NLP-Week0

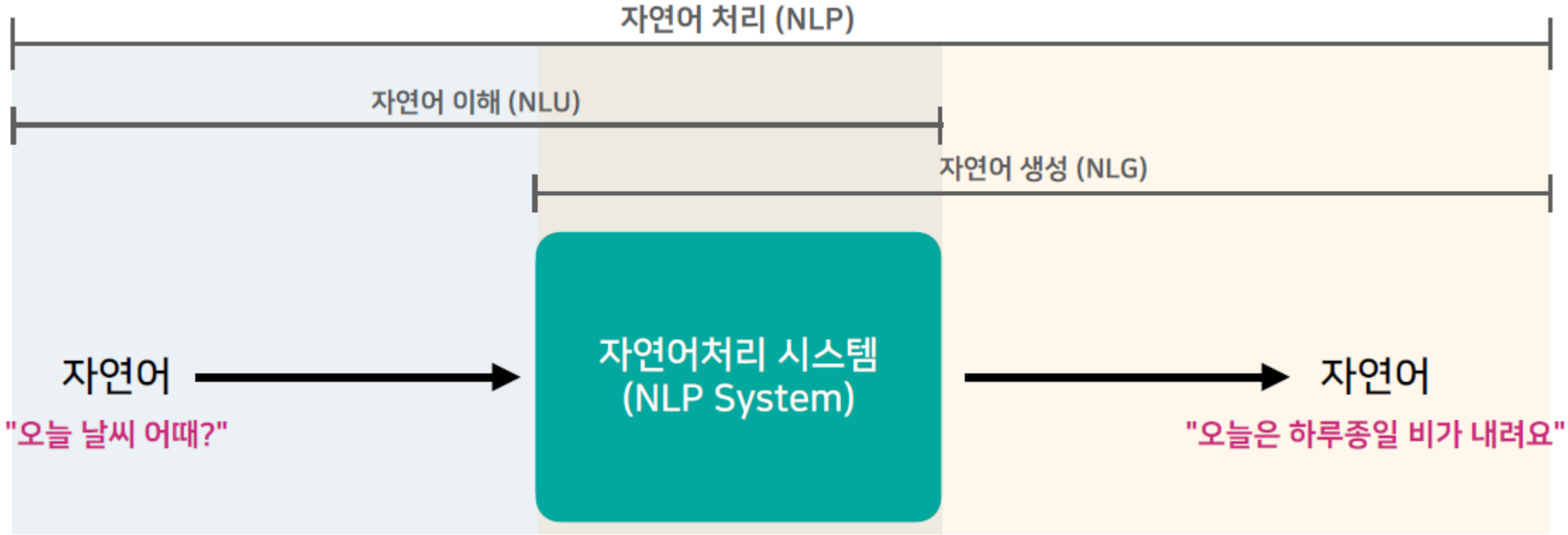
YBIGTA 16기 노시영

Contents

1. 자연어처리

자연어처리란?

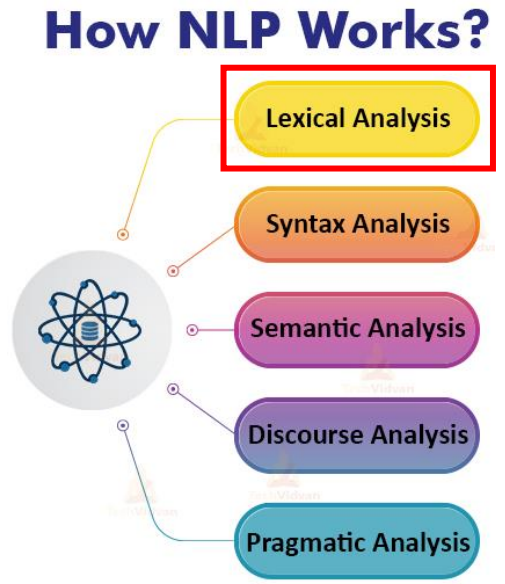
- 자연어: 사람들이 일상적으로 자연스럽게 사용하는 언어 ↔ 인공어(프로그래밍 언어)
- 자연어 처리는 텍스트를 목적에 맞게 내부적으로 처리하는 **자연어 이해**(NLU)와
  주어진 정보를 바탕으로 텍스트를 생성하는 **자연어 생성**(NLG)로 나눌 수 있다.



