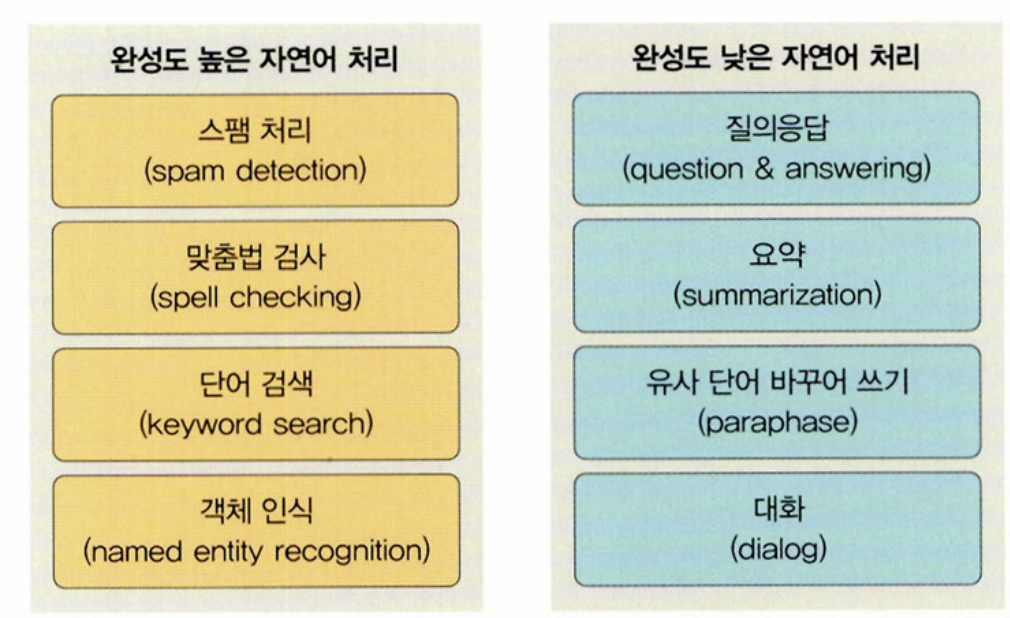
14주차 예습과제

공부할 범위 - 9장 자연어 전처리

9장 자연어 전처리

9.1 자연어 처리란

자연어 처리: 일상에서 사용하는 언어의 의미를 분석해 컴퓨터가 처리할 수 있도록하는 과정. 인간 언어에 대한 이해와 딥러닝에 대한 이해가 모두 필요함.



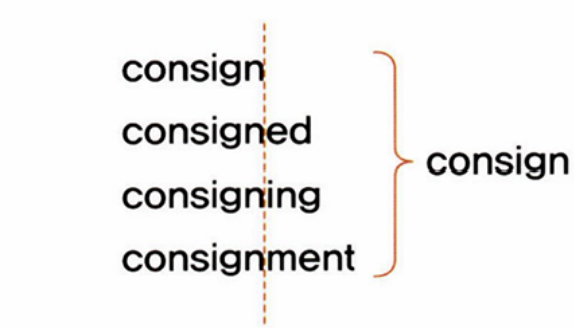
자연어 처리가 가능한 영역과 발전이 필요한 분야

9.1.1 자연어 처리 용어 및 과정

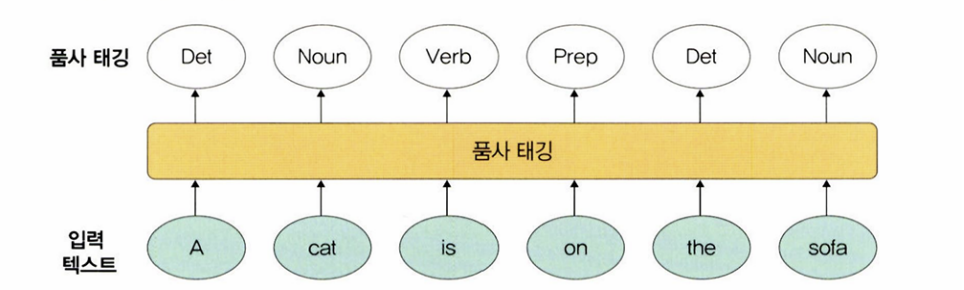
자연어 처리 관련 용어

- 말뭉치(corpus): 자연어 처리에서 모델을 학습시키기 위한 데이터. 자연어 연구를 위해 특정한 목적에서 표본을 추출한 집합.
- 토큰(token): 자연어 처리를 위해 문서를 작은 단위로 나눌 때, 문서를 나누는 단위.
  토큰 생성(tokenizing): 문자열을 토큰으로 나누는 작업.
  토큰 생성 함수: 문자열을 토큰으로 분리하는 함수.
- 토큰화(tokenization): 텍스트를 문장이나 단어로 분리하는 것.

