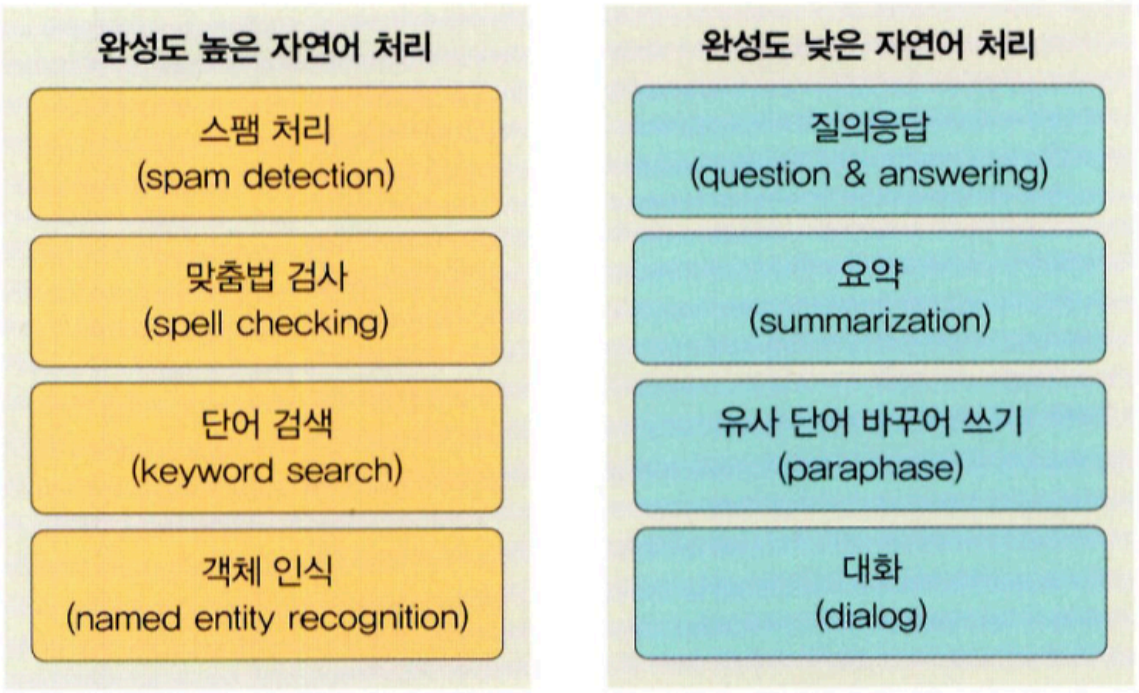
자연어 처리란

- 자연어 처리 : 일상생활에서 사용하는 언어 의미를 분석하여 컴퓨터가 처리할 수 있도록 하는 과정
- 인간 언어에 대한 이해가 필요하기 때문에 접근하기 어려움
- 언어 종류가 다르고 그 형태가 다양하기 때문에 처리가 매우 어려움

