

TEXT MINING for PRACTICE

Python을 활용한 비정형 데이터 분석 - WEEK 04

전병진 FINGEREDMAN (fingeredman@gmail.com)

Part 3.

텍스트 데이터 전처리

텍스트 데이터 전처리

텍스트 전처리에 허투루하면
허투론 분석결과가 나온다

텍스트 마이닝 용어

① 문서 (Document)

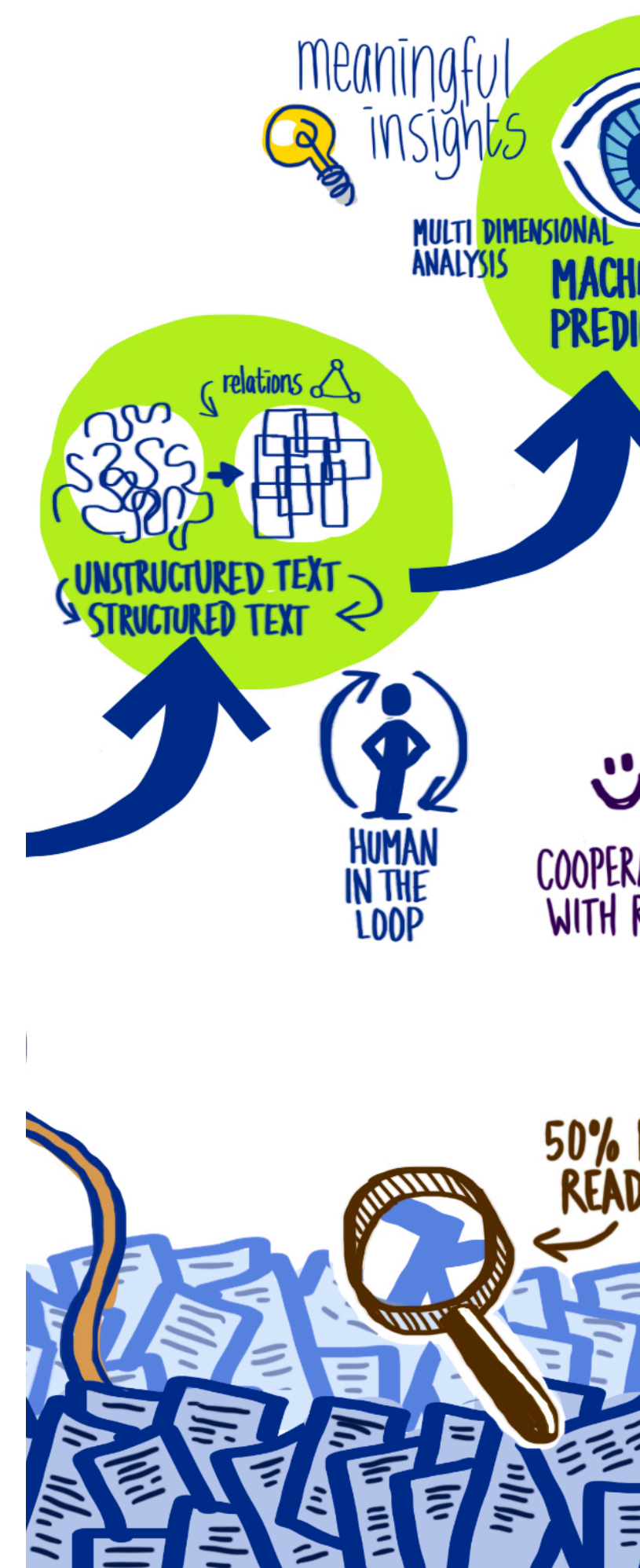
- ▶ 한 덩어리의 텍스트로서 짧은 문장에서부터 긴 문서 까지를 모두 포함하는 의미
- ▶ 문서의 집합을 말뭉치 (corpus)라고 함
- ▶ 문서의 레벨에 따라 말뭉치의 레벨이 바뀔 수 있음(문장, 문단, 페이지, 댓글 등)

② 어휘사전 (Lexicon)

- ▶ 어휘 (lexical)의 집합 또는 어휘에 대한 정의 혹은 설명을 가진 사전
- ▶ 특징 별 어휘사전이 나뉘어서 존재하기도 함(인물사전, 영어사전, 건물사전 등)

③ 불용어 (Stop-word)

