

# AI 개발 실무

## 10. 한국어 분석

김 윤 기    교수



10<sub>week</sub>

A I 개 발 실 무 | 김 윤 기

# 한국어 분석



- » 한국어 자연어 처리 과정을 이해할 수 있다.
- » 한국어의 형태소 분석을 수행할 수 있다.
- » 한국어에 Word2Vec을 적용할 수 있다.

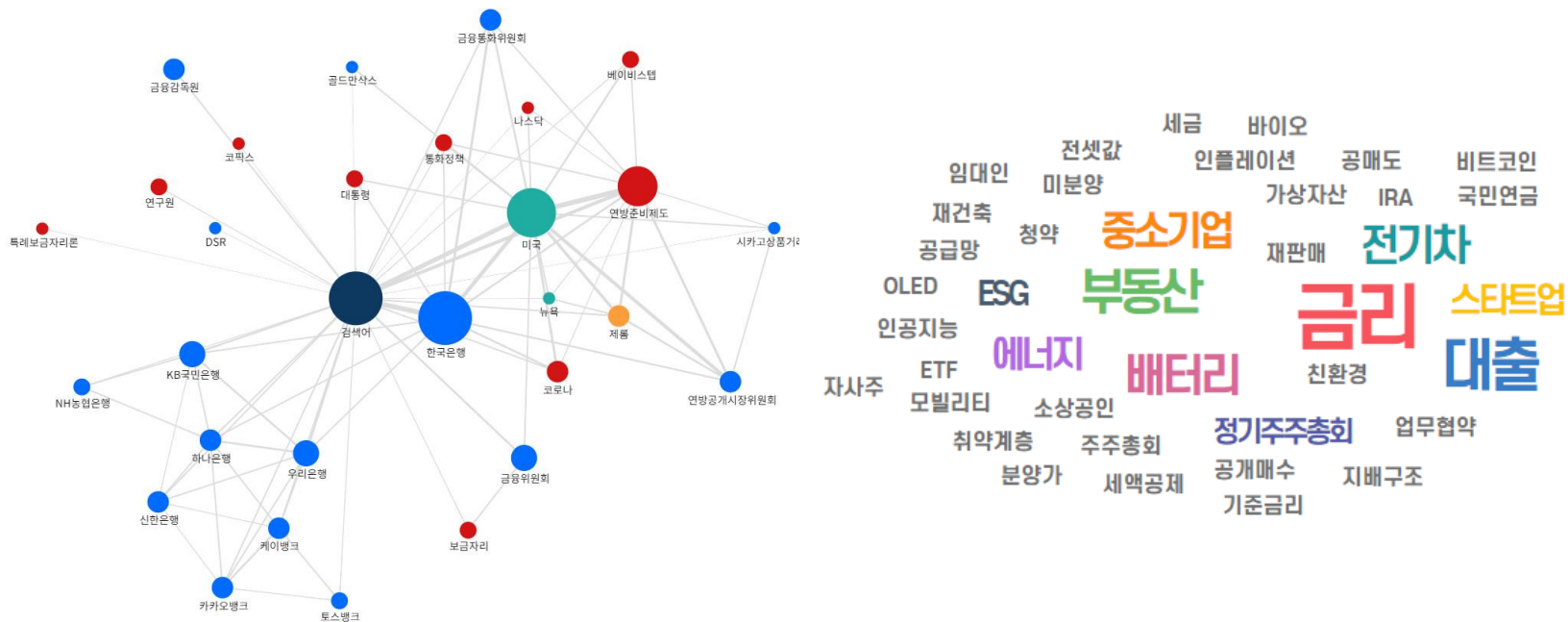
## ① 한국어 분석

---

## ② 형태소 분석 및 Word2Vec 실습

---

수집된 한국어 기사를 분석하는 방법은?



