큰생성함수 : 문자열을 토큰으로 분리하는 함수)
- 토큰화 : 텍스트를 문장이나 단어로 분리하는 것
- 불용어 : 문장 내에서 많이 등장하는 단어, 분석과 관계 없음. a, the, she, he
- 어간 추출 : 단어를 기본형태로 만드는 작업 ex consign, consigned → consign으로 통일
- 품사 태깅 : 주어진 문장에서 품사를 식별하기 위해 붙여 주는 태그(식별 정보)
    - **Det**: 한정사
    - **Noun**: 명사
    - **Verb**: 동사
    - **Prep**: 전치사



9.1.2 자연어 처리를 위한 라이브러리

NLTK 라이브러리가 제공하는 주요 기능

- 말뭉치
- 토큰 생성
- 형태소 분석
- 품사 태깅

- KoNLPy
  - KoNLPy(코엔엘파이라고 읽음)는 한국어 처리를 위한 파이썬 라이브러리
- Gensim
  - Gensim은 파이썬에서 제공하는 워드투벡터(Word2Vec) 라이브러리
- 사이킷런

  사이킷런(scikit-learn)은 파이썬을 이용하여 문서를 전처리할 수 있는 라이브러리를 제공합니다. 특히 자연어 처리에서 특성 추출 용도로 많이 사용됩니다.

  사이킷런에서 제공하는 주요 기능
  - CountVectorizer: 텍스트에서 단어의 등장 횟수를 기준으로 특성을 추출합니다.
  - Tfidfvectorizer: TF-IDF 값을 사용해서 텍스트에서 특성을 추출합니다.

- HashingVectorizer: CountVectorizer와 방법이 동일하지만 텍스트를 처리할 때 해시 함수를 사용하기 때문에 실행 시간이 감소합니다.

▼ 9.2 전처리