- ▶ 텍스트 분석에 있어서, 또는 분석결과에 출현하더라도 아무런 의미가 없는 단어의 집합
- ▶ 정보전달 보다는 주로 기능적인 역할을 하는 단어에 해당함
 - 일반적으로 한국어는 외자 단어, 영어는 알파벳 두 자 단어를 불용어 취급
 - 한국어 예 : 그거, 여기, 이제, 은, 는, 이, ...
 - 영어 예 : a, an, the, of, the, ...
- ▶ 빈출어 (Common-word)
 - 너무 많이 출현하여 분석 결과에서 의미 또는 중요도가 떨어지는 단어의 집합
 - 예 : 기사, 기자, 제목, 사진, 네이버, 검색, 보다, 연기, 평점, 공감, 비공감



텍스트 마이닝 용어

④ 형태소 (Morpheme)

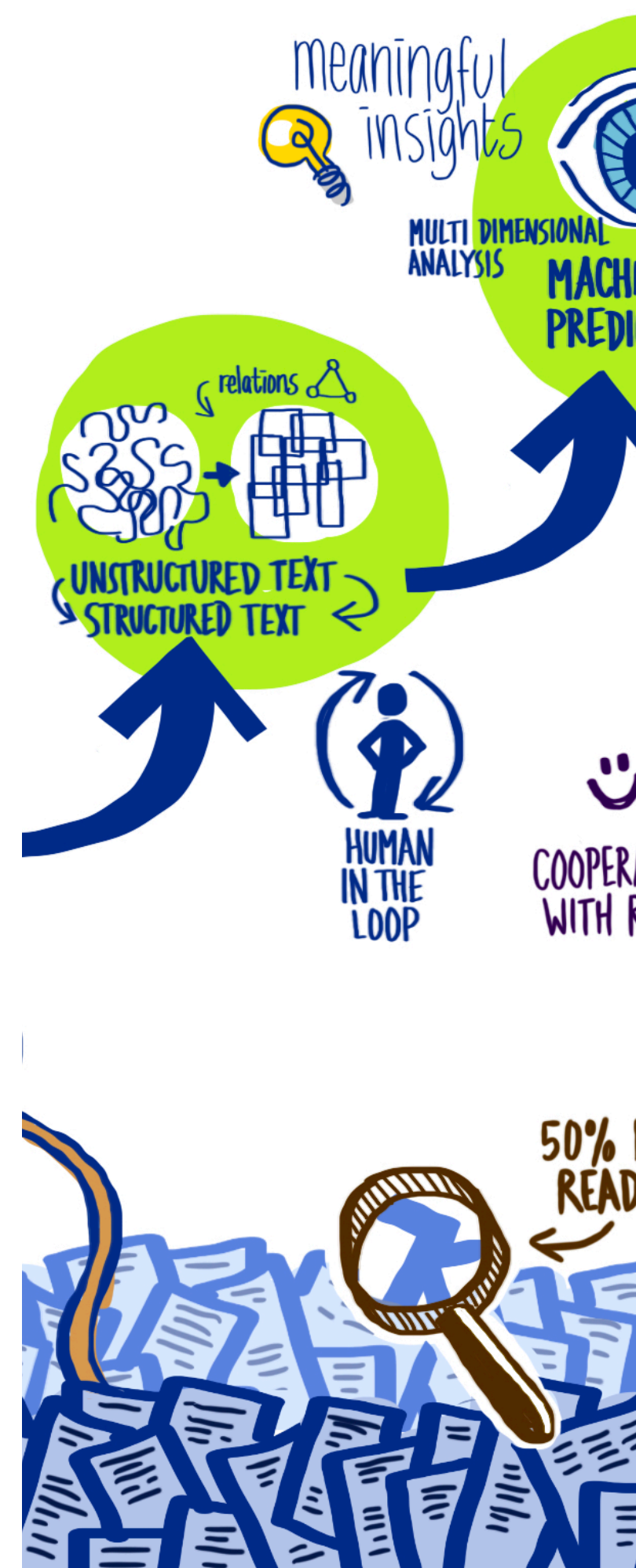
- ▶ 뜻을 가진 가장 작은 말의 단위
- ▶ 동사, 명사, 조사, 문장부호 등 보통 품사 (Part of Speech, POS) 단위를 의미함

⑤ 단어 주머니 (Bag of Words, BoW)