자연어처리 유형: 단어/형태소 분석

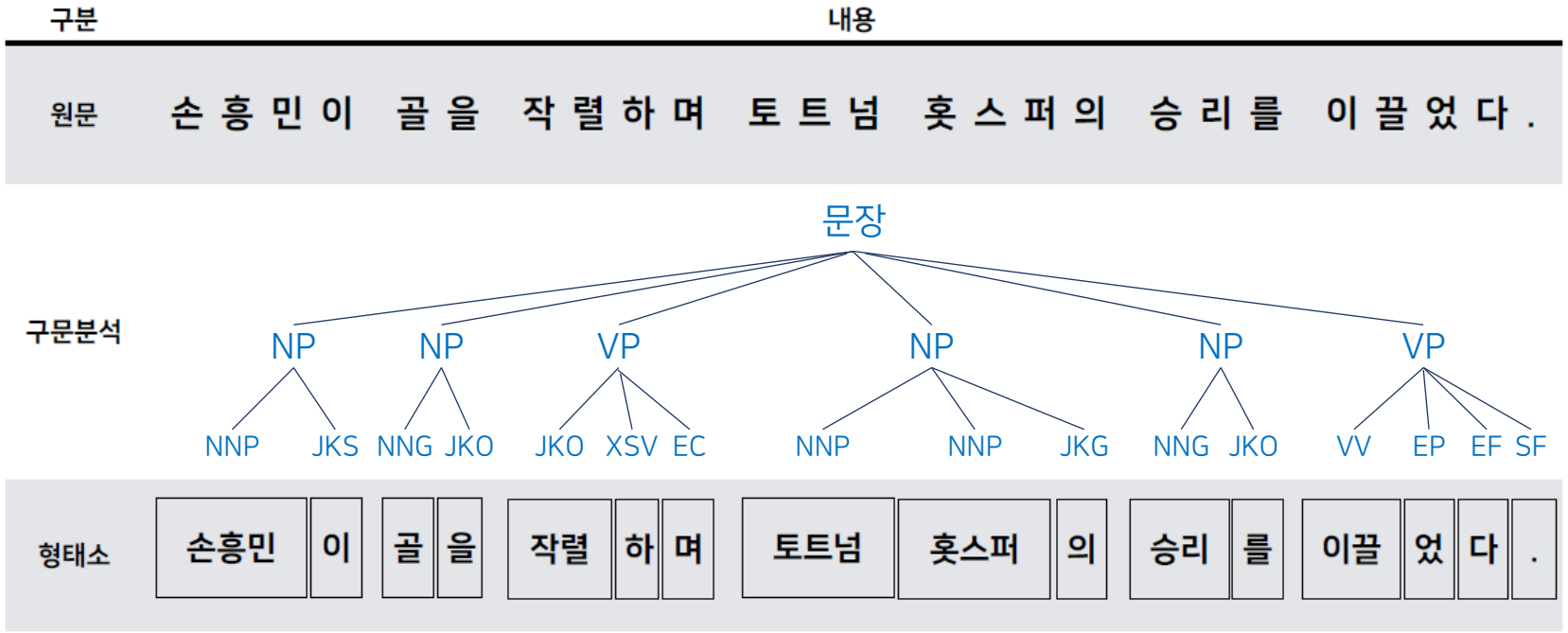
- 문서를 단어 단위로 구분하고 단어의 표면적/내재적 의미를 분석하는 방법
- 형태소(morphome) : 뜻을 가진 가장 작을 말의 단위
- 형태소 분석(morphology analysis) : 문장을 형태소 단위로 분리하고 품사 태깅

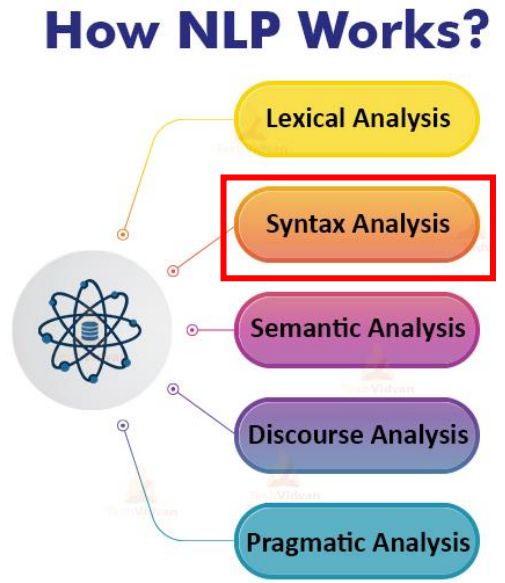
구분	내용																					
원문	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	홋	스	퍼	의	승	리	를	이	끌	었	다	.
음절	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	홋	스	퍼	의	승	리	를	이	끌	었	다	.
형태소	손흥민			이	골	을	작렬		하	며	토트넘			홋스퍼			의	승리		를	이끌		었	다	.
어절	손흥민이				골을		작렬하며				토트넘			홋스퍼의				승리를			이끌었다				.



자연어처리 유형: 구문 분석

- 문장에 출현하는 단어 사이의 구조적 관계를 분석하는 방법
- 한 문장 내 단어 사이의 구조적 관계는 **트리 구조**로 표현할 수 있으며,
  트리 구조로 표현이 어려운 경우에는 문법적으로 맞지 않은 문장이라고 볼 수 있다.

