

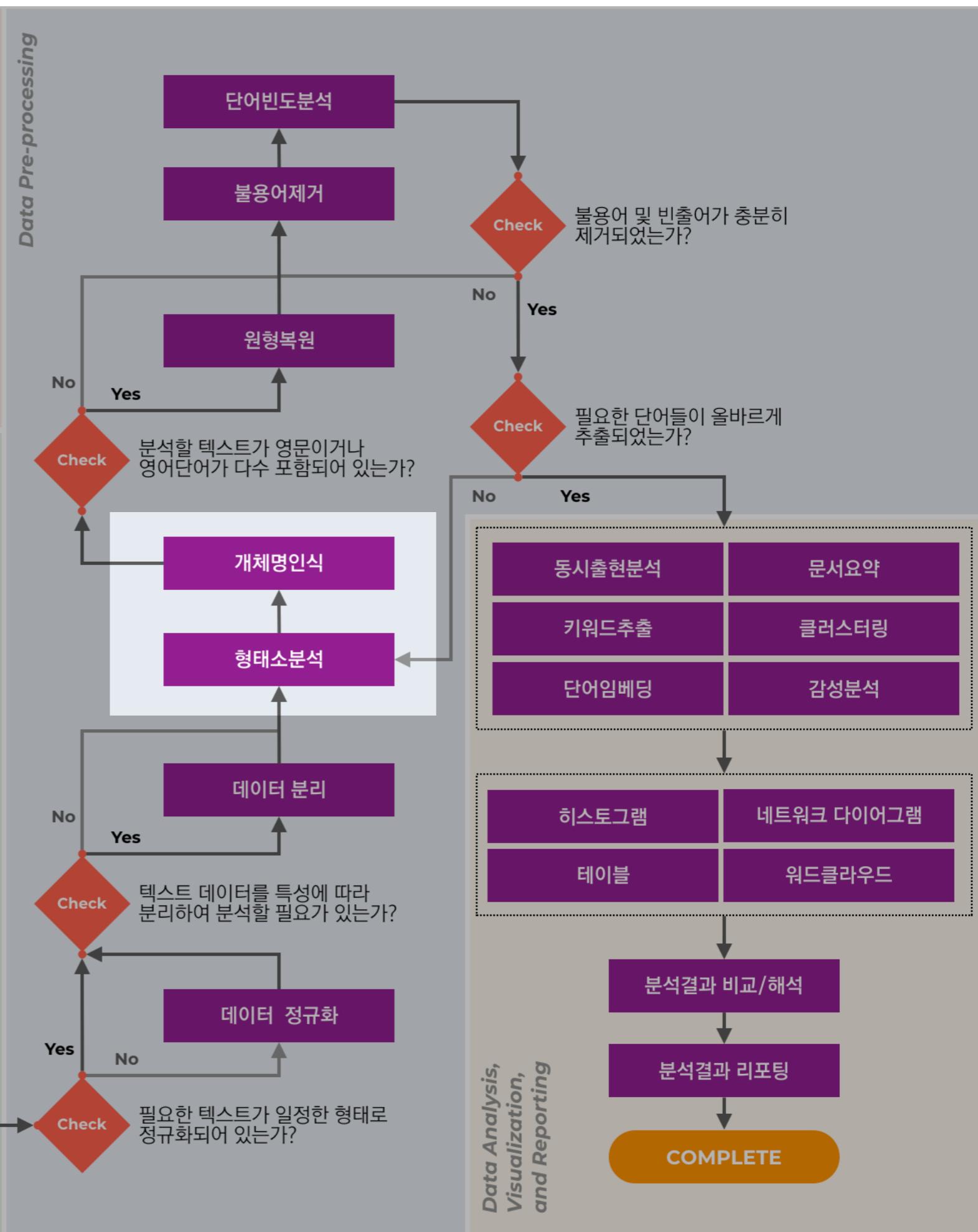
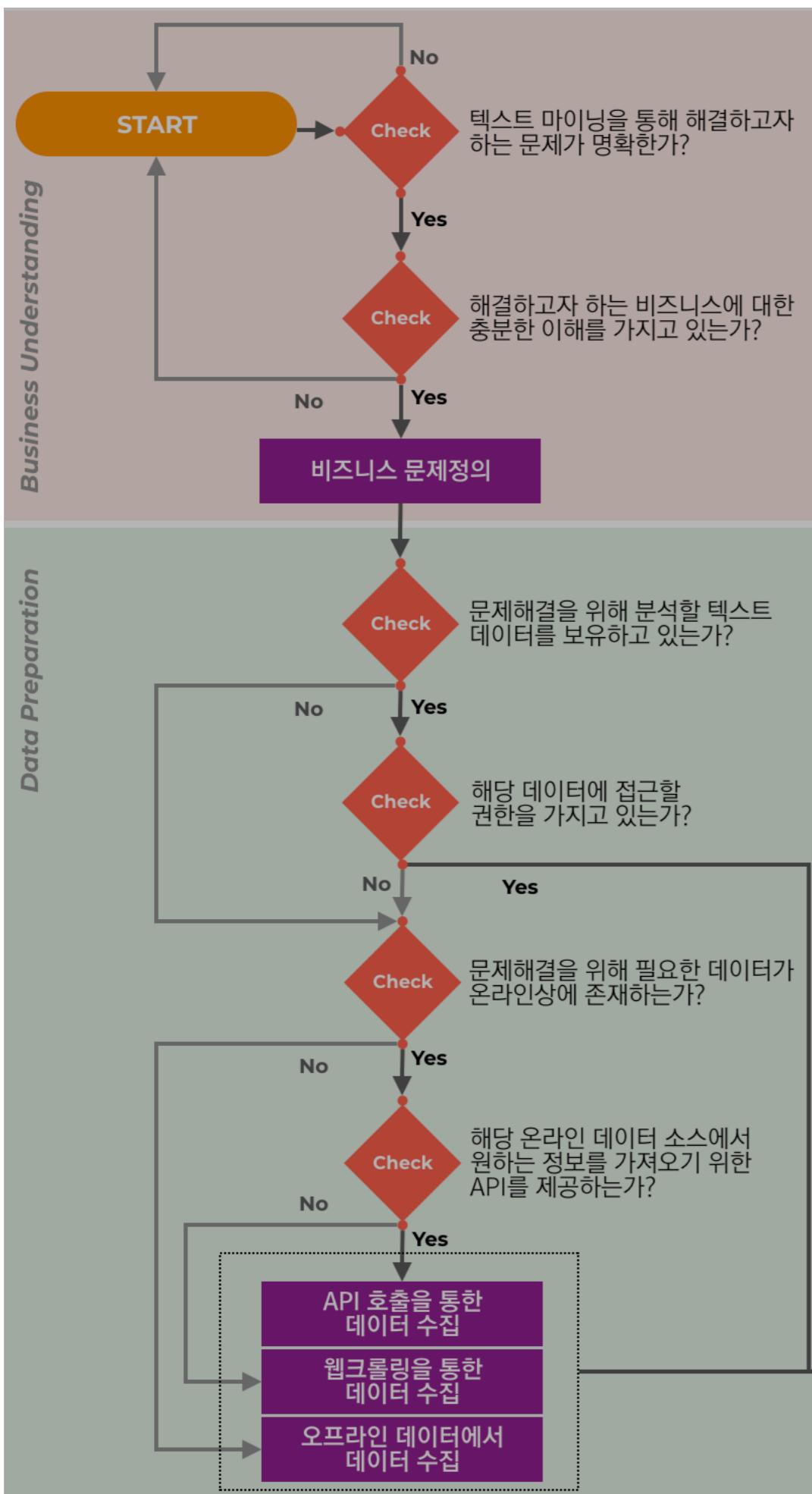
TEXT MINING for PRACTICE

Python을 활용한 비정형 데이터 분석 - WEEK 05

전병진 FINGEREDMAN (fingeredman@gmail.com)

Part 4.