- ▶ 문서에 함께 사용된 단어의 집합
- ▶ 중복된 단어는 하나로 취급하며, 순서에 의미를 고려하지 않음
 - 예 : “아버지가 방에 들어가신다.” → [“아버지”, “방”, “들어가다”]

⑥ 토큰화 (Tokenization)

- ▶ 토큰 (token)
 - 유용한 의미적 단위로 함께 모여지는 일련의 문자열
 - 구분 기호 사이의 글자 시퀀스
- ▶ 문헌 단위의 문자열이 주어졌을 때 토큰들로 문자열을 분리하는 작업
- ▶ 구두점 등 불필요한 글자들을 제외하기도 함
- ▶ 영어는 언어학적 특성상 단어에 조사가 붙지 않아 한글보다 토큰화가 쉬움



N 그램 모형 (N-gram)

Uni-gram

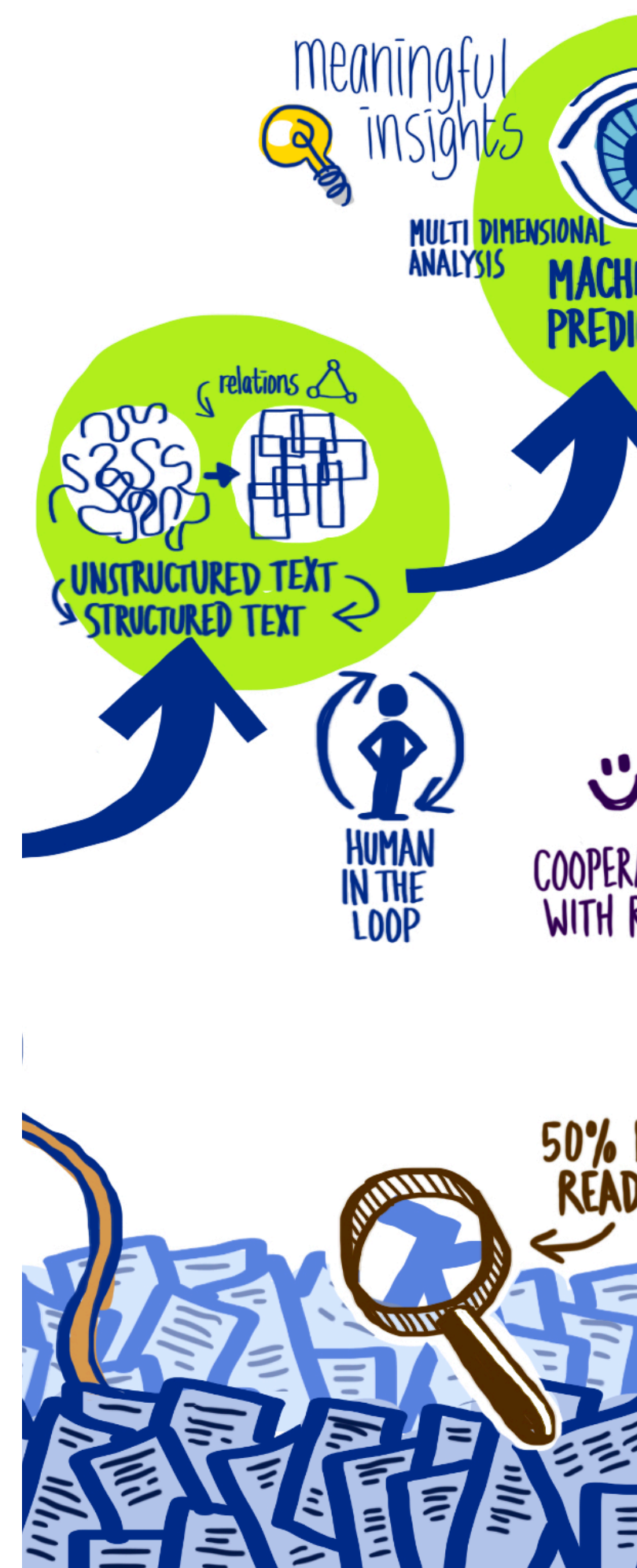
- ▶ 독립된 하나의 단어 또는 형태소 단위를 의미함

구분	내용
원문	· 여러분 안녕하세요. 재미있는 텍스트 마이닝 수업입니다.
형태소 분석	· 여러분 + 안녕 + 하세요 + . · 재미있 + 는 + 텍스트 + 마이닝 + 수업 + 입니다 + .
특징	· 총 11개의 단어 (uni-gram)로 이루어진 문장 · 특수문자를 제외하고 9개의 단어로 구성됨