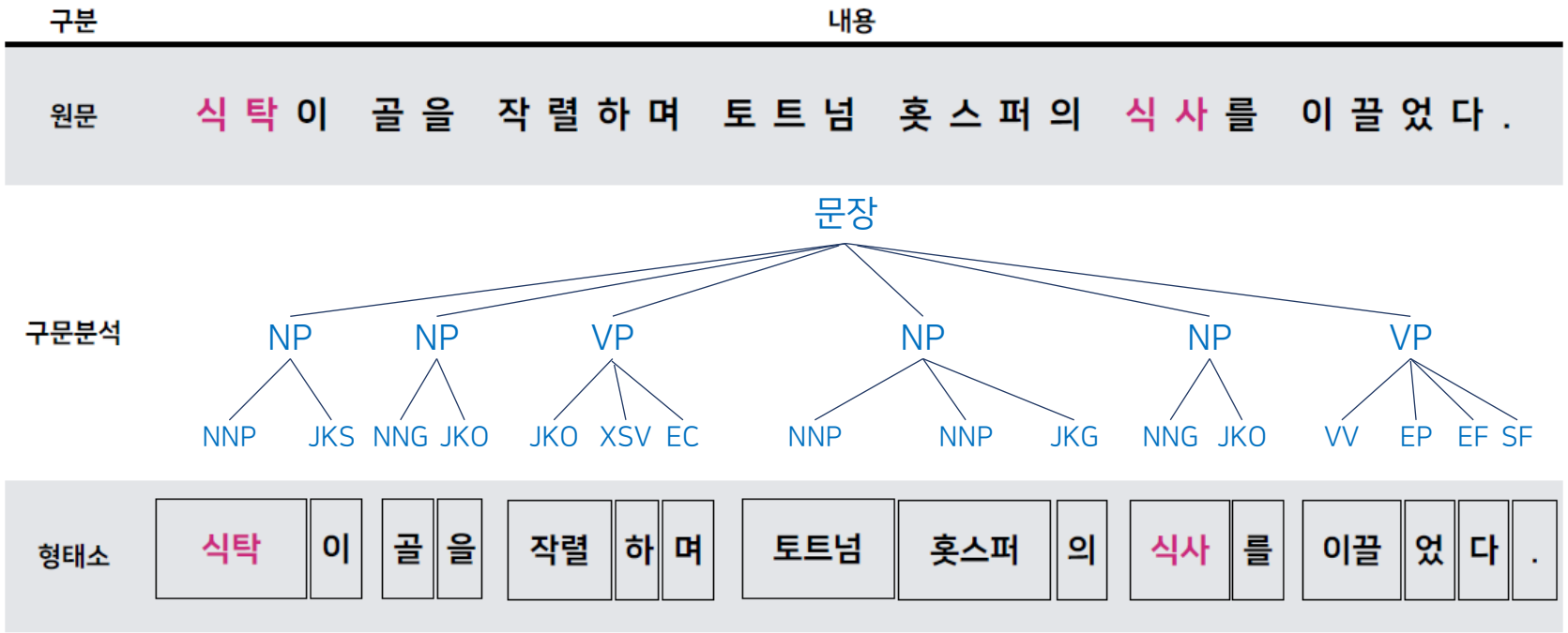


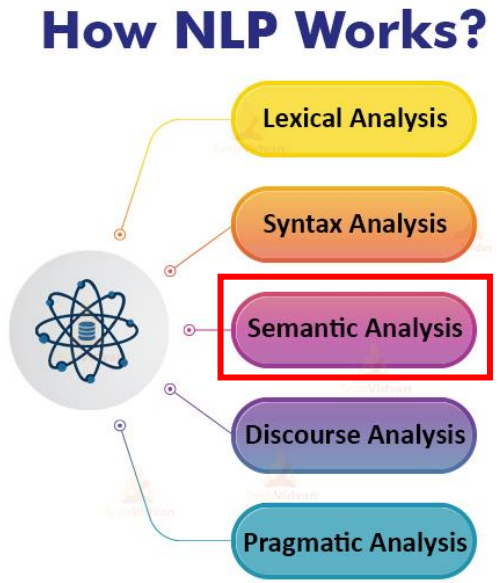


* https://happygrammer.github.io/nlp/postag-set/

자연어처리 유형: 의미 분석

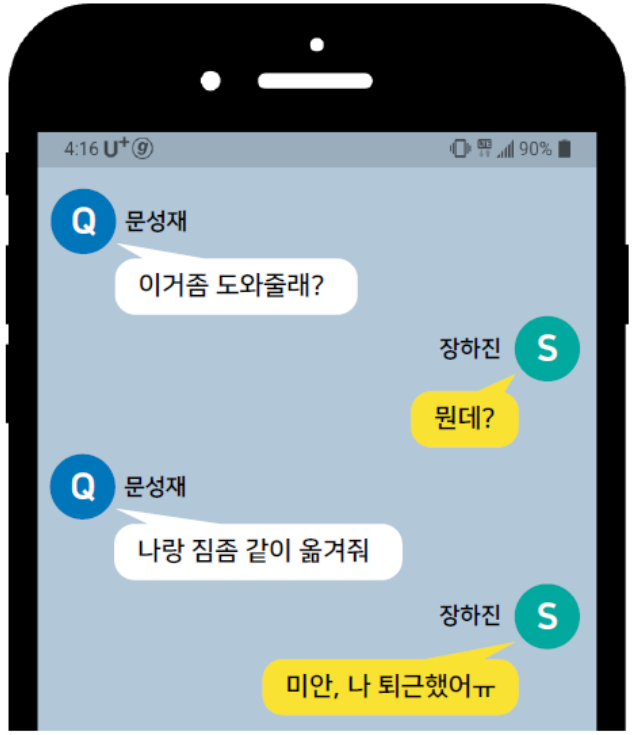
- 문장이 의미적으로 적절한지 여부를 분석하는 방법
- 문법적(Syntax)으로는 옳으나 의미적으로 틀린 경우가 있음
  - 나는 밥을 먹었다. / 나는 자동차를 먹었다.