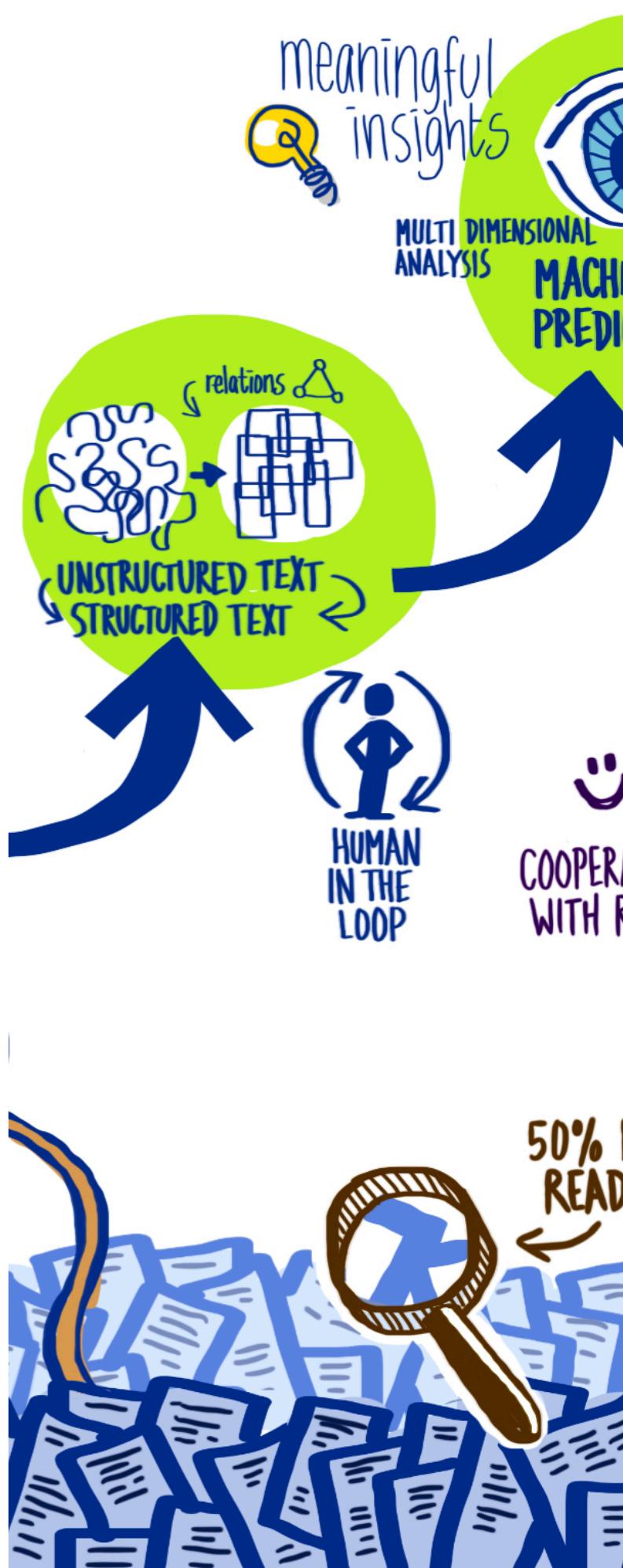
형태소분석 & 개체명인식



어휘분석 (Lexical Analysis)

자연어처리의 유형

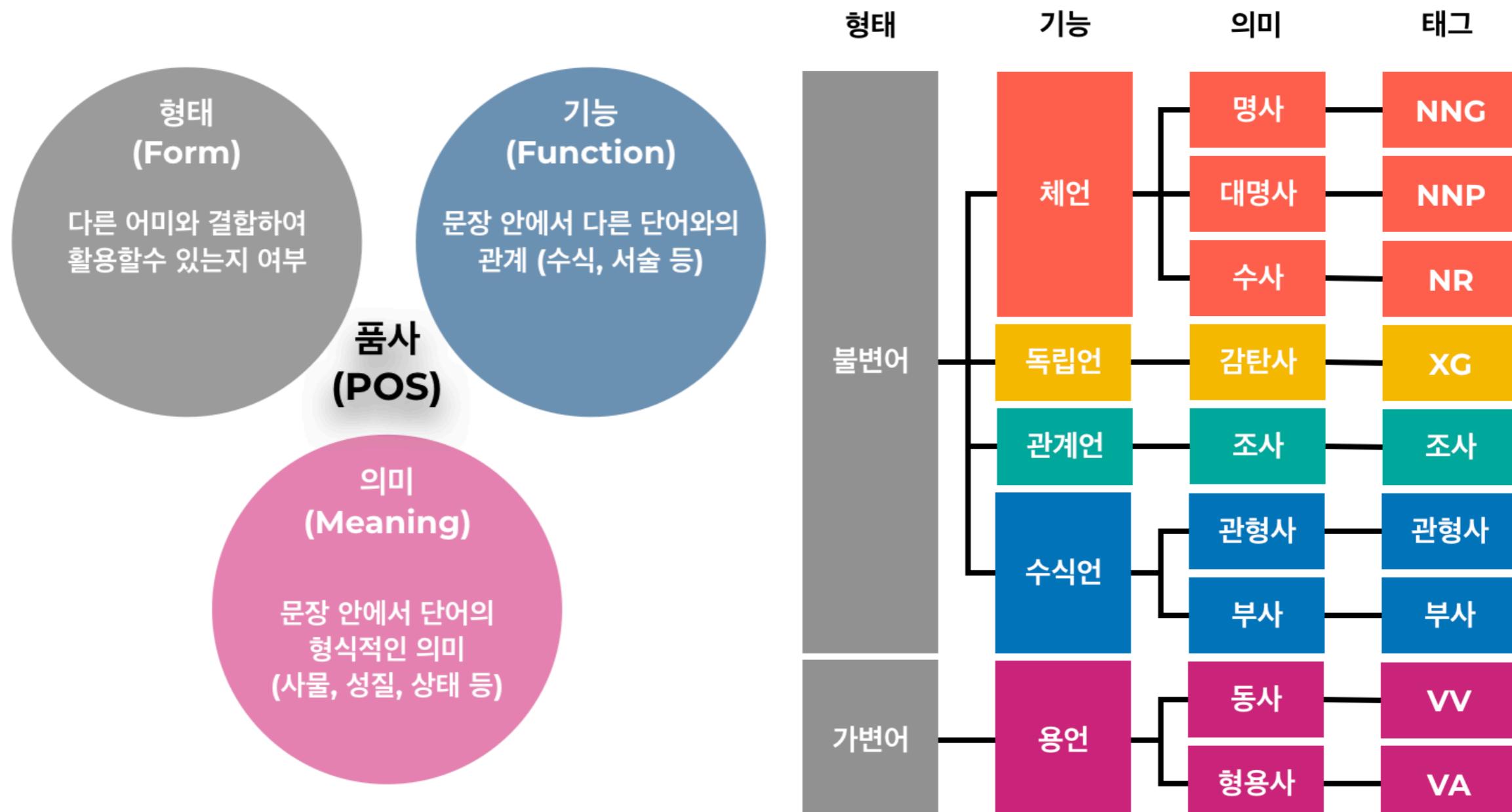
- ▶ Lexical (=Morphology) : The study of meaningful components of words
- ▶ Syntax : The study of structural relationships between words
- ▶ Semantics : The study of meaning
- ▶ Pragmatics : The study of how language is used to accomplish goals



한국어 품사구분

한국어의 5언 9품사

- ▶ 단어를 기능 (function), 의미 (meaning), 형태 (form)의 세 가지 기준에 의해 분류함



*Source : Daum 백과사전, 품사의 분류 기준, <http://igoindol.net/siteagent/100.daum.net/encyclopedia/view/24XXXXX49949/>.

**Source : for textmining, 한국어 품사 분류와 분포(distribution), 2017.4.21., <https://ratsgo.github.io/korean%20linguistics/2017/04/21/wordclass/>.

형태소 분석 (Part of Speech Tagging)

교착어, 굴절어, 그리고. 고립어

- ▶ 교착어 (agglutinative language) : 어근에 접사가 결합되어 각 단어의 기능을 나타내는 언어 (한국어, 일본어, 몽골어, ...)
- ▶ 굴절어 (inflectional language) : 단어 자체의 형태변화로 그 단어의 문법성을 나타내는 언어 (라틴어, 독일어, 러시아어, ...)
- ▶ 고립어 (isolating language) : 단어의 형태변화 없이 문법적 관계는 어순에 의해 정해지는 언어 (영어, 중국어, ...)

형태소 분석이란?