뉴스 기사의 단어 수를 기반으로 접사 등을 제거하여  
키워드를 추출하고, 연관어를 분석함

---

CHAPTER

01

# 한국어 분석

# 1. 한국어 분석이란?

우리가 일상적으로 사용하는 한국어를  
컴퓨터가 이해하는 방식을 사용하여 분석하는  
것

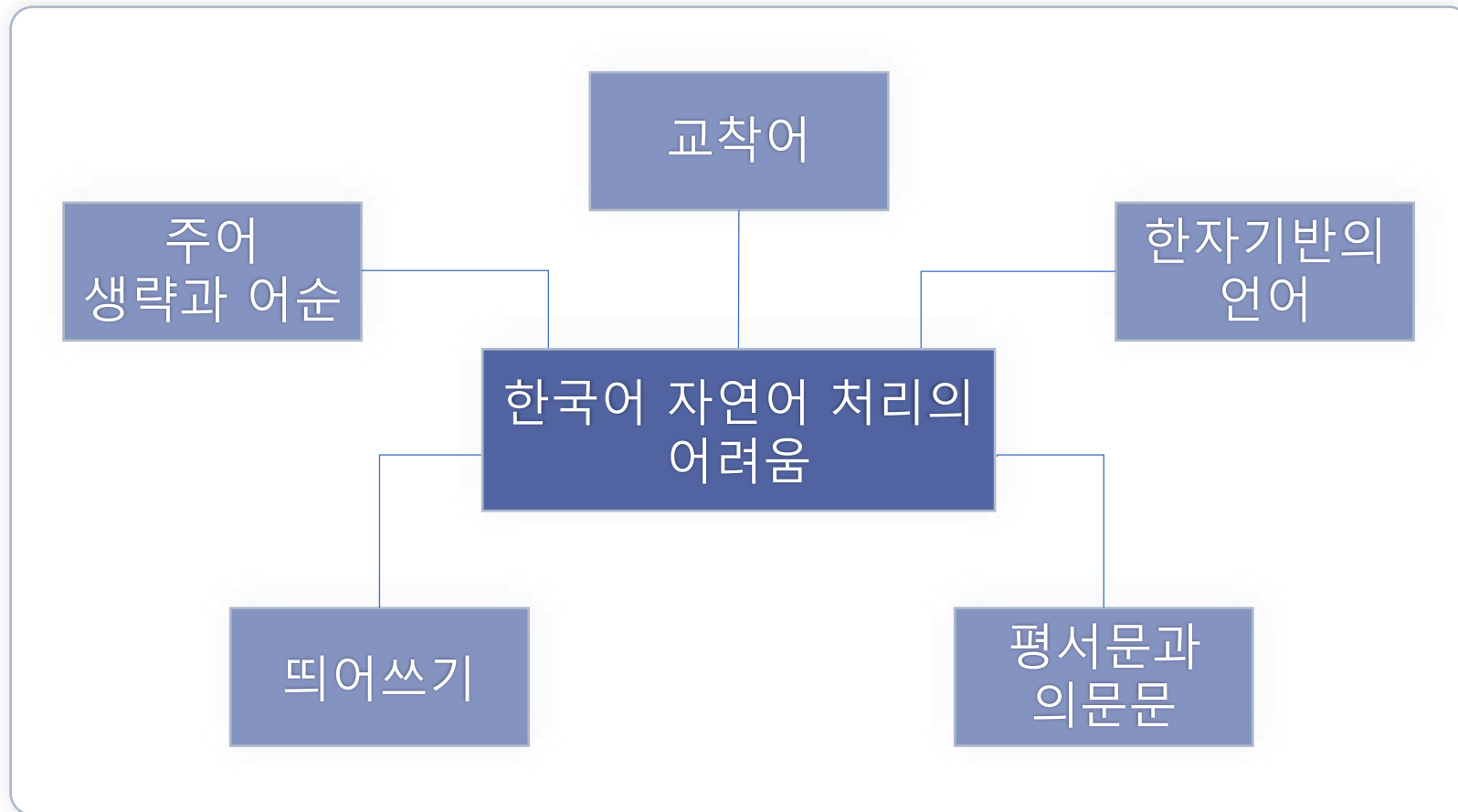
자연어의 의미를 분석하여 많은 양의 데이터를  
사람의 이해하는 수준과 가깝게 처리하는 것을 목적으로  
함

내용 요약, 번역, 사용자의 감성 분석, 텍스트 분류 작업  
(스팸 메일 분류, 뉴스 기사 카테고리 분류), 질의 응답 시  
스템, 챗봇과 같은 곳에서 사용됨

웹에서 수집한 데이터가 텍스트 데이터일 때,  
자연어 처리를 수행하여 분석 가능

## 2. 한국어 자연어 처리의 특징

- » 다른 언어에 비해 비교적 불규칙하며,  
합성어 및 중의적 단어가 많아 컴퓨터로 분석하기 쉽지 않음





## 2. 한국어 자연어 처리의 특징

### 1 교착어

» 언어의 유형론적 분류 중 하나로, 단어의 중심이 되는 형태소(어근)에 접사를 비롯한 다른 형태소들이 덧붙여 단어가 구성되는 것이 특징

» 접사에 의해 다양한 의미가 출현

ex

'정말 이렇게 좋은 옷을 버리시렵니까?'  
-시렵니까 (높임: -시-, 의도: -려-, 의문: -ㅂ니까)

» 접사에 따라 단어의 역할이 정의됨

ex

'다리가 아파서 다리를 주물렀더니 다리의 통증이 괜찮아졌다'  
라는 문장에서 컴퓨터는 <'다리가', '다리를', '다리의'> 세 단어를 모두 다른 단어로 간주할 가능성이 있음

## 2. 한국어 자연어 처리의 특징

### ② 평서문과 의문문의 불명확한 구분

- » 평서문(화자가 문장의 내용을 객관적으로 진술하는 문장)과 의문문(화자가 청자에게 질문을 하여 답을 요구하는 문장)의 구분이 명확하지 않음

어디 갔다 왔어? Vs 어디 갔다 왔어.

## 2. 한국어 자연어 처리의 특징

### 3 띄어쓰기

- » 띄어쓰기에 대한 표준은 계속 바뀌어 왔고, 띄어쓰기 적용 방식은 매우 까다로운 편임
- » 띄어쓰기를 하지 않아도 의미가 전달되는 경우가 많고, 실제로 잘 사용하지 않는 경우도 있음

ex

이번휴가 때 어디로 놀러가?