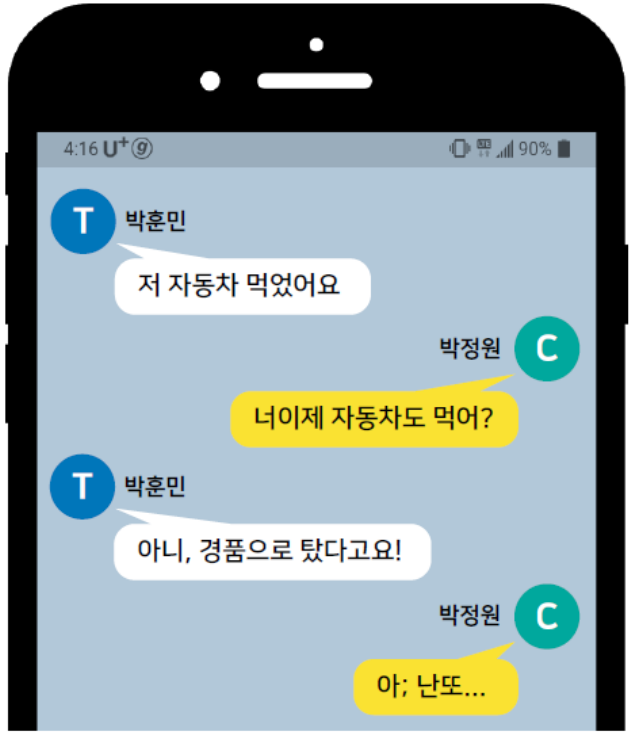


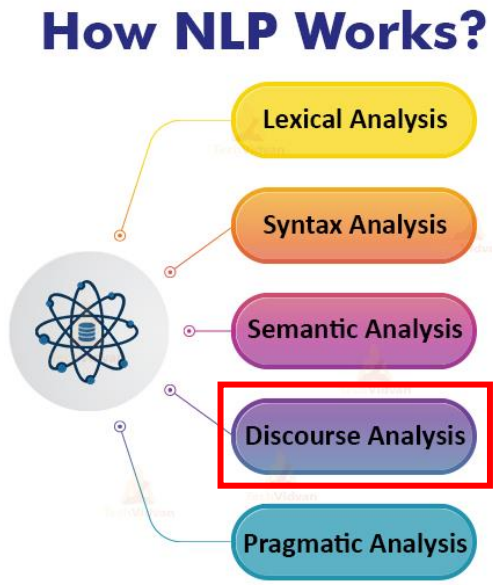


자연어처리 유형: 담화 분석

- 문서 또는 대화 내 여러 개의 문장 속에서 문맥에 따른 문장의 의미를 분석하는 방법
- 문맥에 따라 달라질 수 있는 의미들 중 가장 가능성이 높은 의미를 선정하는 과정