- 불용어(stop words): 문장 내에서 많이 등장하는 단어. 분석과 관계가 없음. 등장 빈도 때문에 성능에 영향을 미침.
  ex) “a”, “the”, “she”, “he”
- 어간 추출(stemming): 단어를 기본 형태로 만드는 작업.
  ex) ‘consign’, ‘consigned’, ‘consignment’ 등이 있을 때 기본 단어인 ‘consign’으로 통일하는 것.



- 품사 태깅(part-of-speech tagging): 주어진 문장에서 품사를 식별하기 위해 붙여 주는 태그.



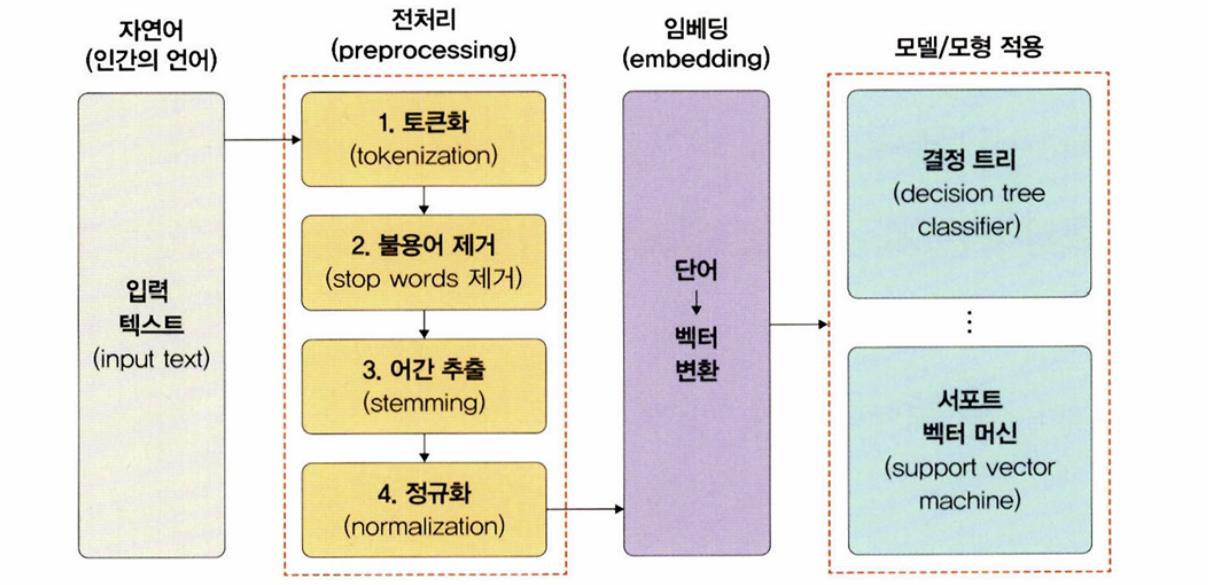
  - Det: 한정사
  - Noun: 명사
  - Verb: 동사
  - Prep: 전치사

자연어 처리 과정

인간 언어인 자연어는 컴퓨터가 이해할 수 없음. 컴퓨터가 이해할 수 있는 언어로 바꾸고 원하는 결과를 얻기까지 크게 네 단계가 필요함.

1. 자연어가 입력 텍스트로 들어옴.
2. 입력된 텍스트에 대해 전처리 과정이 필요함.

3. 전처리가 끝난 단어들을 임베딩(=단어를 벡터로 변환하는 방법).
4. 모델/모형을 이용해 만들어진 데이터에 대한 분류 및 예측을 수행. 데이터 유형에 따라 분류와 예측에 대한 결과가 달라짐.



9.1.2 자연어 처리를 위한 라이브러리

NLTK(Natural Language ToolKit)
: 교육용으로 개발된 자연어 처리 및 문서 분석용 파이썬 라이브러리. 다양한 기능 및 예제를 가지고 있음.

NLTK 라이브러리가 제공하는 주요 기능.

- 말뭉치
- 토큰 생성
- 형태소 분석
- 품사 태깅

Gensim
:
파이썬에서 제공하는 워드투벡터(Word2Vec) 라이브러리. 딥러닝 라이브러리는 아님. 효율적이고 확장 가능.