- ▶ 문장을 형태소 단위로 구분하고 품사를 구별하여 태깅하고 용언의 다양한 활용으로 인한 형태소 탈락현상을 복원하는 과정
- ▶ 분석기마다 형태소 구분 방식이 다르기 때문에 데이터에 맞는 분석기를 선택해야함
- ▶ 모든 언어의 자연어 처리 과정 중 가장 중요하고 기초적인 역할 수행
- ▶ 형태소 분석의 활용
 - 언어학적 측면 : 특정 언어현상의 생성과정을 설명하는 데 용이하게 쓰일 수 있음
 - 전산학적 측면 : 정보검색이나 자연어 처리 자동 처리시스템의 구문 분석의 전 단계 등의 용도로 쓰일 수 있음

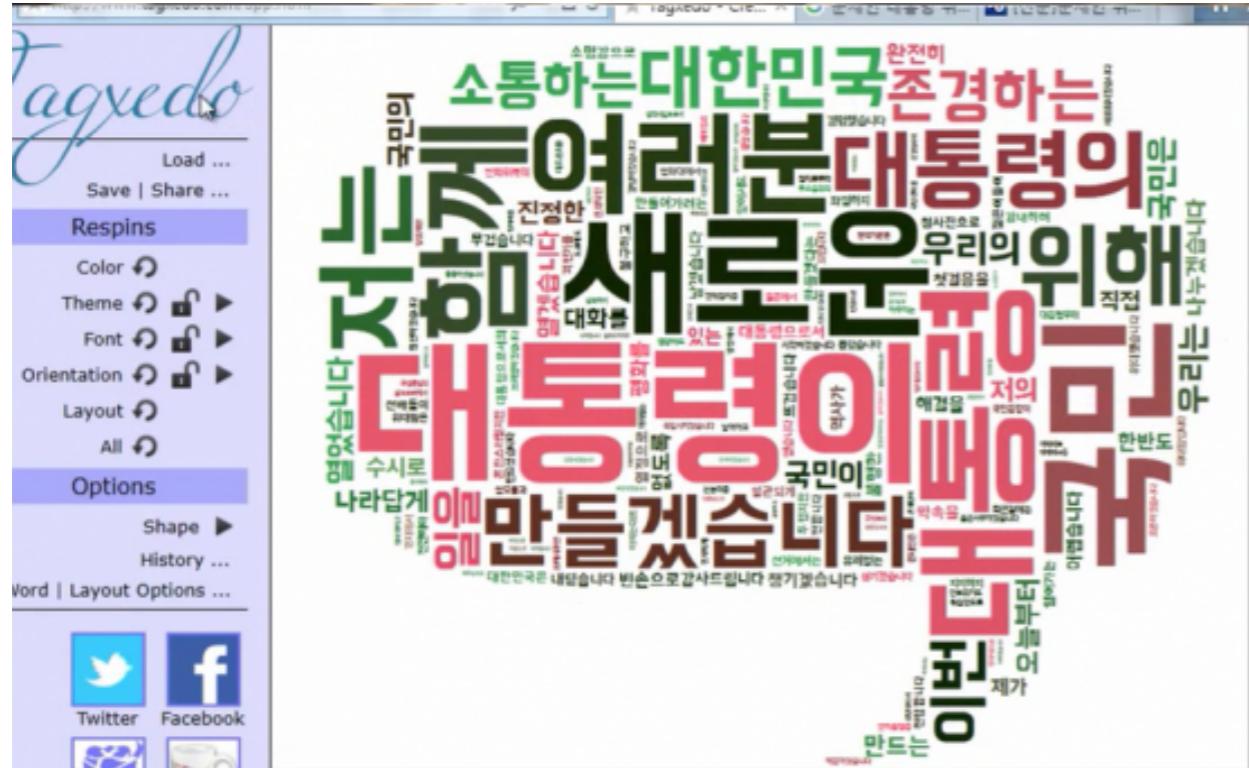
구분	내용
원문	<ul style="list-style-type: none">. 여러분 안녕하세요. 재미있는 텍스트 마이닝 수업입니다.
형태소 분석	<ul style="list-style-type: none">. 여러분/NP + 안녕/NNG + 하세요/EF + ./SF. 재미있/VA + 는/ETM + 텍스트/NNG + 마이닝/NNG + 수업/NNG + 입니다/EF + ./SF

형태소 분석 (Part of Speech Tagging)

대분류	세종 품사 태그		KKMA 단일 태그 V 1.0						
	태그	설명	묶음 1	묶음 2	태그	설명	활용태그	저장사전	
체언	NNG	일반 명사	N	NN	NNG	보통 명사	simple.dic	noun.dic	
	NNP	고유 명사			NNP	고유 명사			
	NNB	의존 명사			NNB	일반 의존 명사			
	NNM	단위 의존 명사			NNM	단위 의존 명사			
	NR	수사		NR	NR	수사			
용언	NP	대명사			NP	대명사	verb.dic	simple.dic	
	VV	동사	V	VV	VV	동사			
	VA	형용사			VA	형용사			
	VX	보조 용언		VX	VXV	보조 농사	raw.dic		
	VCP	긍정 지정사			VXA	보조 형용사			
관형사	VCN	부정 지정사	M	VC	VCP	긍정 지정사, 서술격 조사 '이다'			
	MM	관형사			VCN	부정 지정사, 형용사 '아니다'			
부사	MAG	일반 부사	MA	MD	MDT	일반 관형사	simple.dic	simple.dic	
	MAJ	접속 부사			MDN	수 관형사			
감탄사	IC	감탄사	I	IC	IC	감탄사	IC		
조사	JKS	주격 조사	J	JK	JKS	주격 조사	JKS	simple.dic	
	JKC	보격 조사			JKC	보격 조사	JKC		
	JKG	관형격 조사			JKG	관형격 조사	JKG		
	JKO	목적격 조사			JKO	목적격 조사	JKO		
	JKB	부사격 조사			JKM	부사격 조사	JKM		
	JKV	호격 조사			JKI	호격 조사	JKI		
	JKQ	인용격 조사			JKQ	인용격 조사	JKQ		
	JX	보조사		JC	JX	보조사	JX		
	JC	접속 조사			JC	접속 조사	JC		
선어말 어미	EP	선어말 어미	EP	EPH	존칭 선어말 어미	EP	raw.dic	simple.dic	
				EPT	시제 선어말 어미				
				EPP	공손 선어말 어미				

대분류	세종 품사 태그		KKMA 단일 태그 V 1.0						
	태그	설명	묶음 1	묶음 2	태그	설명	활용태그	저장사전	
어말 어미	EF	종결 어미	E	EF	EFN	평서형 종결 어미	EF	simple.dic	
	EFQ	의문형 종결 어미			EFO	명령형 종결 어미			
	EFA	청유형 종결 어미			EFI	감탄형 종결 어미			
	EFR	존칭형 종결 어미			ECE	대등 연결 어미	EC		
	ECD	의존적 연결 어미			ECS	보조적 연결 어미			
	ETN	명사형 전성 어미			ETM	관형형 전성 어미			
접두사	XPN	체언 접두사	XP	XP	XPN	체언 접두사	XPN	simple.dic	
	XPV	용언 접두사			XSV	동사 파생 접미사	XSV		
접미사	XSN	명사 파생 접미사	XS	XS	XSA	형용사 파생 접미사	XSA		
	XSV	동사 파생 접미사			XSM	부사 파생 접미사	XSM		
	XSA	형용사 파생 접미사			XSO	기타 접미사	XSO		
					XR	어근	XR		
					SF	마침표물음표, 느낌표	SF		
부호	SP	쉼표, 가운뎃점, 콜론, 빛금	S	SE	SS	따옴표, 괄호표, 줄표	SS	Symbol class	
	SE	줄임표			SO	불임표(물결, 숨김, 빠짐)	SO		
	SW	기타기호 (논리수학기호, 화폐기호)			SW	기타기호 (논리수학기호, 화폐기호)	SW		
	NF	명사추정범주		U	UN	명사추정범주	NNA		
	NV	용언추정범주			UV	용언추정범주	N/A		
	NA	분석불능범주			UE	분석불능범주	N/A		
한글 이외	SL	외국어	O	OL	OL	외국어	NNA	N/A	
	SH	한자			OH	한자	NNA		
	SN	숫자		ON	ON	숫자	NR		

형태소 분석 (Part of Speech Tagging)



번역

실현

Python 한국어 형태소 분석기

① 고꼬마 형태소 분석기: Kkma

- ▶ 서울대학교 IDS (Intelligent Data Systems) 연구실에서 자연어 처리를 위한 모듈구축과제로 개발한 형태소 분석기
- ▶ Java 언어를 기반으로 하며, Python-Java 연동을 통해 Python에서 사용 가능함
- ▶ 동적 프로그래밍 (Dynamic Programming) 방식으로 가능한 모든 형태소 후보를 모두 찾아 가장 적합한 형태소를 판단함 → 매우느림