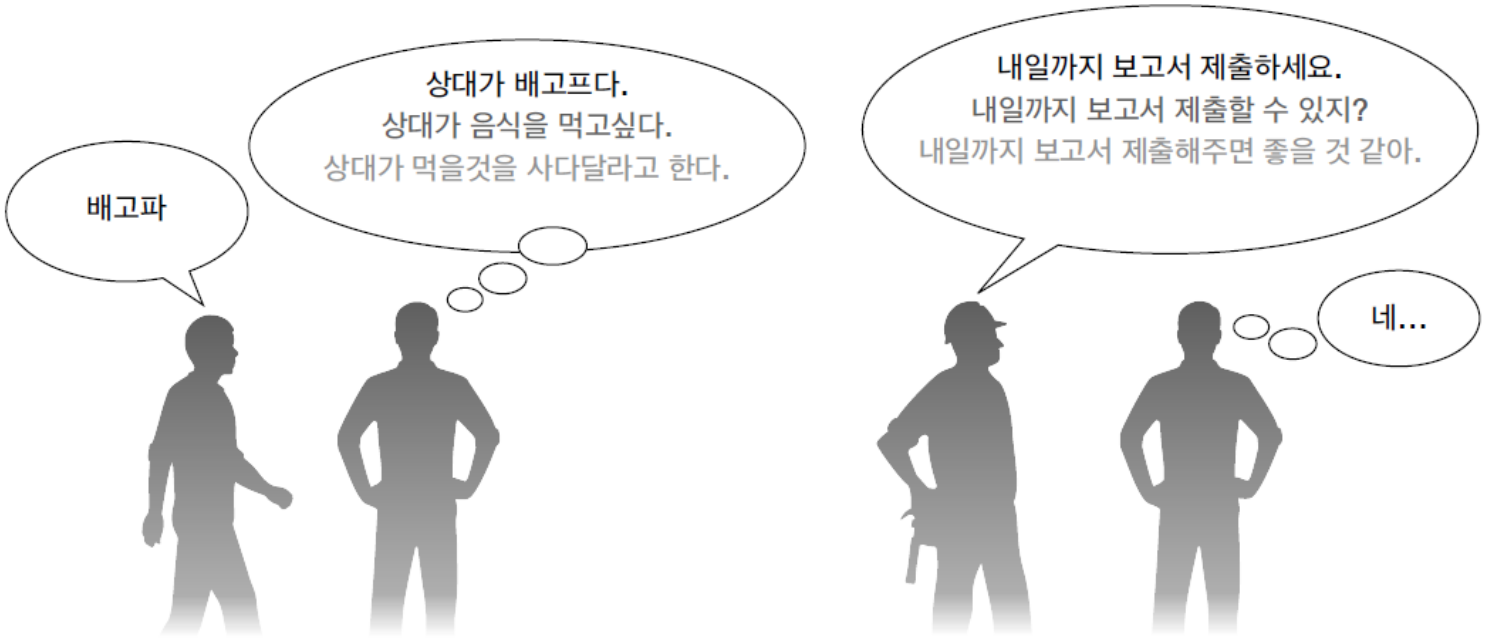


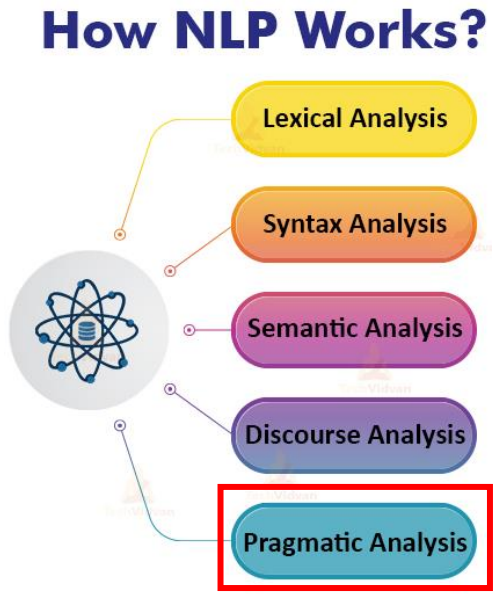




자연어처리 유형: 화용 분석

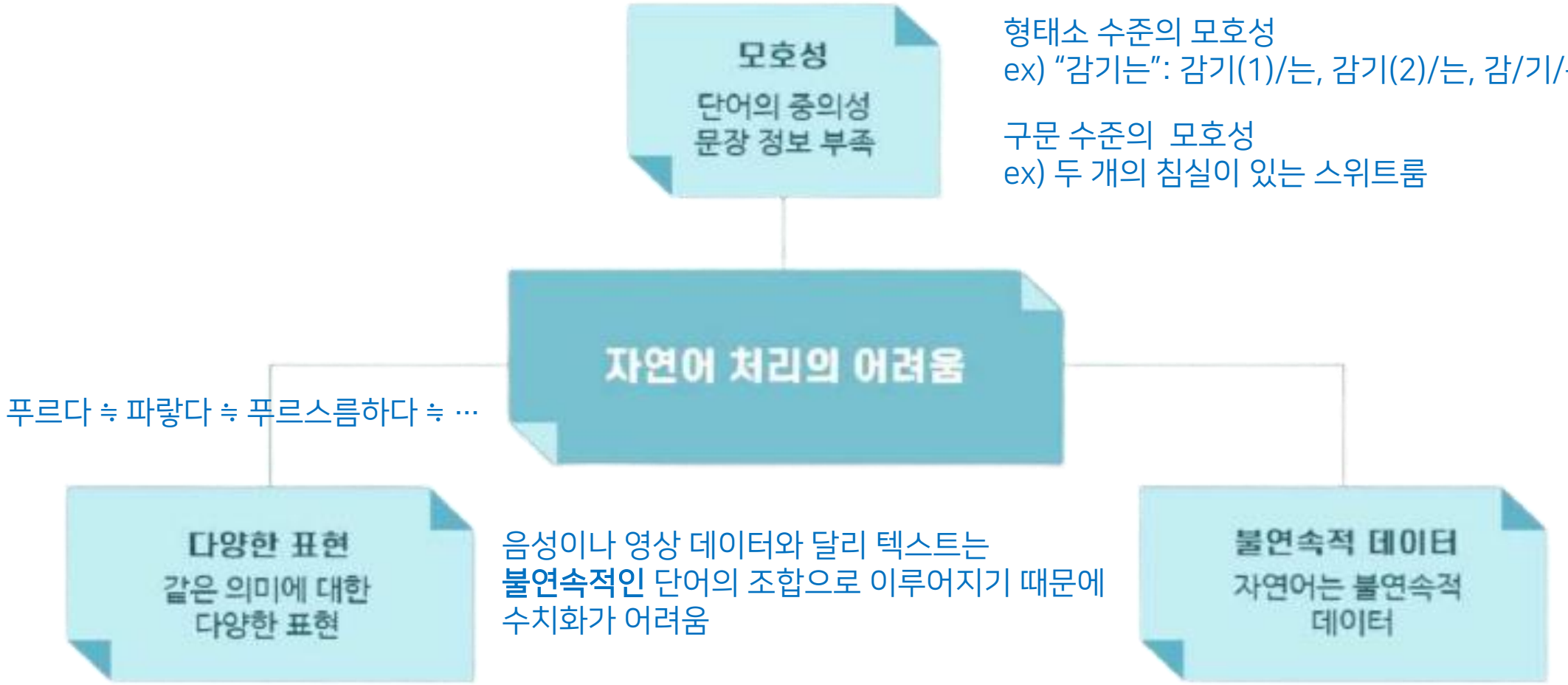
- 언어가 특정 목적을 달성하기 위해 어떻게 사용되는지 분석하는 방법
- 의사소통 과정에서 대화상대나 문맥을 고려하여 어떤 표현을 사용해야 하는지 연구하는 분야
  - "죄송하지만 힘들 것 같습니다." = "도와드릴 수 없다"





