Natural Language Processing with PyTorch
-2장-

정 시 열

목차

말뭉치, 토큰, 타입

모든 NLP 작업은 **말뭉치(Corpus)**라 부르는 텍스트 데이터로 시작한다.

- 말뭉치 = 원시데이터 + 메타데이터

- 원시데이터: 문자 시퀀스 (토큰 단위로 묶었을 때 유용).
  영어에서 토큰은 공백문자나 구두점으로 구분되는 단어와 숫자를 의미.
- 메타데이터: 식별자, 레이블, 타임스탬프 등 텍스트와 관련된 부가 정보

머신러닝 분야에서는 메타데이터가 붙은 텍스트를 **샘플/ 데이터 포인트**라고 부른다. 샘플의 모음인 말뭉치는 **데이터셋**이라고 부른다.

토큰화: 텍스트를 토큰으로 나누는 과정
*토큰화의 기준은 자유롭지만 정확도에 영향을 미칠 수 있다.

타입: 말뭉치에 등장하는 고유한 토큰. 말뭉치에 있는 모든 타입의 집합이 어휘 사전 또는 어휘이다.

n-그램

n-그램: 텍스트에 있는 고정 길이의 연속된 토큰 시퀀스 (n: 시퀀스 길이)

- bigram: 토큰 2개, unigram: 토큰 1개.

An adorable little boy is spreading smiles

unigrams : an, adorable, little, boy, is, spreading, smiles
bigrams : an adorable, adorable little, little boy, boy is, is spreading, spreading smiles
trigrams : an adorable little, adorable little boy, little boy is, boy is spreading, is spreading smiles
4-grams : an adorable little boy, adorable little boy is, little boy is spreading, boy is spreading smiles

표제어와 어간

표제어: 단어의 기본형
ex) flow, flew, flies, … -> fly
- 토큰을 표제어로 바꾸어 벡터 표현의 차원을 줄이는 것을 **표제어 추출**이라고 한다.

어간 추출: 표제어 추출대신 사용하는 축소기법.
- 수동으로 만든 규칙을 사용해 단어의 끝을 잘라 어간 형태로 축소

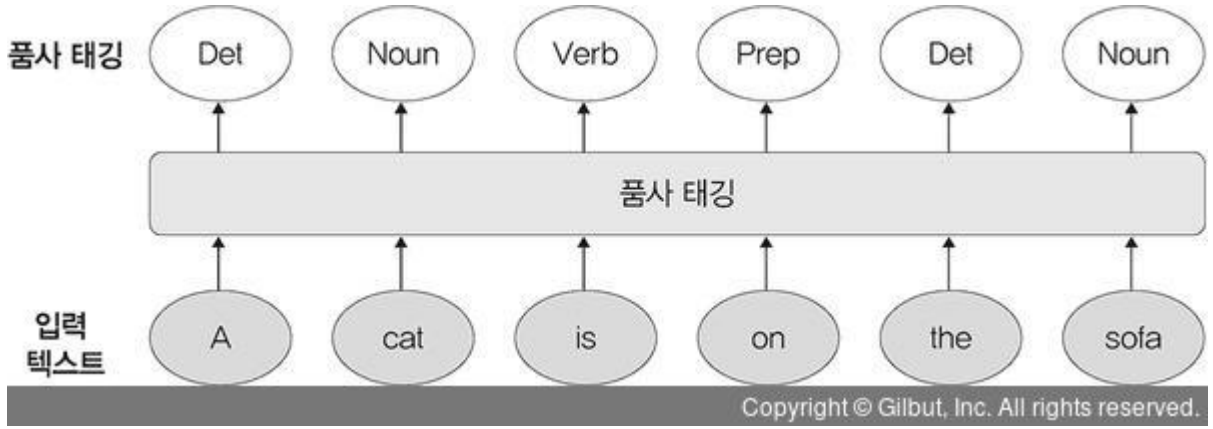
‘geese‘ – 표제어 추출: ‘goose’, 어간 추출: ‘gees’

문장과 문서 분류

TF / TF-IDF표현이 문서나 문장 같은 긴 뭉치를 분류하는데 유용하다.
문서 분류 문제에는 토픽 레이블 할당, 리뷰의 감성 예측, 스팸 이메일 필터링, 언어 식별, 이메일 분류 등이 있다.

단어 분류하기: 품사 태깅

문서에 레이블을 할당하는 개념을 단어나 토큰으로 확장할 수 있다.
단어 분류 작업의 예로는 **품사 태깅**이 있다.