### 4 주어 생략

- » 한국어에서는 주어를 생략하는 경우가 많아, 컴퓨터가 생략된 정보를 메꿀 수 없어 문장의 정확한 의미 파악이 힘들

ex

(너) 밥 먹어

## 2. 한국어 자연어 처리의 특징

### ⑤ 어순의 자유로움

» 접사에 따라 단어의 역할이 정의 되어 어순이 중요하지 않음

나는 기쁘게 결승선을 통과했다.

나는 결승선을 기쁘게 통과했다.

통과했다 결승선을 나는 기쁘게.

기쁘게 결승선을 통과했다 나는.

결승선을 기쁘게 통과했다 나는.

## 2. 한국어 자연어 처리의 특징

### ⑥ 한자 기반의 언어

한자

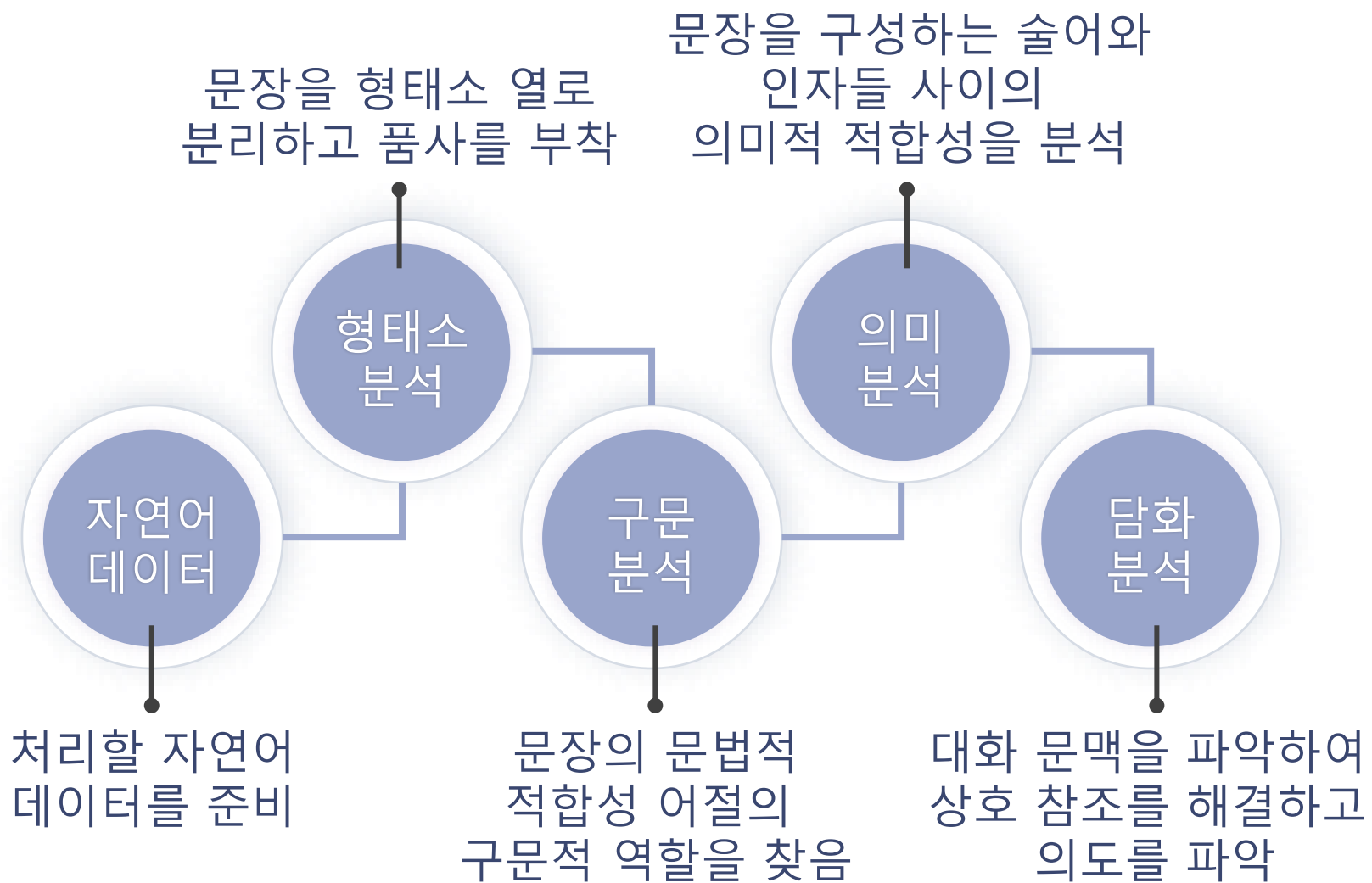
&

한글

- » 하나의 문자로 단어 또는 형태소를 나타내는 표어 문자
- » 읽는 방법은 같지만 문자의 모양이나 의미는 다른 경우가 있음
- » 사람의 말소리를 기호로 나타낸 표음 문자