자연어처리의 활용 분야

자연어처리의 어려움

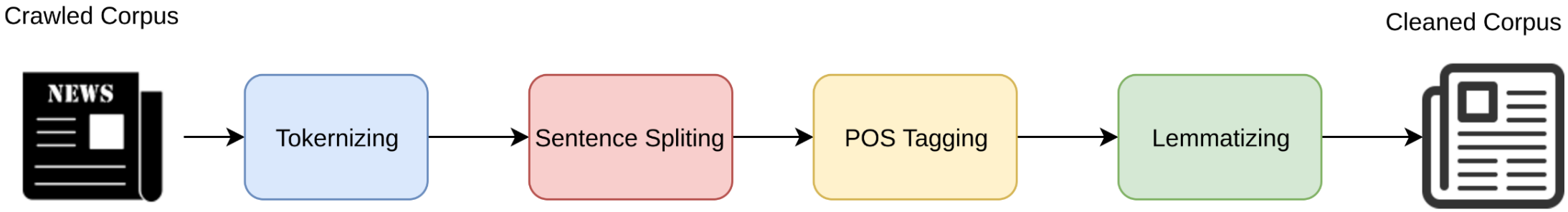


+) 한국어는 교착어로서 어근에 접사가 결합되어 다양한 활용이 가능하므로 타 언어에 비해 텍스트 분석이 어려움

2. 텍스트 전처리

텍스트 전처리

- 데이터 전처리란 어떤 작업을 진행하기 이전에 주어진 데이터를 목적에 따라 변형/가공하는 과정
- 크롤링을 통해 얻은 데이터는 비정형 데이터이므로 데이터 전처리가 필수적이다.



Crawled Corpus

'// flash 오류를 우회하기 위한 함수 추가₩nfunction _flash_removeCallback() {}₩n₩n4일 인천국제공항 2터미널 출국장에 코로나19 음성확인서 무인발급기가 설치돼 있다. /문호남 기자 munonam@[아시아경제 서소정 기자] 4일 국내 코로나19 신규 확진자 수가 424명으로 이틀 연속 400명대를 기록했다. 중앙방역대책본부(방대본)는 이날 0시 기준으로 국내 코로나19 신규 확진자가 424명 늘어 누적 9만1240명이라고 밝혔다. 국내 발생 401명, 해외유입 23명이다.지역별로는 서울 117명, 경기 177명, 인천 18명으로 수도권에서만 312명이 나왔다. 비수도권 지역에서는 대구 19명, 부산 17명, 충북 12명, 경남 9명, 광주·강원·전북 각 6명, 경북 5명, 대전·제주 각 3명, 충남 2명, 전남 1명이다. 울산과 세종에서는 확진자가 발생하지 않았다. 해외유입 확진자는 23명으로 6명은 검역 과정에서, 나머지 17명은 지역 사회에서 확인됐다. 이 가운데 내국인은 9명, 외국인은 14명이다.코로나19 누적 사망자는 전날보다 7명 늘어 1619명이 됐다. 위중증 환자는 140명이다.신규 격리해제자는 462명 증가해 누적 8만2162명이다.이날 0시 기준 6만5446명이 코로나19 백신 추가 접종을 받아 총 15만4421명이 1차 접종을 완료했다. 아스트라제네카 백신은 15만1679명, 화이자 백신은 2742명이 맞았다.서소정 기자 ssj@asiae.co.kr▶ 2021년 신축년(辛丑年) 신년운세와 토정비결은?▶ 발 빠른 최신 뉴스, 네이버 메인에서 바로 보기▶

주요 전처리 과정

토큰화(Tokenization)