  ❯ 그림 9-1 자연어 처리 완성도



자연어 처리 용어 및 과정

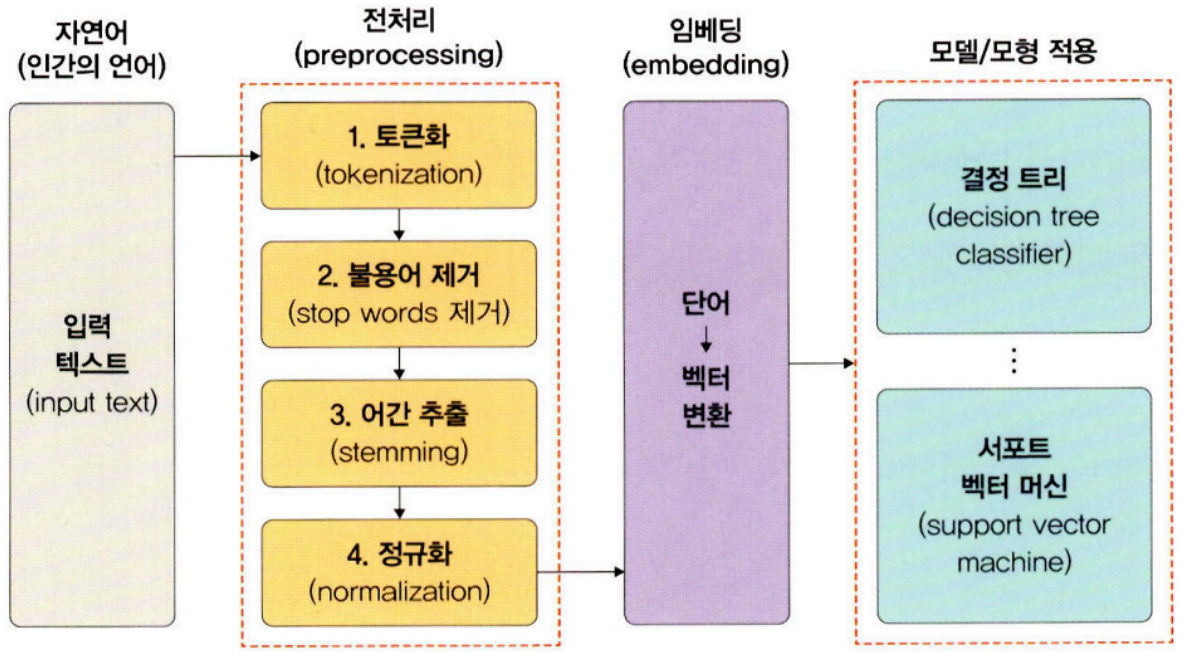
자연어 처리 용어

- 말뭉치 (코퍼스) : 자연어 처리에서 모델을 학습시키기 위한 데이터, 특정한 목적에서 표본을 추출한 집합
- 토큰 : 문서를 나누는 단위
- 토큰화 : 텍스트를 문장이나 단어로 분리하는 것
- 불용어 : 문장 내에서 많이 등장하는 단어, 분석과 관계 없으며, 자주 등장하는 빈도 때문에 성능에 영향을 미치므로 사전에 제거해 주어야 함.
- 어간 추출 : 단어를 기본 형태로 만드는 작업
- 품사 태깅 : 주어진 문장에서 품사를 식별하기 위해 붙여주는 태그

자연어 처리 과정

1. 자연어가 입력 텍스트로 들어옴
2. 입력된 텍스트에 대한 전처리 과정을 거침
3. 전처리가 끝난 단어들을 임베딩
4. 모델/모형을 이용하여 데이터에 대한 분류 및 예측을 수행

❯ 그림 9-6 자연어 처리 과정



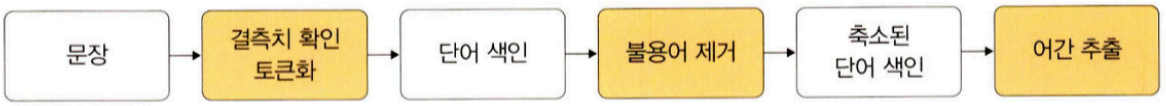
자연어 처리를 위한 라이브러리

- **NLTK**
  - 교육용으로 개발된 자연어 처리 및 문서 분석용 파이썬 라이브러리
  - 말뭉치, 토큰 생성, 형태소 분석, 품사 태깅을 지원
- **KoNLPy**
  - 한국어 처리를 위한 파이썬 라이브러리
- **Gensim**
  - 파이썬에서 제공하는 워드투벡터 라이브러리
  - 워드투 벡터 임베딩, 토픽 모델링, LDA를 지원
- 사이킷런
  - 파이썬을 이용하여 문서를 전처리할 수 있는 라이브러리를 제공
  - 특성 추출 용도로 많이 사용됨

전처리

- 자연어 처리에서 텍스트 자체를 특성으로 사용할 수는 없음.
- 텍스트 데이터에 대한 전처리 작업이 필요.

그림 9-15 전처리 과정



결측치 확인

- 결측치 : 주어진 데이터셋에서 데이터가 없는 것

토큰화

- 토큰화 : 주어진 텍스트를 단어/문자 단위로 자르는 것
  - 단어 토큰화, 문장 토큰화로 구분됨.

불용어 제거

- 불용어 : 문장 내에서 빈번하게 발생하여 의미를 부여하기 어려운 단어들
- 자연어 처리에 있어 효율성을 감소기키고 처리 시간이 길어지는 단점이 있기 때문에 반드시 제거가 필요.

어간 추출

- 단어 원형을 찾아주는 것
- 어간 추출 : 품사가 달라도 사용 가능
- 표제어 추출 : 품사가 같아야 사용 가능

정규화

- 데이터셋이 가진 특성의 모든 데이터가 동일한 정도의 범위를 갖도록 하는 것
- 스케일 차지가 크면 값이 큰 데이터가 더 많은 영향을 미치게 됨 -> but 그렇다고 해서 분석에 더 중요한 요소라고 간주할 수 없음 -> 정규화 필요