Python 형태소 분석 예시

```
from konlpy.tag import Kkma
```

```
text = “나 정말 이거 엔시티 팬 분들께 나눠 드리고 싶은데 ㅠㅠ 아무도 알티 안 해 주실 것 같아 ㅠㅠ”
```

```
...
```

```
kkma = Kkma()
```

```
pos_result = kkma.morphs(text)
```

```
...
```

Result : [나/NP 정말/MAG 이거/NP 엔/NNG 시티/NNG 팬/NNG 분/NNG 들/XSN 께/JKM 나누/VV
어/ECS 드리/VXV 고/ECE 싶/VXA 은데/ECD ㅠㅠ/EMO 아무/NP 도/JX 알/VV ㄹ/ETD 티/NNG 안/
MAG 하/VV 어/ECS 주/VXV 시/EPH ㄹ/ETD 것/NNB 같/VA 아/ECD ㅠㅠ/EMO]

Python 한국어 형태소 분석기

② 한나눔 형태소 분석기: Hannanum

- ▶ KAIST Semantic Web Research Center (SWRC)에서 개발한 형태소 분석기
- ▶ 자동 띄어쓰기 모듈을 제공해 형태소 분석 결과를 활용하여 한글 문장에 대한 자동 띄어쓰기 수행 가능
- ▶ 사전 기반의 맞춤법 교정 모듈로 형태소 분석 결과를 활용하여 한글 단어에 대한 맞춤법 교정 수행 가능

Python 형태소 분석 예시

```
from konlpy.tag import Hannanum
```

```
text = “나 정말 이거 엔시티 팬 분들께 나눠 드리고 싶은데 ㅠㅠ 아무도 알티 안 해 주실 것 같아 ㅠㅠ”
```

```
...
```

```
hannanum = Hannanum()  
pos_result = hannanum.morphs(text)
```

```
...
```

Result : [나/N 정말/M 이거/N 엔시티/N 팬/P 를/E 불/P 를/E 들/N 께/J 나누/P 어/E 드리/P 고/E 싶/E
은/E 데/N ㅠㅠ/N 아무/N 도/J 알티/N 안/M 하/P 어/E 주/P 시르/E 것/N 같/P 아/E ㅠㅠ/N]

Python 한국어 형태소 분석기

③ 코모란 형태소 분석기: Komoran

- ▶ 서울대학교 IDS (Intelligent Data Systems) 연구실에서 자연어 처리를 위한 모듈구축과제로 제작한 형태소 분석기
- ▶ Shineware에서 개발된 한국어 형태소 분석기로서 Java Library 형태(jar)로 제공됨
- ▶ 타 형태소 분석기와 달리 여러 어절을 하나의 품사로 분석 가능함으로써 공백이 포함된 고유명사(영화 제목, 음식점명 등)를 정확하게 분석

Python 형태소 분석 예시

```
from konlpy.tag import Komoran
```

```
text = “나 정말 이거 엔시티 팬 분들께 나눠 드리고 싶은데 ㅠㅠ 아무도 알티 안 해 주실 것 같아 ㅠㅠ”
```

```
...
```

```
komoran = Komoran()
```

```
pos_result = komoran.morphs(text)
```

```
...
```

Result : [나/NP 정말/MAG 이것/NP 엔/NNB 시티/NNP 팬/NNG 분/NNB 들/XSN 께/JKB 나누/VV
어/EC 드리/VV 고/EC 싶/VX 은데/EC ㅠㅠ/NA 아무도/NNP 알티/NNG 안/MAG 하/VV 아/EC 주/VX
시/EP ㄹ/ETM 것/NNB 같/VA 아/EC ㅠㅠ/NA]

Python 한국어 형태소 분석기

④ 은전한닢 형태소 분석기: Mecab

- ▶ 검색에서 쓸만한 오픈소스 한국어 형태소 분석기를 목적으로 개발된 한국어 형태소 분석기
- ▶ 오픈소스 검색엔진 Elasticsearch에 적용되어 활용되고 있음
- ▶ 사용자 사전 등록기능을 제공하여 다양한 도메인에서 생성되는 단어들을 인식할 수 있도록 도와줌

Python 형태소 분석 예시

```
from konlpy.tag import Mecab
```

```
text = “나 정말 이거 엔시티 팬 분들께 나눠 드리고 싶은데 ㅠㅠ 아무도 알티 안 해 주실 것 같아 ㅠㅠ”
```

```
...
```

```
mecab = Mecab()
```

```
pos_result = mecab.morphs(text)
```

```
...
```

Result : [나/NP 정말/MAG 이거/NP 엔시/NNG 티/NNG 팬/VV+ETM 분/NNB 들/XSN 께/JKB 나
눠/VV+EC 드리/VX 고/EC 싶/VX 은데/EC ㅠㅠ/UNKNOWN 아무/NP 도/JX 알/VV 티/EC 안/MAG
해/VV+EC 주/VX 실/EP+ETM 것/NNB 같/VA 아/EC ㅠㅠ/UNKNOWN]

Python 한국어 형태소 분석기

⑤ 카이 형태소 분석기: Khaiii (Kakao Hangul Analyzer III)

- ▶ 카카오에서 DHA2 (Daumkakao Hangul Analyzer 2)를 계승하여 개발하고 2018년 공개된 두 번째 버전의 형태소분석기
- ▶ 속도를 매우 중요시하여 신경망 알고리즘들 중에서 Convolutional Neural Network (CNN)을 사용하여 개발됨
- ▶ 사용자 사전 등록기능을 제공하여 다양한 도메인에서 생성되는 단어들을 인식할 수 있도록 도와줌

Python 형태소 분석 예시

```
from khaiii import KhaiiiApi
```

```
text = “나 정말 이거 엔시티 팬 분들께 나눠 드리고 싶은데 ㅠㅠ 아무도 알티 안 해 주실 것 같아 ㅠㅠ”
```

```
...
```

```
api = KhaiiiApi()
```

```
pos_result = api.analyze(text)
```

```
...
```

Result : [나/NP 정말/MAG 이거/NP 엔시/NNG + 티/NNP 팬/NNG 분/NNB + 들/XSN + 께/JKB 나
누/VV + 어/EC 드리/VX + 고/EC 싶/VX + 은데/EC ㅠㅠ/NNG 아무/NP + 도/JX 알티/NNG 안/MAG
하/VV + 여/EC 주/VX + 시/EP + 르/ETM 것/NNB 같/VA + 아/EC ㅠㅠ/NNG]

Python 한국어 형태소 분석기

⑥ 트위터 형태소 분석기: Twitter (Okt)

- ▶ 트위터에서 개발한 한국어 형태소 분석기
- ▶ SNS에서 발생하는 언어에서 자주 발생하는 인물명, 신조어 등을 잘 인식하는 편이며, 속도가 빠르지만 형태소 분석 품질은 상대적으로 낮음

Python 형태소 분석 예시

```
from konlpy.tag import Twitter
```

```
text = “나 정말 이거 엔시티 팬 분들께 나눠 드리고 싶은데 ㅠㅠ 아무도 알티 안 해 주실 것 같아 ㅠㅠ”
```

```
...
```

```
twitter = Twitter()
```

```
pos_result = twitter.morphs(text)
```

```
...
```

Result : [나/Noun 정말/Noun 이/Determiner 거/Noun 엔시/Noun 티/Noun 팬/Noun 분/Noun
들/Suffix 께/Josa 나눠/Verb 드리고/Verb 싶은데/Verb ㅠㅠ/KoreanParticle 아무/Noun 도/Josa
알티/Noun 안/Noun 해/Noun 주실/Verb 것/Noun 같아/Adjective ㅠㅠ/KoreanParticle]