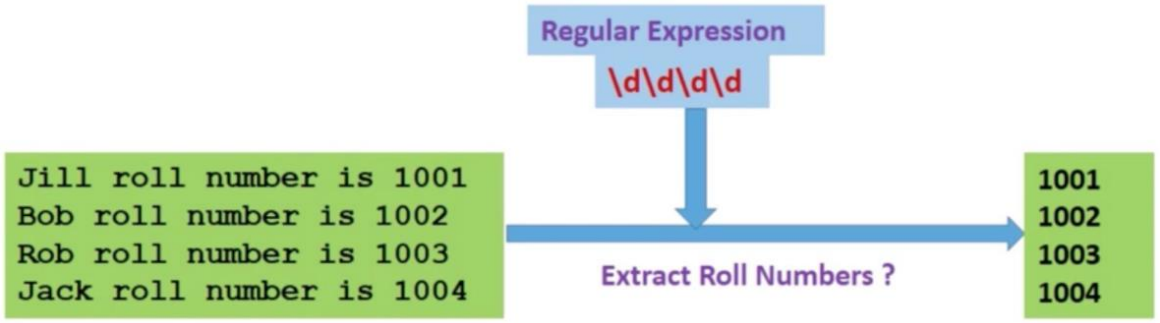
- 주어진 코퍼스(corpus)에서 문장 or 단어 단위로 나누는 작업
  - 입력: Time is an illusion. Lunchtime double so!
  - 출력 : "Time", "is", "an", "illustion", "Lunchtime", "double", "so"
- NLTK(영어), KoNLPy(한글) 등 패키지에서는 다양한 토큰화 모듈을 제공한다.
  - 토큰화 툴에 따라서 기준이 다르기 때문에 상황에 따라 적절한 방법을 사용해야 함
- 한글 토큰화의 어려움
  - 교착어: 접사에 따라 어근이 다양한 형태로 활용됨
  - 띄어쓰기: 영어에 비해 띄어쓰기가 잘 지켜지지 않음

    ex1) 제가이렇게띄어쓰기를전혀하지않고글을썼다고하더라도글을이해할수있습니다.

    ex2) Tobeornottobethatisthequestion

정제 및 정규화

- 정제: 데이터에서 부정확하거나 불필요한 부분을 대체/삭제하는 과정
  - ex) 맞춤법 교정, 불용어 처리
- 정규화: 표현 방법이 다른 단어들을 통합시켜 같은 단어로 만드는 과정
  - 대/소문자 통합: text.lower()
  - 불필요한 단어 제거
    - 등장 빈도가 적은 단어 ex) 5회 이하
    - 길이가 짧은 단어 ex) 1~2글자 단어 제외
  - **정규표현식**:동일한 패턴의 단어들을 일괄적으로 처리



문법 용어

- 어간 추출(Stemming): 동사/형용사에서 변하지 않는 부분을 추출
- 어근(실질형태소) + 접사(형식형태소) => 단어

어간과 어미, 어근과 접사			
동사와 형용사에서 사용하는 개념	어간	용언 활용 시 변하지 않는 부분	'덧붙이다'에서 '덧붙이-'
	어미	용언 활용 시 변하는 부분	'덧붙이다'에서 '-다'
모든 단어에서 쓸 수 있는 개념	어근	단어의 중심 의미를 나타내는 부분	'덧붙이다'에서 '붙-'
	접사	단어에 붙어 그 뜻을 제한하거나 문법적 기능을 하는 형식 형태소	'덧붙이다'에서 '덧-, -이'

어간 및 표제어 추출

- 어간 추출(Stemming): 문장 내 각 단어의 접미사를 삭제하거나 대체하여 어근 형태로 변환하는 과정

```
words=['formalize', 'allowance', 'electricical']
print([s.stem(w) for w in words])
```

```
['formal', 'allow', 'electric']
```

- 표제어 추출(Lemmatizing): 의도된 품사(POS)와 단어의 의미를 식별하는 과정
  - 활용 어미를 없애고 단어를 기본형으로 변환
  - 단어의 형태가 같아도 품사에 따라 의미가 달라질 수 있음

```
print(n.lemmatize('dies', 'v'))
print(n.lemmatize('watched', 'v'))
print(n.lemmatize('watched', 'v'))
print(n.lemmatize('flies', 'n'))
print(n.lemmatize('flies', 'v'))
```