- » 표어 문자인 한자를 표음 문자인 한글로 표현하는 과정에서 정보의 손실이 존재함
- » 같은 글자처럼 보이지만 하나의 음절이 다른 의미를 지니는 경우가 존재함

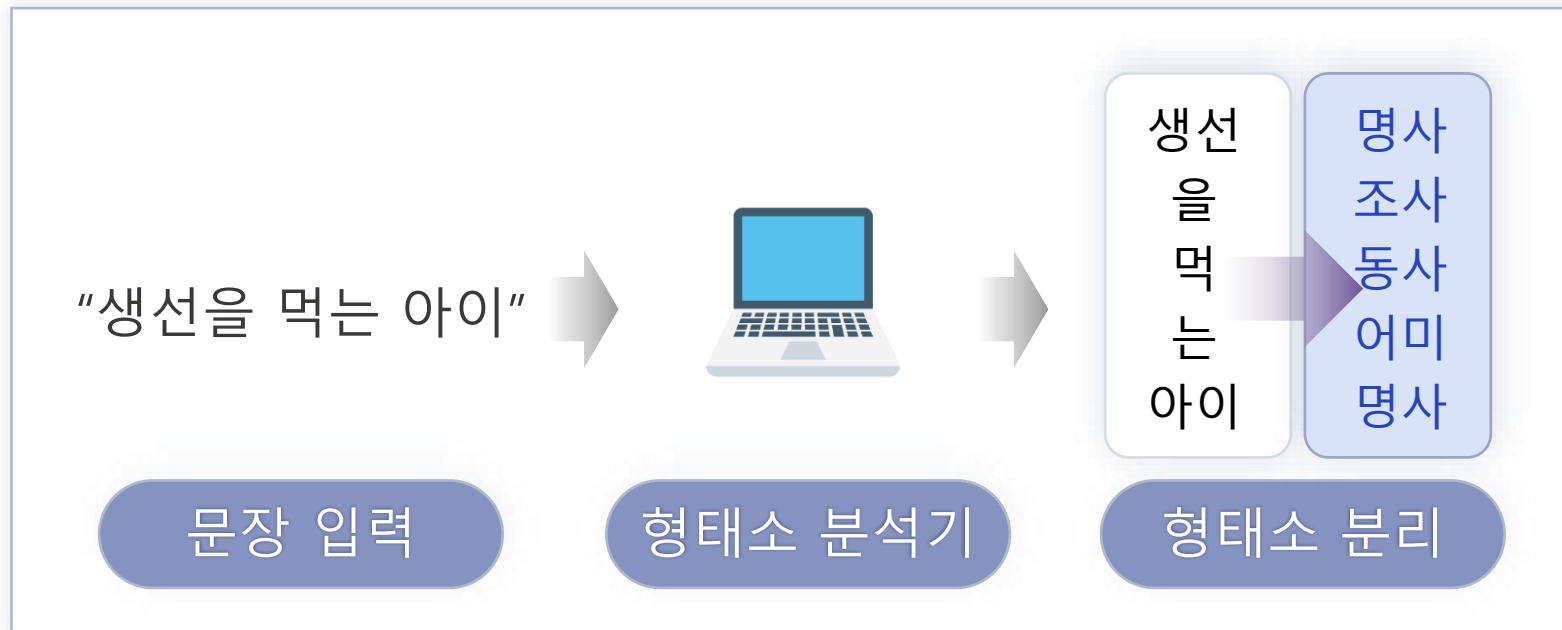
### 3. 자연어 처리의 과정



### 3. 자연어 처리의 과정

#### ① 형태소 분석

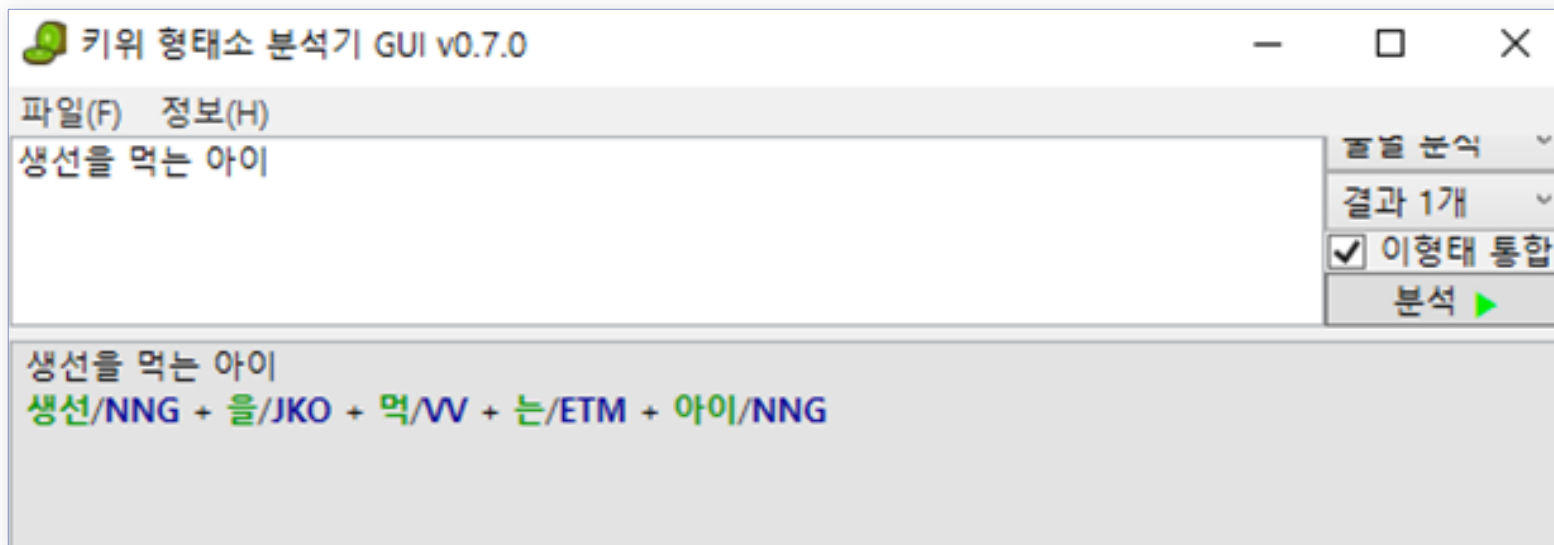
- » '형태소 분석(Morphological Analysis)'은 문장을 '형태소' 라는 최소 의미 단위로 분리하는 것
- » 예컨대 '생선을 먹는 아이' 라는 문장이 입력되면 아래와 같이 명사, 조사, 동사, 어미 등을 분리한다.



### 3. 자연어 처리의 과정

#### 1 형태소 분석

- » '형태소 분석(Morphological Analysis)'은 문장을 '형태소' 라는 최소 의미 단위로 분리하는 것