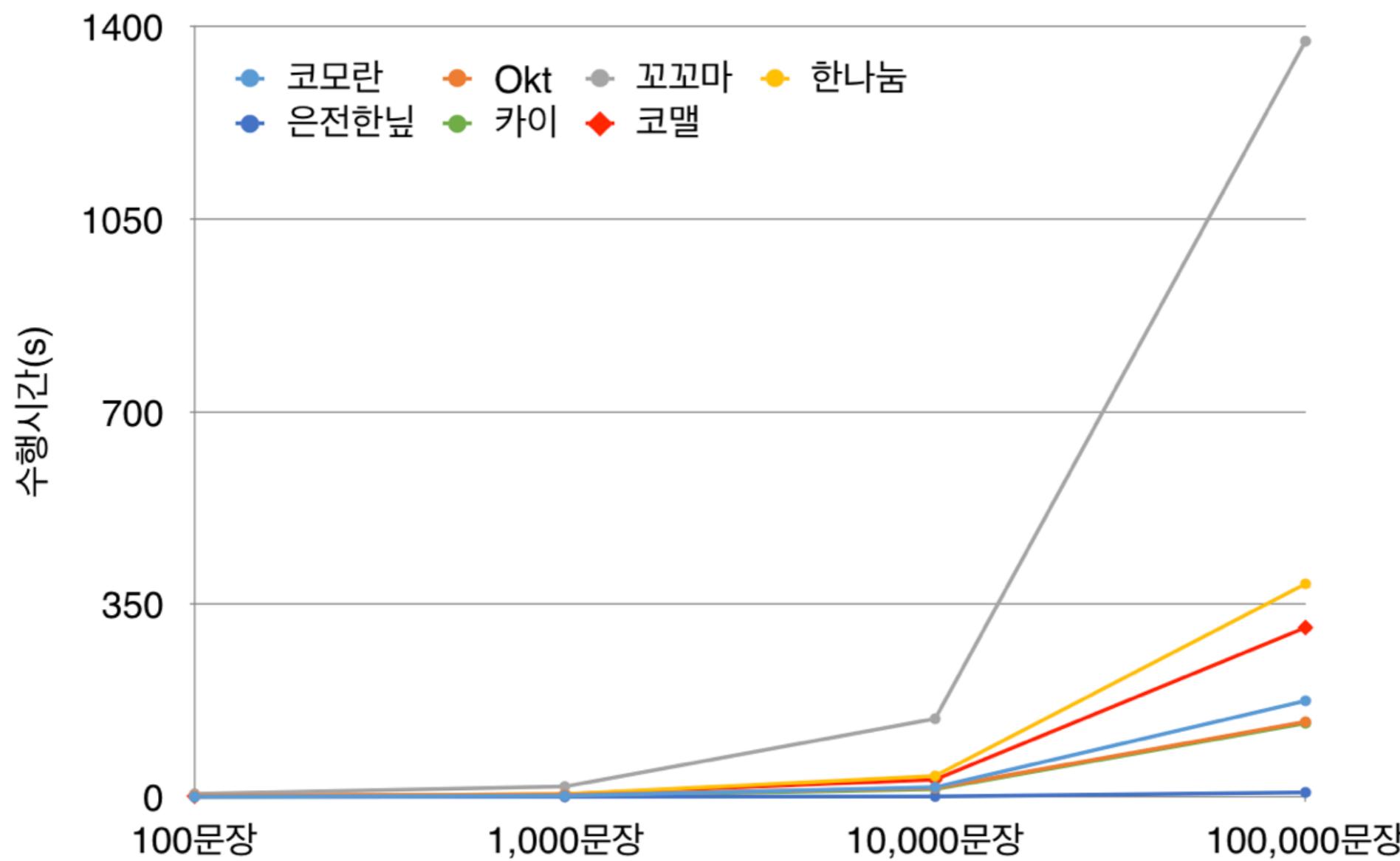
형태소 분석기 비교

형태소 분석 결과 비교

구분	형태소 분석 결과
원문	나 정말 이거 엔시티 팬 분들께 나눠 드리고 싶은데 ㅠㅠ 아무도 알티 안 해 주실 것 같아 ㅠㅠ
코모란	나/NP 정말/MAG 이것/NP 엔/NNB 시티/NNP 팬/NNG 분/NNB 들/XSN 께/JKB 나누/VV 어/EC 드리/VV 고/EC 싶/VX 은데/EC ㅠㅠ/NA 아무도/NNP 알티/NNG 안/MAG 하/VV 아/EC 주/VX 시/EP 르/ETM 것/NNB 같/VA 아/EC ㅠㅠ/NA
Okt	나/Noun 정말/Noun 이/Determiner 거/Noun 엔시/Noun 티/Noun 팬/Noun 분/Noun 들/Suffix 께/Josa 나눠/Verb 드리고/Verb 싶은데/Verb ㅠㅠ/KoreanParticle 아무/Noun 도/Josa 알티/Noun 안/Noun 해/Noun 주실/Verb 것/Noun 같아/Adjective ㅠㅠ/KoreanParticle
꼬꼬마	나/NP 정말/MAG 이거/NP 엔/NNG 시티/NNG 팬/NNG 분/NNG 들/XSN 께/JKM 나누/VV 어/ECS 드리/VXV 고/ECE 싶/VXA 은데/ECD ㅠㅠ/EMO 아무/NP 도/JX 알/VV 르/ETD 티/NNG 안/MAG 하/VV 어/ECS 주/VXV 시/EPH 르/ETD 것/NNB 같/VA 아/ECD ㅠㅠ/EMO
한나눔	나/N 정말/M 이거/N 엔시티/N 패/P ㄴ/E 불/P ㄴ/E 들/N 께/J 나누/P 어/E 드리/P 고/E 싶/P 은/E 데/N ㅠㅠ/N 아무/N 도/J 알티/N 안/M 하/P 어/E 주/P 시르/E 것/N 같/P 아/E ㅠㅠ/N
은전한닢	나/NP 정말/MAG 이거/NP 엔시/NNG 티/NNG 팬/VV+ETM 분/NNB 들/XSN 께/JKB 나눠/VV+EC 드리/VX 고/EC 싶/VX 은데/EC ㅠㅠ/UNKNOWN 아무/NP 도/JX 알/VV 티/EC 안/MAG 해/VV+EC 주/VX 실/EP+ETM 것/NNB 같/VA 아/EC ㅠㅠ/UNKNOWN
카이	나/NP 정말/MAG 이거/NP 엔시/NNG + 티/NNP 팬/NNG 분/NNB + 들/XSN + 께/JKB 나누/VV + 어/EC 드리/VX + 고/EC 싶/VX + 은데/EC ㅠㅠ/NNG 아무/NP + 도/JX 알티/NNG 안/MAG 하/VV + 여/EC 주/VX + 시/EP + 르/ETM 것/NNB 같/VA + 아/EC ㅠㅠ/NNG

형태소 분석기 비교

수행시간 비교 (Time Analysis)



개체명 인식 (Named Entity Recognition)

문장에서 하나의 개체로써 인식되어야하는 단어를 구별하는 과정

- ▶ 데이터에서 개체명을 구별하고 태깅함(지명, 사명, 인물명, 약자, 기관명 등)
- ▶ 사전 기반의 개체명 인식에서 개체명은 매일 새롭게 생겨나고 변형되므로, 개체명 사전을 유지하는 것이 매우 중요함
- ▶ 분석의 목적에 따라서 머신러닝 기반의 개체명 인식을 사용할 수 있으나 새로 생겨나거나 변형되는 단어에 취약함

#KoNLPy 형태소 분석

```
from konlpy.tag import Kkma
```

```
text = “호날두 한명이 주는 효과가 세리에 전체 인기도 영향을 미치다니.. 역시 개드립월클의 힘”
```

```
...
```

```
kkma = Kkma()
```

```
pos_result = kkma.pos(text)
```

```
...
```

Result :

```
[(호, NNG), (날, NNG), (두, MDN), (한명, NNG), (이, JKS), (줄,VV), (는, ETD), (효과, NNG), ... ,  
(세리, NNG), (에, JKM), (전체, NNG), (인기도, NNG), (영향, NNG), (을, JKO), (미, NNG), ... ,  
(역시, MAG), (개, NNG), (드립, UN), (월, NNM), (크, VA), (ㄹ, ETD), (의, NNG), (힘, NNG)]
```

개체명 인식 (Named Entity Recognition)

문장에서 하나의 개체로써 인식되어야하는 단어를 구별하는 과정

- ▶ 데이터에서 개체명을 구별하고 태깅함(지명, 사명, 인물명, 약자, 기관명 등)
- ▶ 사전 기반의 개체명 인식에서 개체명은 매일 새롭게 생겨나고 변형되므로, 개체명 사전을 유지하는 것이 매우 중요함
- ▶ 분석의 목적에 따라서 머신러닝 기반의 개체명 인식을 사용할 수 있으나 새로 생겨나거나 변형되는 단어에 취약함

#Eucalyptus 형태소 분석

```
from Eucalyptus.NerTagger import NerTagger
```

```
input_file, output_file = "output_pos.txt", "output_ner.txt"  
ner_tagger = euc.NerTagger(input_file, output_file)  
ner_tagger.tagging()
```

Result (output_ner.txt):

...

