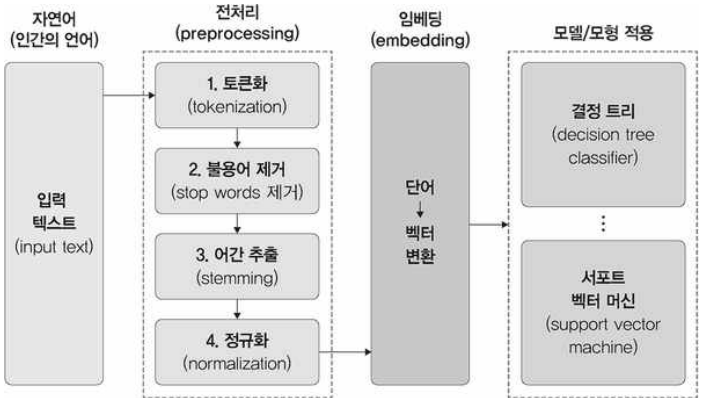
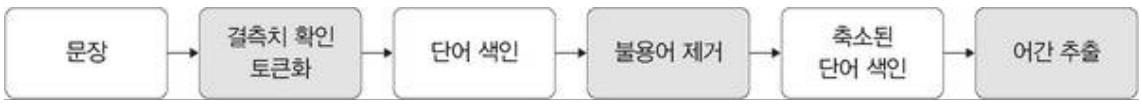
9장 자연어처리(nlp)

용어
말뭉치: 데이터
토큰: 문서를 나누는 단위
토근화: 텍스트를 문장이나 단어로 분리하는 것
불용어: 문장 내에서 많이 등장하는 단어, 성능에 영향을 미치므로 사전 제거해야 함.
어간 추출: 단어를 기본 형태로 만드는 작업
품사 태깅: 품사를 식별하기 위해 붙여주는 태그

자연어처리과정



전처리



토큰화
A, cat, is, on, the, sofa. 단어/문자 단위로 자르는 것.

문장토큰화: 마침표등과 같은 기호에 따라 분리
단어토큰화: 띄어쓰기를 기준으로 구분

불용어제거
from nltk.corpus import stopwords

어간추출
from nltk.stem import PorterStemmer

from nltk.stem import LancasterStemmer

표제어추출
from nltk.stem import WordNetLemmatizer

정규화 scaling