- » 예컨대 '생선을 먹는 아이' 라는 문장이 입력되면 아래와 같이 명사, 조사, 동사, 어미 등을 분리한다.





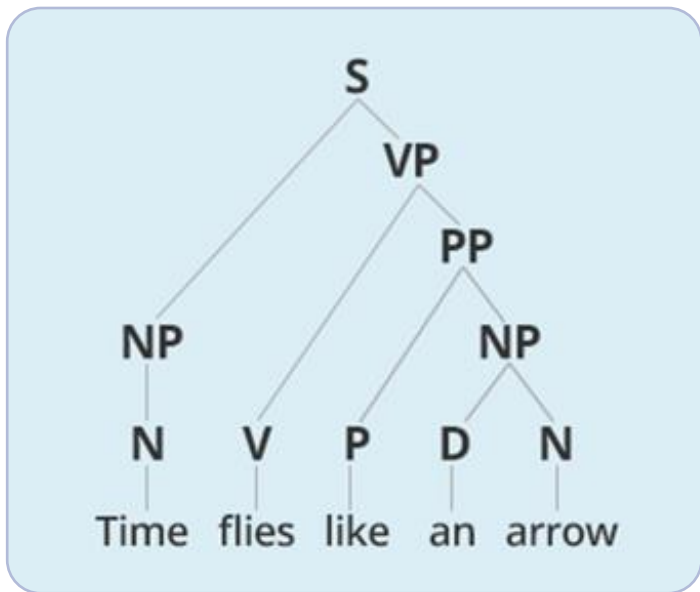
### 3. 자연어 처리의 과정

#### ② 구문 분석

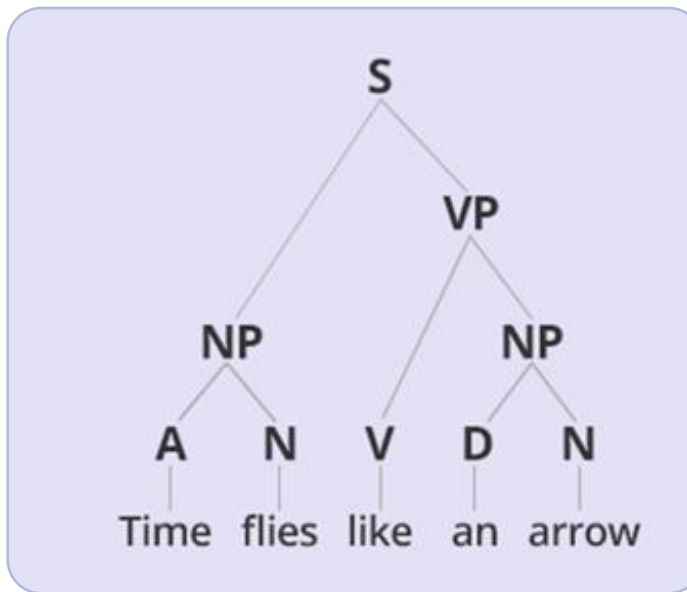
- » '구문 분석(Syntax Analysis)'은 문장의 구성요소를 분해하고, 그들 사이의 위계 관계를 분석해 문장의 구조를 찾아내는 것
- » 보통은 각각의 어절 단위로 나눠 Parsing tree를 이용해 해당하는 tag를 부여해서 분류함
- » 'Time flies like an arrow.'라는 문장을 구문 분석하는 경우, flies를 명사 또는 동사로 분석하는지, like를 동사 또는 전치사로 분석하는지에 따라 분석 결과가 확연히 달라짐