{result:

```
[(호날두, NNP, Person), (한명, NNG), (0|, JKS), (줄, VV), (는, ETD), (효과, NNG), (가, JKS),  
(세리에, NNP, Sports), (에, JKM), (전체, NNG), (인기도, NNG), (영향, NNG), (을, JKO), ... ,  
(역시, MAG), (개드립월클, NNG, Neologism), (의, NNG), (힘, NNG)]}
```

개체명 사전 (NER Corpus)

[단순 개체명 사전]

구분	의학	인물	고유명사	블록체인
1	불량 식품	사나	서울플랜트엔지니어링	블록체인
2	진행 암	쯔위	서울플리머	블럭체인
3	전진 피판	정연	서울피브이시상사	비트코인
4	유해 효과	나연	서울피브이씨	이더리움
5	무력증	황민현	서울피비씨	알트코인
6	유산소 운동	강다니엘	서울피앤씨	추격매수
7	산소 호흡	옹성우	서울하이테크	풀매수
8	공기 삼킴증	전병진	서울학연구	총알
9	분무제	진상형	고광엔지니어링	운전수
10	에어로졸	서지석	서울합금	고점
11	분무 주입법	배현진	서울합판	저점
12	대기 요법	현빈	서울합판목재상사	장투
13	정동 장애	진세연	서울합판상사	단타
14	정감성	남지현	서울해체산업	떡상
15	정동성	주상욱	서울행정신문사	떡락
16	들신경	김태희	서울행정학회	횡보
17	협력 병원	허맹호	서울화성	손절
18	친화력	유아인	서울화인테크	익절
19	친화 크로마토그래피	이승기	서울화학	반등
20	무섬유소원 혈증	한예슬	고광훈	패닉셀

[부가정보를 포함하는 개체명 사전]

구분	지역명	영문 지역명	구분
1	서울	Seoul	Metropolitan
2	종로	Jongno	district
3	중	Jung	district
4	용산	Yongsan	district
5	성동	Seongdong	district
6	광진	Gwangjin	district
7	동대문	Dongdaemun	district
8	중랑	Jungnang	district
9	성북	Seongbuk	district
10	강북	Gangbuk	district
11	도봉	Dobong	district
12	노원	Nowon	district
13	은평	Eunpyeong	district
14	서대문	Seodaemun	district
15	마포	Mapo	district
16	양천	Yangcheon	district
17	강서	Gangseo	district
18	구로	Guro	district
19	금천	Gumcheon	district
20	영등포	Yeongdeungpo	district

미등록단어 주출

형태소 분석과 개체명 인식은 새로운 단어를 인식하기가 어려움

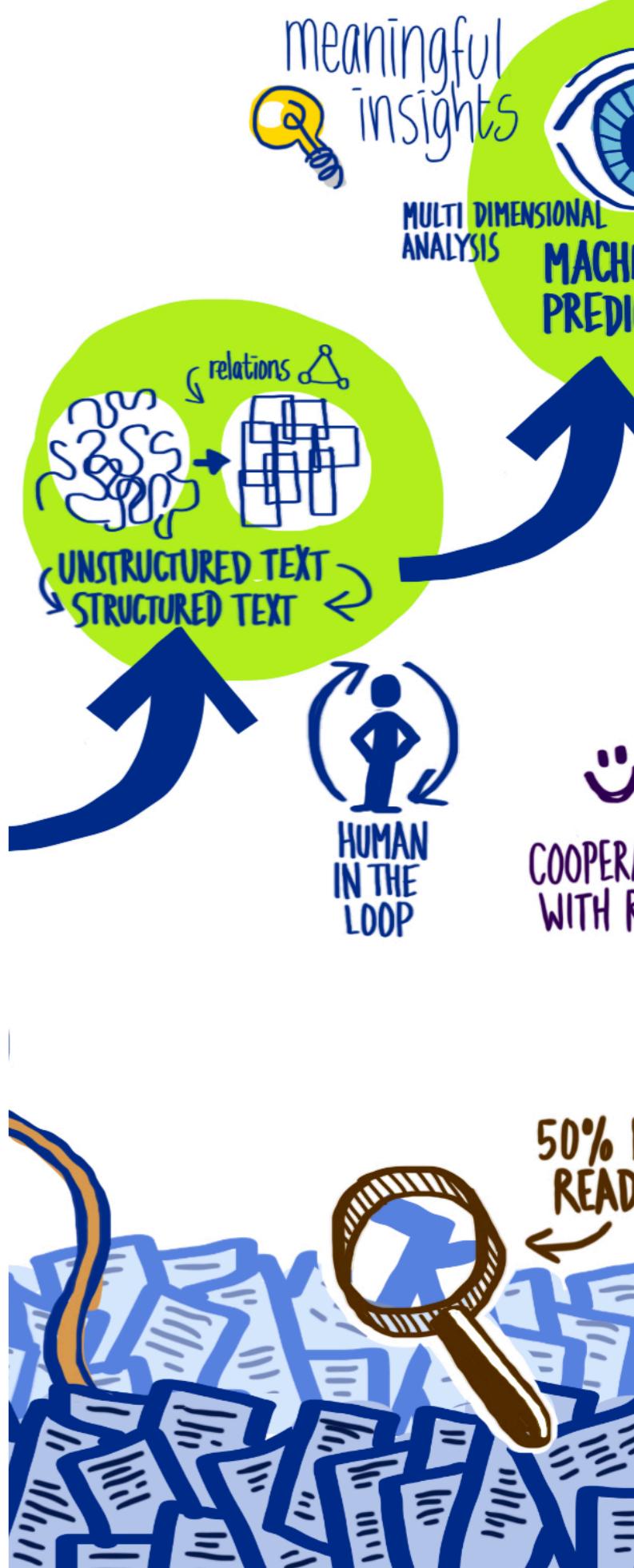
[가정의 달! 든든한 금융]KB손보, The간편한치매간병보험 출시

경증부터 중증까지 폭 넓은 보장

등록 2019-05-18 오전 10:06:05
수정 2019-05-18 오전 10:06:05

가 가

구분	내용
원문	<ul style="list-style-type: none">KB손해보험은 치매에 대해 경증부터 중증까지 폭넓게 보장하는 'The간편한치매간병보험'을 판매 중이다. 이 상품은 경증치매, 중등도치매, 중증치매, 알츠하이머병, 파킨슨병까지 치매와 관련된 질병들을 포괄적으로 보장하는 게 가장 큰 특징이다. ...
형태소 분석	<ul style="list-style-type: none">kb/SL 손해/NNG 보험/NNG 은/JX 치매/NNG 에/JKB 대하/VV 아/EC 경증/NNG 부터/JX 중증/NNG 까지/JX 폭넓/VA 게/EC 보장/NNG 하/XSV 는/ETM '/SO the/SL 간편/XR 하/XSA 르/ETM 치매/NNG 간병/NNG 보험/NNG '/SO 을/JKO 판매/NNG 중/NNB 이/VCP 다/EF ./SF 이/MM 상품/NNG 은/JX 경증/NNG 치매/NNG ,/SP 중등/NNG 도/JX 치매/NNG ,/SP 중증/NNG 치매/NNG ,/SP 알츠하이머병/NNG ,/SP 파킨슨병/NNG 까지/JX 치매/NNG 와/JC 관련/NNG 되/XSV 르/ETM 질병/NNG 들/XSN 을/JKO 포괄/NNG 적/XSN 으로/JKB 보장/NNG 하/XSV 는/ETM 것/NNB 이/JKS 가장/MAG 크/VA 르/ETM 특징/NNG 이/VCP 다/EF ./SF ...