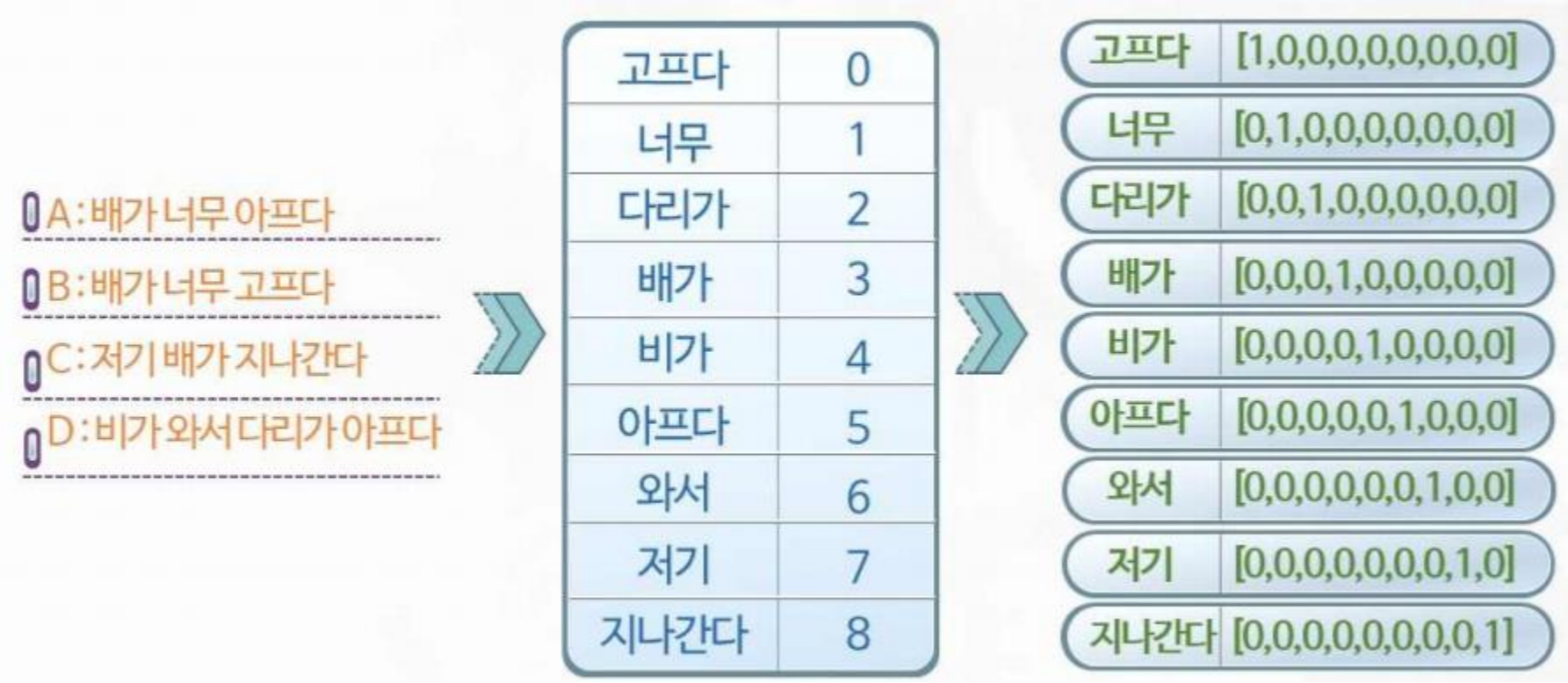
```
die
watch
watch
fly
fly
```

정수 인코딩

- 자연어를 컴퓨터가 이해할 수 있는 숫자 형태로 변환하는 과정
- 정수 인코딩 후에는 문장의 길이를 맞춰주는 패딩(Padding) 작업이 필요함

원-핫 인코딩

- 주어진 단어 집합에서 해당 단어의 인덱스에는 1, 이외에는 0을 부여
- 원-핫 인코딩은 메모리 측면에서 비효율적이며, 단어간 유사도를 반영하지 못한다는 한계가 있음



3. 크롤링 및 실습

크롤링

- 정제되지 않은 웹페이지에서 필요한 데이터를 추출/수집하는 행위
- 크롤링하는 3가지 방법

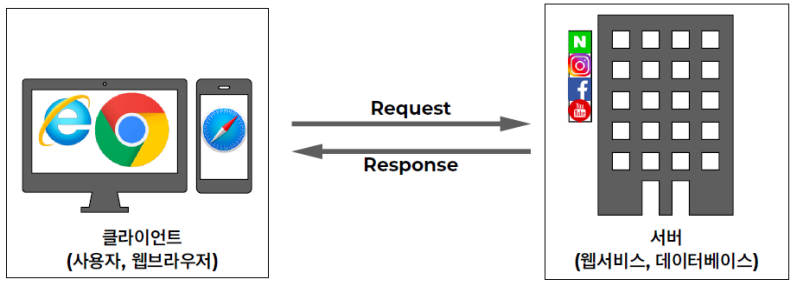
  ① **requests**를 통해 HTML 페이지를 가져와서 **BeautifulSoup**으로 파싱하고 필요한 데이터 추출하기

  ② **Selenium**으로 브라우저에 접속해서 필요한 데이터 추출하기

  ③ **Open API**(Rest API)를 호출하여 필요한 데이터 추출하기

  * 일반적으로 API를 통한 수집은 제약이 많아서 requests나 selenium을 통한 크롤링이 요구됨

- 실습 진행: petition_crawling.ipynb
  - 국민청원 데이터를 requests 방식으로 크롤링하는 코드입니다.
  - 드라이브에 사본으로 저장 후 실행해주세요!



과제

- 과제 코드: week0_assignment.ipynb
  - 드라이브에 사본으로 저장 후 실행해주세요!
- Task List
  - 텍스트 정제
  - 토크나이징
  - 워드 클라우드 생성
- 한글/영어 워드 클라우드(이미지) 제출
  - 파일명: WC_이름_검색어1.jpg, WC_이름_검색어2.jpg
  - 제출기한: 12일 23:59분까지

예시

References

- TEANAPS, Text Mining for Practice
- 위키독스, 딥러닝을 이용한 자연어 처리 입문
- KOCW, AI 서비스 개발을 위한 딥러닝 자연어처리
- NLTK 3.5 documentation
- lovit/soynlp
- 코딩도장, 정규표현식 사용하기