Bi-gram

- ▶ 두 개의 단어 또는 형태소 단위 하나의 단어로 취급하는 단위를 의미함
- ▶ Uni-gram에 비해 하나의 단위에 더 많은 정보를 내포할 수 있음

구분	내용
원문	· 여러분 안녕하세요. 재미있는 텍스트 마이닝 수업입니다.
형태소 분석	· (여러분 + 안녕), (안녕 + 하세요) · (재미있 + 는), (는 + 텍스트), (텍스트 + 마이닝), (마이닝 + 수업), (수업 + 입니다)
특징	· 총 6개의 단어 (bi-gram)로 이루어진 문장



N 그램 모형 (N-gram)

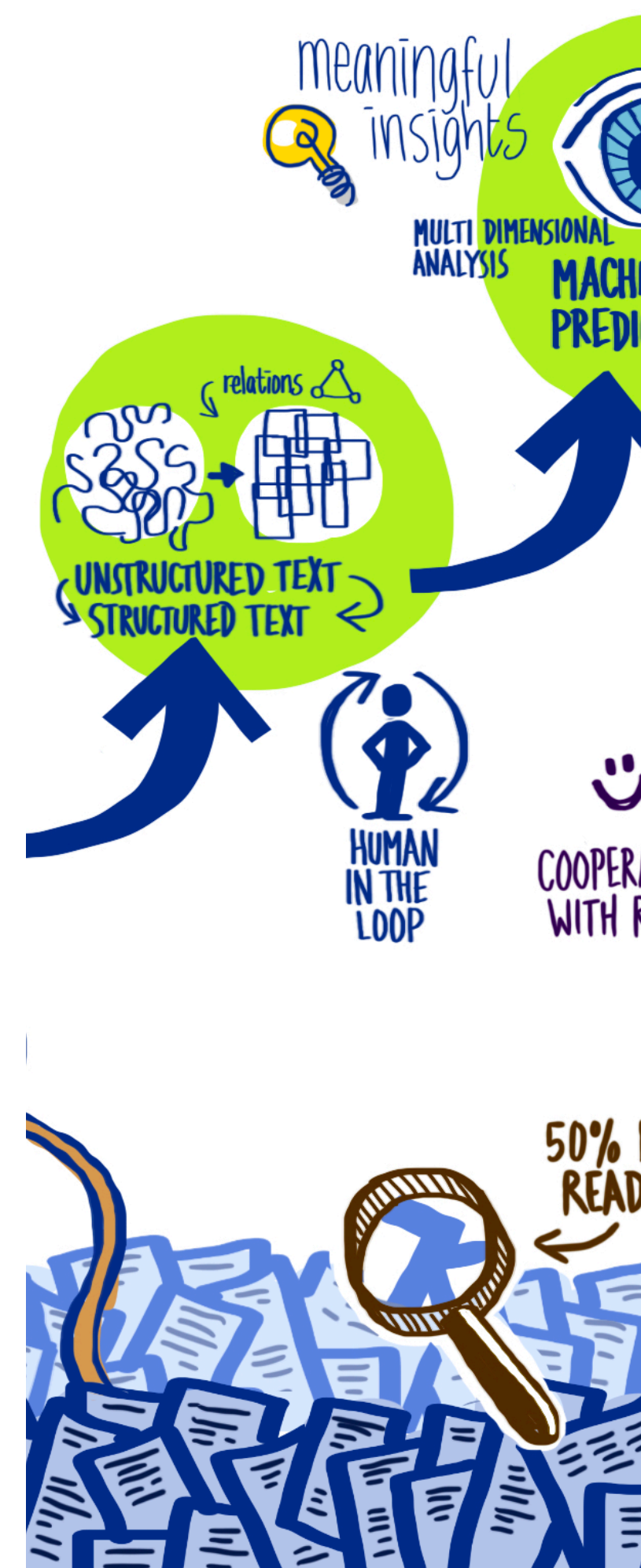