주요 기능

- 임베딩: 워드투벡터
- 토픽 모델링

- LDA(Latent Dirichlet Allocation)

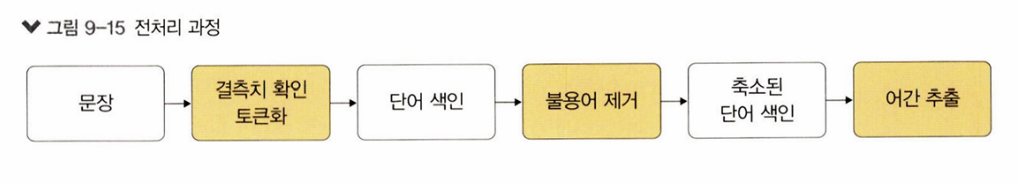
사이킷런(scikit-learn)
: 파이썬을 이용해 문서를 전처리할 수 있는 라이브러리 제공. 특성 추출 용도로 많이 사용됨.

주요 기능

- CountVectorizer: 텍스트에서 단어의 등장 횟수를 기준으로 특성 추출.
- Tfidfvectorizer: TF-IDF 값을 이용해 특성 추출.
- HashingVectorizer: 텍스트 처리 시 해시 함수 사용해 실행 시간 감소.

9.2 전처리

머신 러닝/딥러닝에서 텍스트 자체를 특성으로 사용할 수는 없음. 텍스트 데이터에 대한 전처리 작업이 필요. 이때 전처리를 위해 토큰화, 불용어 제거 및 어간 추출 등이 필요하게 됨.



9.2.1 결측치 확인

결측치: 데이터셋에 데이터가 없는(NaN) 것.

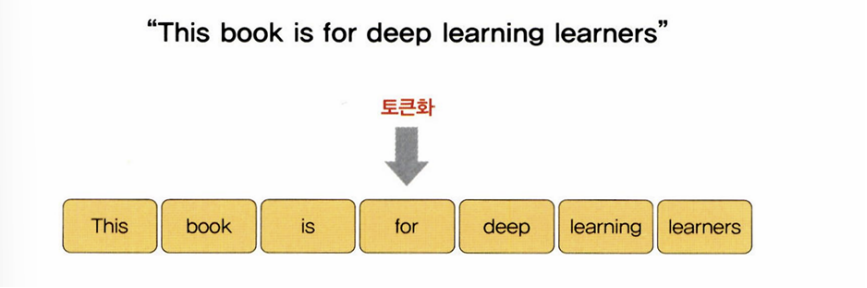
ID	이름	몸무게	키
1	홍길동	76	177
2	성춘향	NaN	155
3	이도령	65	170

9.2.2 토큰화

토큰화: 주어진 텍스트를 단어/문자 단위로 자르는 것. 토큰화는 문장 토큰화와 단어 토큰화로 구분됨.

- 문장 토큰화
  : 마침표(.), 느낌표(!), 물음표(?) 등 문장의 마지막을 뜻하는 기호에 따라 분리.

- 단어 토큰화
  : 띄어쓰기를 기준으로 문장을 구분. 한국어는 띄어쓰기만으로 토큰을 구분하기 어려운 단점이 있음.



9.2.3 불용어 제거

불용어(stop word): 문장 내에서 빈번하게 발생해 의미를 부여하기 어려운 단어들.
'a', 'the' 같은 단어들은 모든 구문에 매우 많이 등장해 의미가 없음. 뿐만아니라 불용어는 자연어 처리의 효율성을 감소시키고 처리 시간이 길어지는 단점이 있음. 따라서 반드시 제거가 필요함.

9.2.4 어간 추출

어간 추출(stemming)과 표제 추출(lemmatization)은 모두 단어의 원형을 찾아주는 것.

- 어간 추출
  : 단어 그 자체만 고려. 품사가 달라도 사용 가능.
  ex) Automates, automatic, automation → automat
- 표제어 추출
  : 단어가 문장 속에서 어떤 품사로 쓰였는지도 고려. 품사가 같아야 사용 가능.
  ex) am, are, is → be / car, cars, car's, cars' → car

둘 다 어근 추출이 목적이지만 어간 추출은 사전에 없는 단어도 추출 가능하고 표제어 추출은 사전에 있는 단어만 추출할 수 있음.

9.2.5 정규화

정규화(normalization)
: 데이터셋이 가진 특성(또는 칼럼)의 모든 데이터가 동일한 정도의 범위(스케일 혹은 중요도)를 갖도록 하는 것.
특성들의 범위가 다르면 더 큰 범위를 가진 특성이 더 많은 영향을 미치게 됨. 그러나 값이 크다고 반드시 더 중요한 요소는 아니기 때문에 이는 정확하지 못한 학습을 초래함. 그러므로 정규화는 중요.