### 3. 자연어 처리의 과정

#### ② 구문 분석



"시간은 화살과 같이 날아  
간다"



"시간파리들은 화살을 좋아  
한다"

### 3. 자연어 처리의 과정

#### ③ 의미 분석

- » '의미 분석(Semantic Analysis)'은 문장의 뜻을 파악하는 작업
- » 의미 분석을 통해 문법에는 맞지만, 의미가 잘못된 문장을 파악하는 것

문법 규칙에는 맞지만 의미가 올바르지 않을 것을 파악하는 작업

① 사람이 사과를 먹는다.

② 사람이 비행기를 먹는다.

③ 비행기가 사과를 먹는다.

### 3. 자연어 처리의 과정

#### 4 담화 분석

- » 담화 분석(Discourse Analysis)은 문장을 전체 문맥과 연결하여 정확한 의미를 분석하는 작업
- » 예를 들어 “그는 울고 말았다.”라는 동일한 문장은 문맥에 따라 의미가 달라질 수 있음
- » 쉽게 말해 발화자의 의도에 맞도록 대화의 흐름을 파악하고 응답하게 해주는 기능이 담화 분석임

문장의 연관관계를 분석하여 대화의 흐름상 어떤 의미인지 파악하는 작업

민수가 도자기를 떨어뜨렸다. 그는 울고 말았다.

민수가 경기에서 우승을 했다. 그는 울고 말았다.

## 4. 단어 수치화

텍스트를 컴퓨터가 이해하고, 효율적으로 처리하게 하기  
위해서는 컴퓨터가 이해할 수 있도록 텍스트를 적절히 숫자로  
변환해야 함

단어 수치화를 거쳐 신경망 모델을 활용하여 자연어 처리가  
가능

단어를 표현하는 방법에 따라서 자연어 처리의 성능이 크게  
달라지기 때문에 단어를 수치화 하기 위한 많은 방법이  
존재

## 4. 단어 수치화

### 1 One-hot 인코딩

- » 단어별로 하나의 좌표축을 대응 시킨 공간에서,  
해당 되는 단어 위치에만 1을 설정하고 나머지에는 0을 설정하여,  
공간상에 단어를 표현, 단어 간의 유사도를 계산하기 곤란

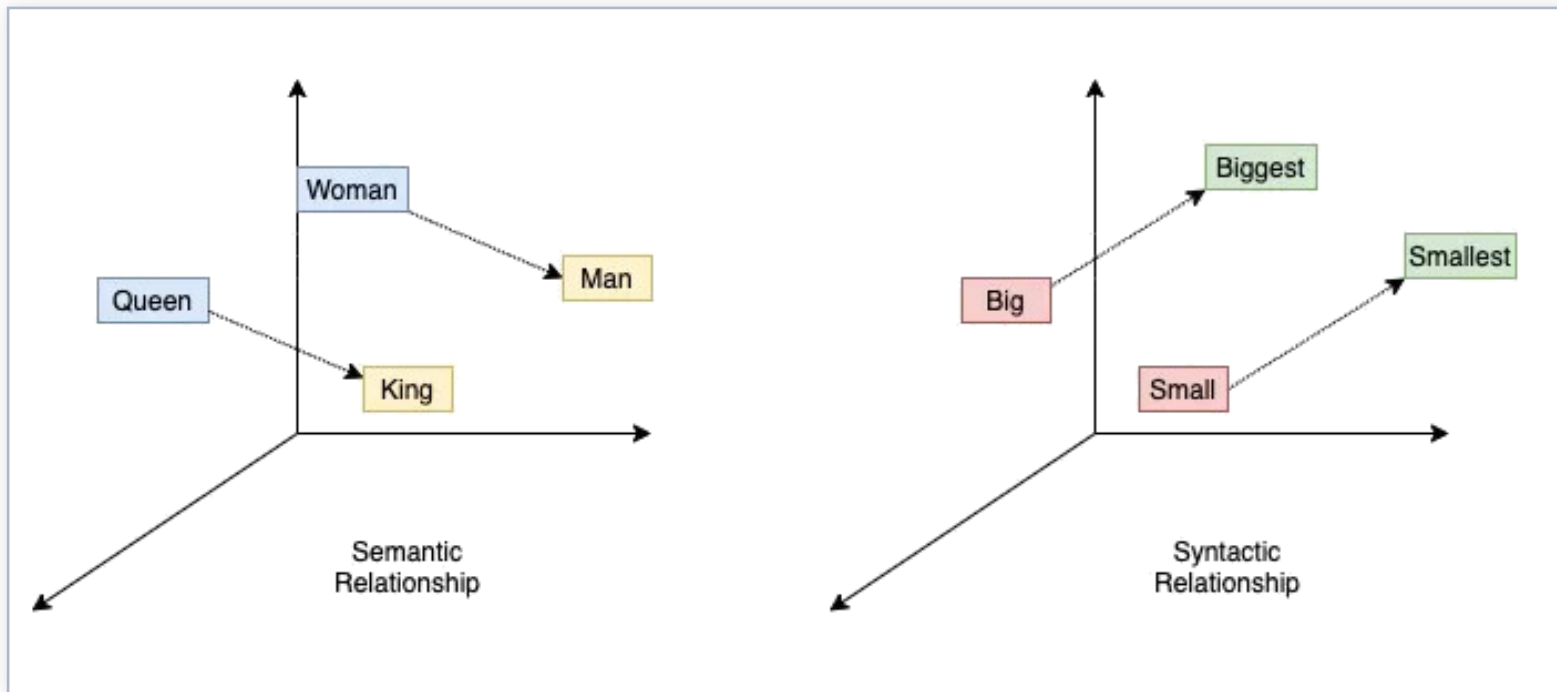
0	0	0	1	0	...	0	0	0	0
딸기	오이	참외	수박	배	...	복숭아	사과	귤	바나나