청크 나누기와 개체명 인식

종종 여러 토큰으로 구분되는 **텍스트 구에** 레이블을 할당하는 경우를
청크 나누기 혹은 **부분 구문분석**이라고 한다.
목적은 명사, 동사, 형용사 같은 문법 요소로 구성된 고차원 단위의 유도

또 다른 고차원의 단위는 **개체명**
- 사람, 장소, 회사와 같은 실제 세상의 개념을 의미하는 문자열이다.

```
import spacy
nlp = spacy.load('en')
doc = nlp(u"Mary slapped the green witch.")
for chunk in doc.noun_chunks:
  print('{} - {}'.format(chunk, chunk.label_))
# Mary - NP
# the green witch - NP
```

	개체명 범주	태그	정의
1	PERSON	PER	실존, 가상 등 인물명에 해당 하는 것
2	FIELD	FLD	학문 분야 및 이론, 법칙, 기술 등
3	ARTIFACTS_WORKS	AFW	인공물로 사람에 의해 창조된 대상물
4	ORGANIZATION	ORG	기관 및 단체와 회의/회담을 모두 포함
5	LOCATION	LOC	지역명칭과 행정구역 명칭 등
6	CIVILIZATION	CVL	문명 및 문화에 관련된 용어
7	DATE	DAT	날짜
8	TIME	TIM	시간
9	NUMBER	NUM	숫자
10	EVENT	EVT	특정 사건 및 사고 명칭과 행사 등
11	ANIMAL	ANM	동물
12	PLANT	PLT	식물
13	MATERIAL	MAT	금속, 암석, 화학물질 등
14	TERM	TRM	의학 용어, IT관련 용어 등 일반 용어를 총칭

문장 구조

구 사이의 관계를 파악하는 작업을 **구문 분석**이라고 한다
[그림 2-3] : 문장 안의 문법 요소가 계층적으로 어떻게 이루어지는 보여준다
[그림 2-4]: 각 성분이 서로 의존 관계를 형성하여 하나의 구문을 형성함을 보여준다.

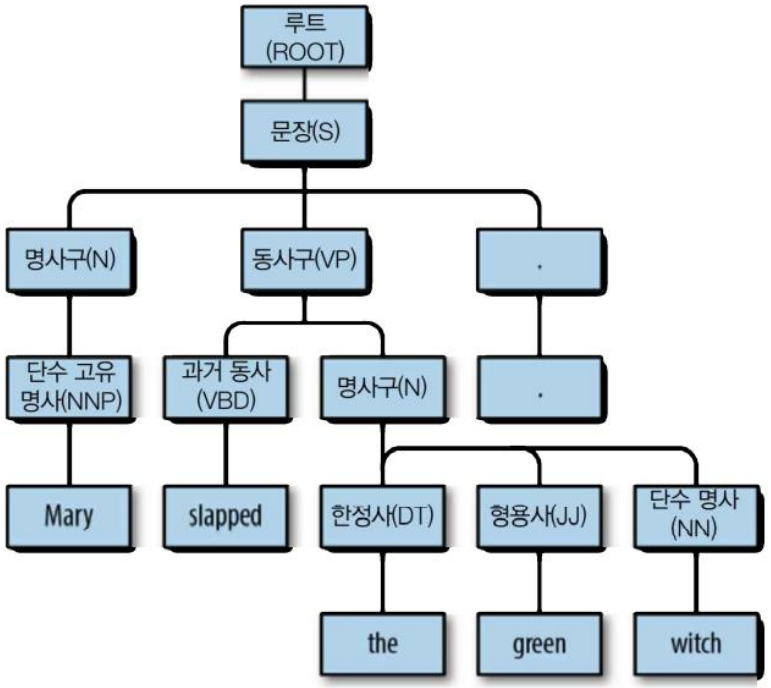


그림 2-3 "Mary slapped the green witch."의 구성 구문 분석

그림 2-4 "Mary slapped the green witch."의 의존 구문 분석

단어 의미와 의미론

단어에는 의미가 하나 이상 존재
- 단어가 나타내는 각각의 뜻을 단어의 의미라고 한다.
- 프린스턴에서 진행하고 있는 WordNet 프로젝트도 모든 영어 단어의 관계와 의미를 수집하는 것이 목표

단어 의미는 문맥으로 결정될 수도 있다.
텍스트 내에서 단어의 의미를 자동으로 찾는 일은 실제로 NLP에 적용된 첫번째 준지도 학습이다.