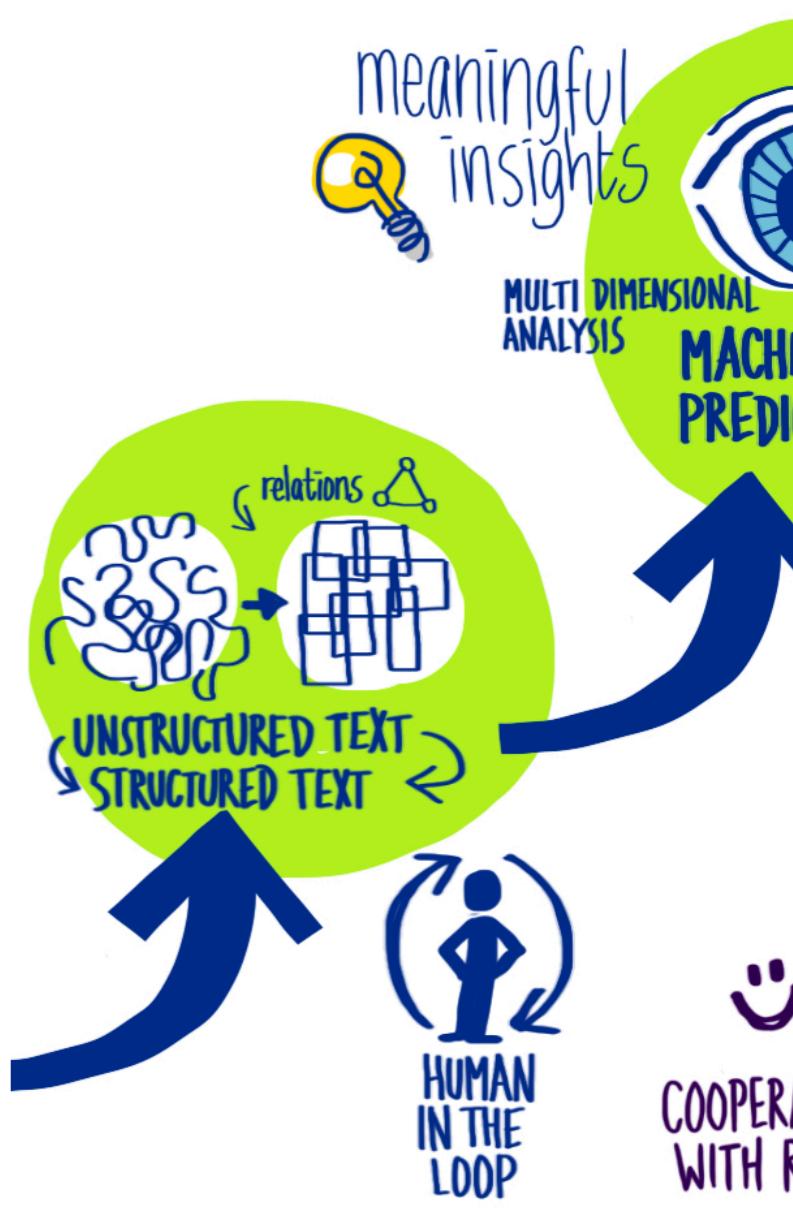
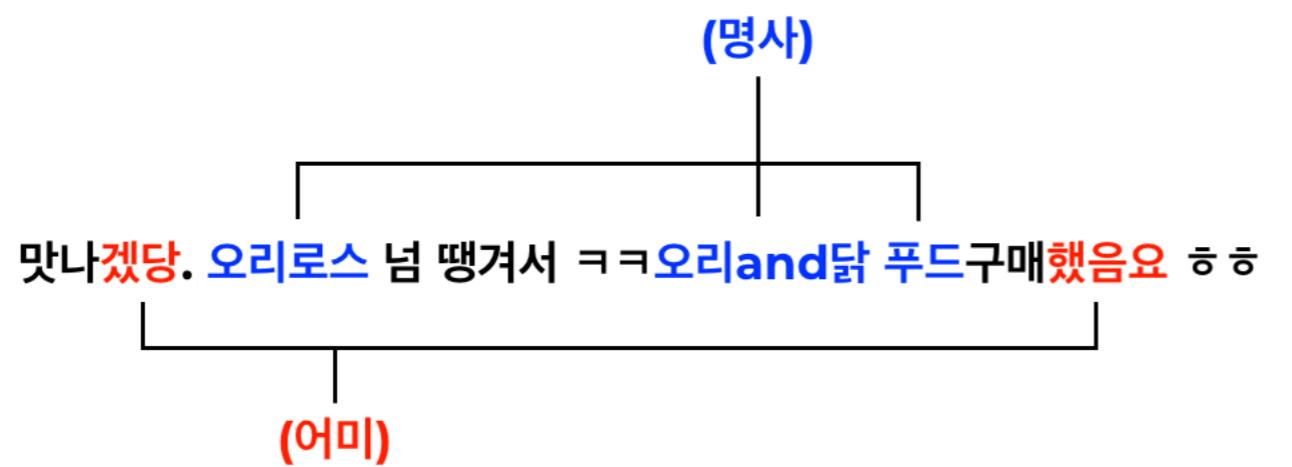
미등록단어 주출

새로운 단어를 인식하기가 어려운 이유

- ▶ 형태소 사전 기반으로 형태소 분석을 수행하는 경우, 미등록단어를 알려진 형태소 단위로 분해함
- ▶ 특히 한국어는 한자어의 조합으로 구성된 단어들이 많기 때문에 작은 의미단위로 분해될 가능성이 큼
- ▶ 좋은 품질의 형태소 분석을 위해서는 새로운 단어들을 사전에 추가하는 과정이 반드시 필요함

신규 단어등록의 자동화 또는 반자동화

- ▶ 사용자 사전을 만들되, 효율적으로 구성하는 방법을 찾아야 텍스트 전처리 시간을 줄이고 전처리 결과의 질을 높힐 수 있음
- ▶ 신규 단어의 대부분은 명사 또는 어미로 이루어지는 경우가 많음
 - 명사 : 새로운 개념을 표현하기 위해 생성됨
 - 어미 : 새로운 말투를 표현하기 위해 생성됨 (동사/형용사에 영향)



미등록단어 추출

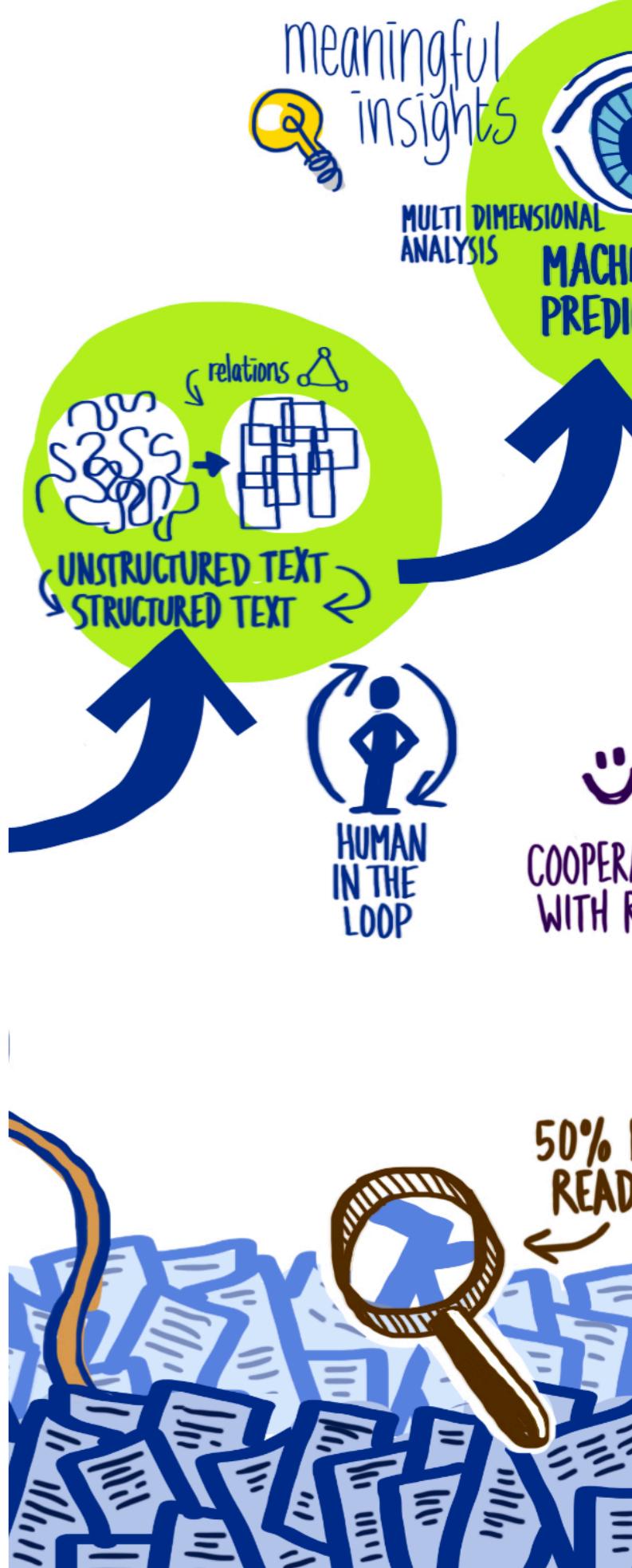
비지도학습과 Char N-gram을 활용한 미등록단어 추출방법

- ▶ Char N-gram : 일반적으로 사용하는 단어 단위의 N-gram과 달리 음절 단위로 묶는 방법
- ▶ 한국어에서 의미를 지니는 단어는 어절의 왼쪽에 존재함
 - 오리고기를 먹다 → (오리고기/NNG + 를/JKO) + (먹/VV + 다/EF)
- ▶ 단어에 대한 정보가 충분하지 않을 경우에 다음 글자가 등장할 확률을 통해 문맥의 모호성을 판별함