'수박' = (0,0,0,1,0,...,0,0,0,0)

## 4. 단어 수치화

### ② Word2Vec

» 단어의 의미를 충분히 잘 나타내도록 단어를 공간 상의 실수 벡터로 표현

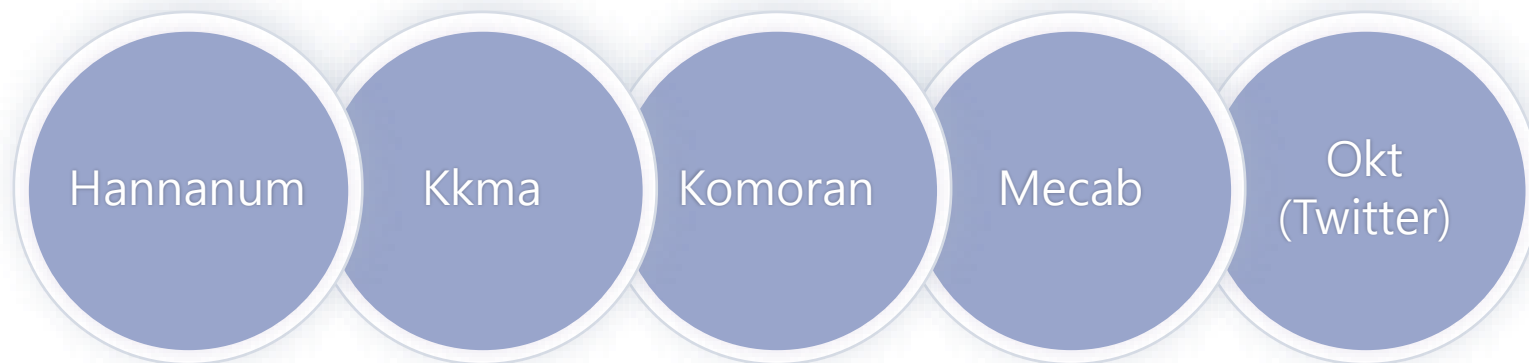


 <https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eccc30>

## 5. 자연어 처리를 위한 라이브러리

### 1 KoNLPy

- » **형태소 분석**을 할 수 있는 좋은 라이브러리들을 파이썬 라이브러리로 통합해서 사용할 수 있도록 하여 한국어 구문 분석을 쉽게 할 수 있도록 만들어진 라이브러리
- » 다양한 형태소 분석기들이 객체 형태로 포함돼 있으며 다음과 같은 각 형태소 분석기들이 있다.





## 5. 자연어 처리를 위한 라이브러리

### 1 KoNLPy

#### » 품사 태깅의 성능 분석

ex

"아버지가방에들어가신다"

Hannanum	Kkma	Komoran	Mecab	Twitter
아버지가방에 들어가 / N	아버지 / NNG	아버지가방에 들어가신다 / NNP	아버지 / NNG	아버지 / Noun
이 / J	가방 / NNG		가 / JKS	가방 / Noun
시ㄴ다 / E	에 / JKM		방 / NNG	에 / Josa
	들어가 / VV		에 / JKB	들어가신 / Ver b
	시 / EPH		들어가 / VV	다 / Eomi
	ㄴ다 / EFN		신다 / EP+EC	

## 5. 자연어 처리를 위한 라이브러리

### ② Gensim

- » Gensim은 파이썬에서 제공하는 자연어 처리 라이브러리
- » Word2Vec, 토픽 모델링, LDA 등의 자연어처리를 위한 패키지 제공
- » 말뭉치(Corpus)라고 불리는 단어 사전을 만들어 데이터를 학습시켜, 단어를 벡터화 시킴



개발 실무  
**실습하기**

P r a c t i c a l P r o g r a m m i n g f o r A I

형태소 분석 및 Word2Vec

## ① 한국어 분석이란?

우리가 일상적으로 사용하는 한국어를  
컴퓨터가 이해하는 방식을 사용하여  
분석하는 것

## ② 한국어 자연어 처리의 특징

교착어

평서문과 의문문의 불명확한 구분

띄어쓰기

주어 생략

어순의 자유로움

한자 기반의 언어

### ③ 자연어 처리의 과정

자연어 데이터 준비 → 형태소 분석  
→ 구문 분석 → 의미 분석 → 담화 분석

### ④ 단어 수치화

컴퓨터로 자연어 처리를 하기 위해 필요

단어 수치화를 거쳐 신경망 모델을 활용하여  
자연어 처리가 가능

## ⑤ 자연어 처리를 위한 라이브러리

KoNLPy

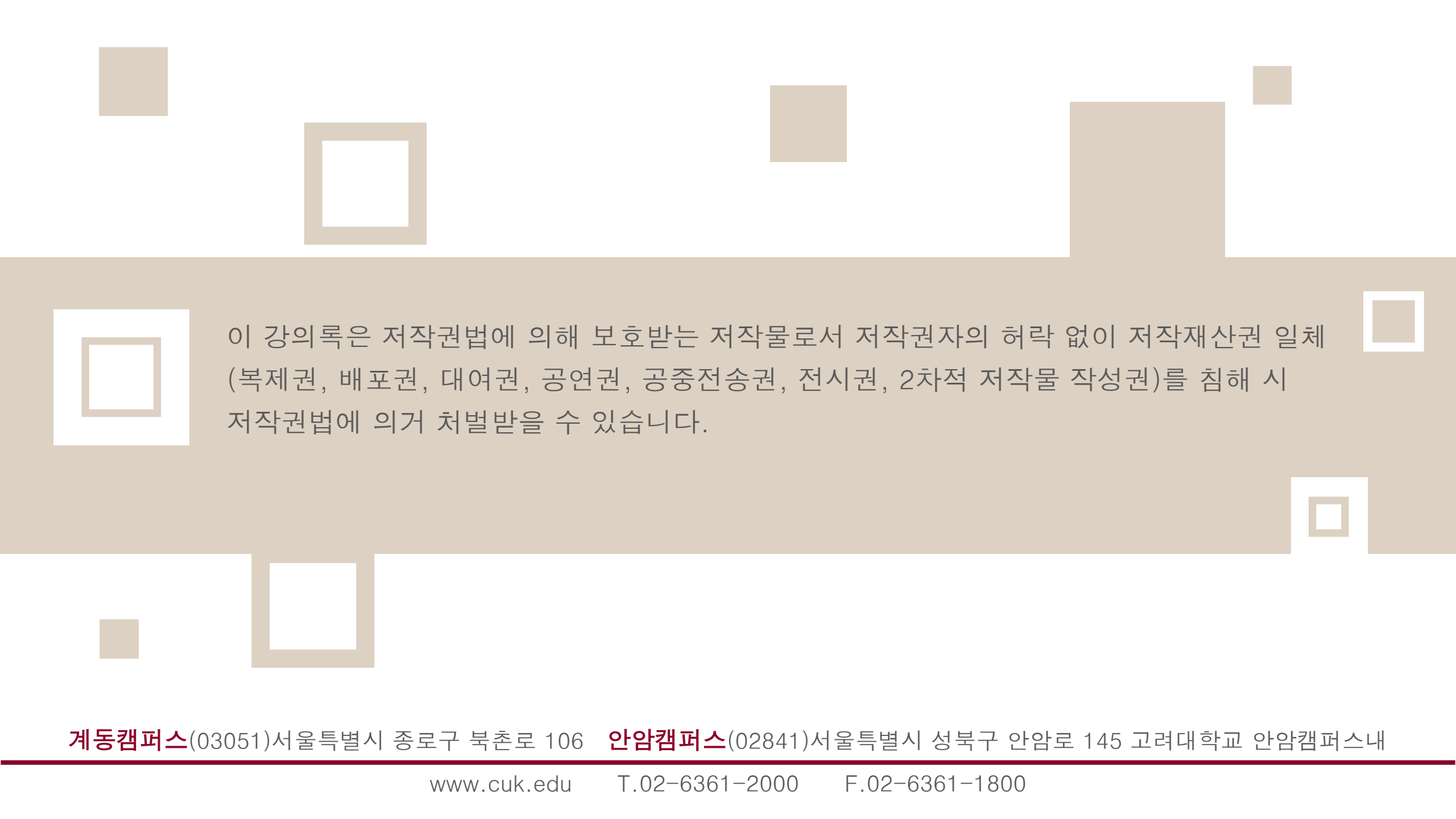
Gensim

## ⑥ 형태소 분석 및 Word2Vec 실습

한국어 분석

- ❏ 퍼블릭에이아이([www.public.co.kr](http://www.public.co.kr))
- ❏ 딥러닝을 이용한 자연어 처리 입문(<https://wikidocs.net/book/2155>)





이 강의록은 저작권법에 의해 보호받는 저작물로서 저작권자의 허락 없이 저작재산권 일체 (복제권, 배포권, 대여권, 공연권, 공중전송권, 전시권, 2차적 저작물 작성권)를 침해 시 저작권법에 의거 처벌받을 수 있습니다.