Char N-gram	아이	아이오	아이오아	아이오아이	아이오아이는
등장확률 (%)	아니/17.15	아이폰/16.60	아이오아/87.95	아이오아이/100	아이오아이의/31.97
	아이/14.86	아이들/13.37	아이오닉/7.49		아이오아이는/27.21
	아시/8.06	아이디/9.66	아이오와/3.26		아이오아이와/13.61
	아닌/4.74	아이돌/6.77	아이오빈/0.65		아이오아이가/12.24
	아파/4.43	아이뉴/6.77	아이오페/0.33		아이오아이에/9.52
	아직/3.85	아이오/6.33	아이오케/0.33		아이오아이까/1.36

* 2016.10.22. 뉴스기사 기준 측정치

$$\text{cohesion}(\text{"아이오아이"}) = \{ P(\text{아} \rightarrow \text{아이}) * P(\text{아이} \rightarrow \text{아이오}) \\ * P(\text{아이오} \rightarrow \text{아이오아}) \\ * P(\text{아이오아}) \rightarrow (\text{아이오아이}) \}^{1/(5-1)}$$

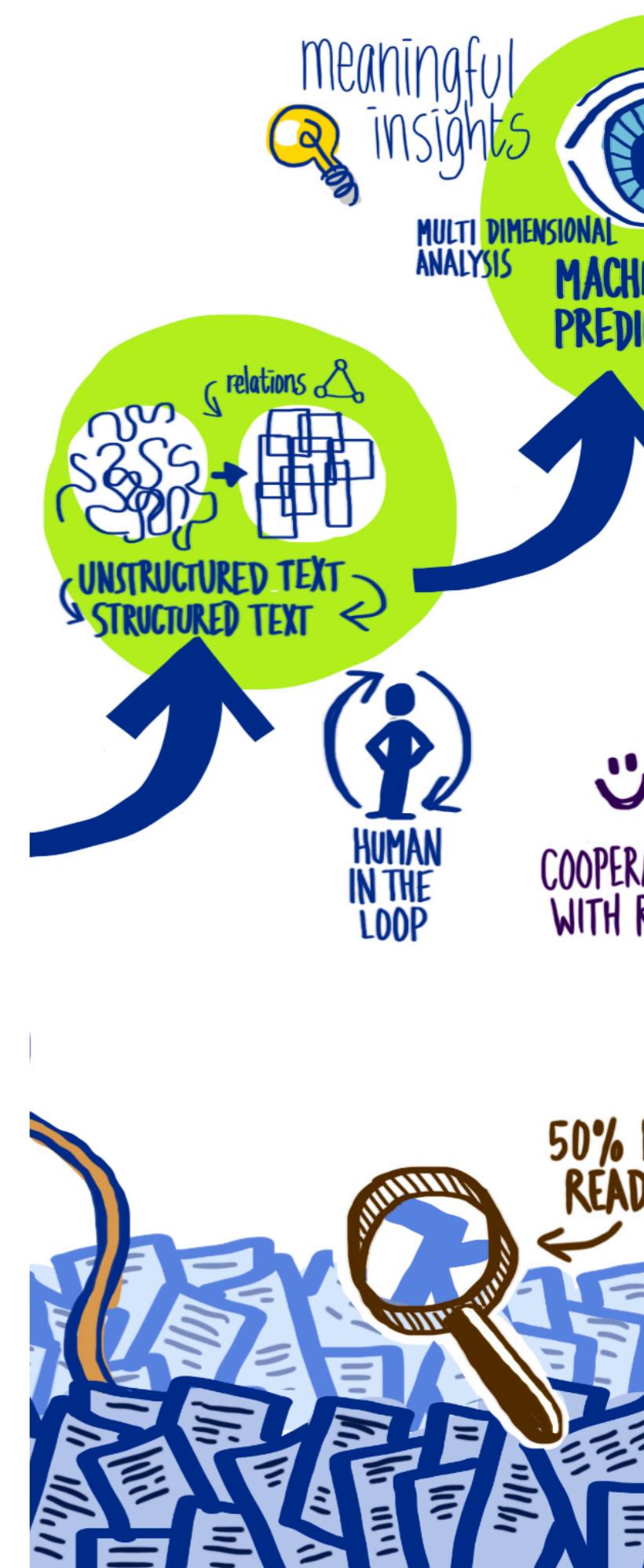
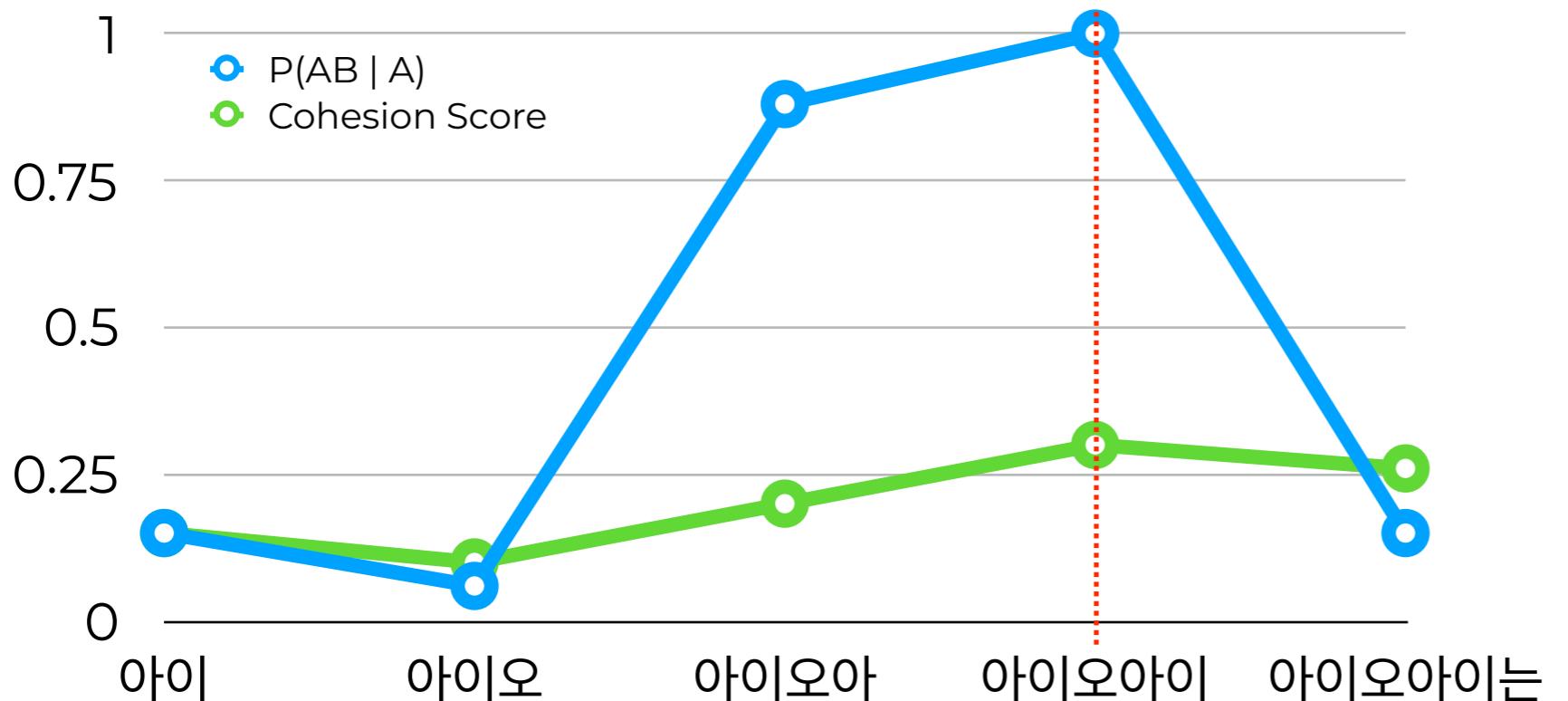


미등록단어 추출

하루치 뉴스에서 계산한 Cohesion Score

- “아이오아이”로 검색된 하루치 뉴스로 부터 학습된 결과

Char N-gram	Frequency	$P(AB A)$	Cohesion Score
아이	4,910	0.15	0.15
아이오	307	0.06	0.10
아이오아	270	0.88	0.20
아이오아이	270	1.00	0.30
아이오아이는	40	0.15	0.26



E.O.D