N-gram

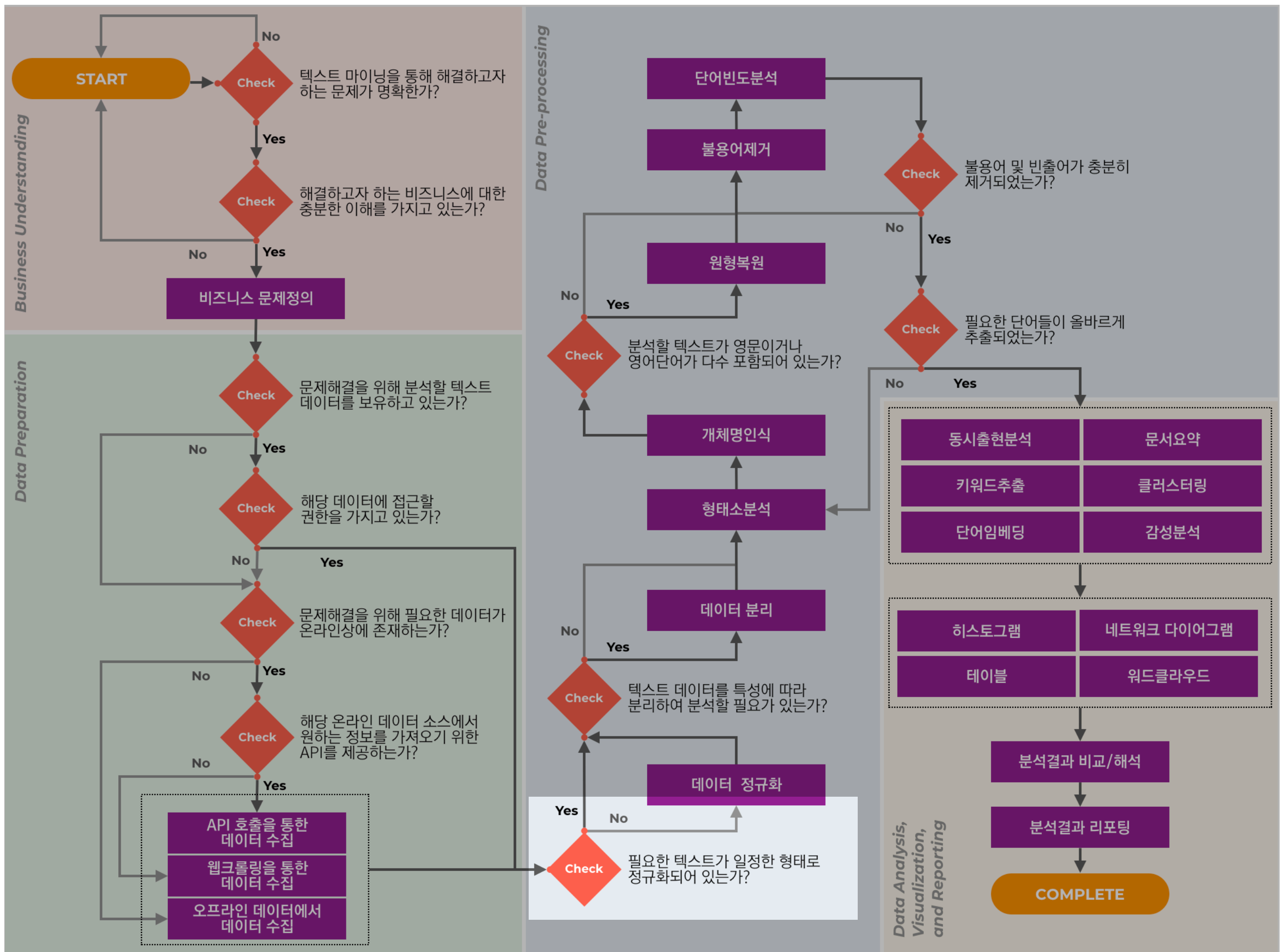
- ▶ N 개의 단어 또는 형태소 단위를 하나의 단어로 취급하는 단위를 의미함
- ▶ N > 2 일때 N-gram 이라는 표현을 사용하고 그 외는 Uni-gram, Bi-gram을 사용함
- ▶ N이 증가할 수록 텍스트 처리에 필요한 연산량이 기하급수적으로 증가하기 때문에 일반적으로 1~3 범위의 N을 지정하여 사용함(Bi-gram으로도 충분)
- ▶ 의미없는 불용어가 포함된 N-gram을 방지하기 위해 (명사+명사), (형용사+명사) 등의 조합을 가진 N-gram 단어만 사용할 필요가 있음

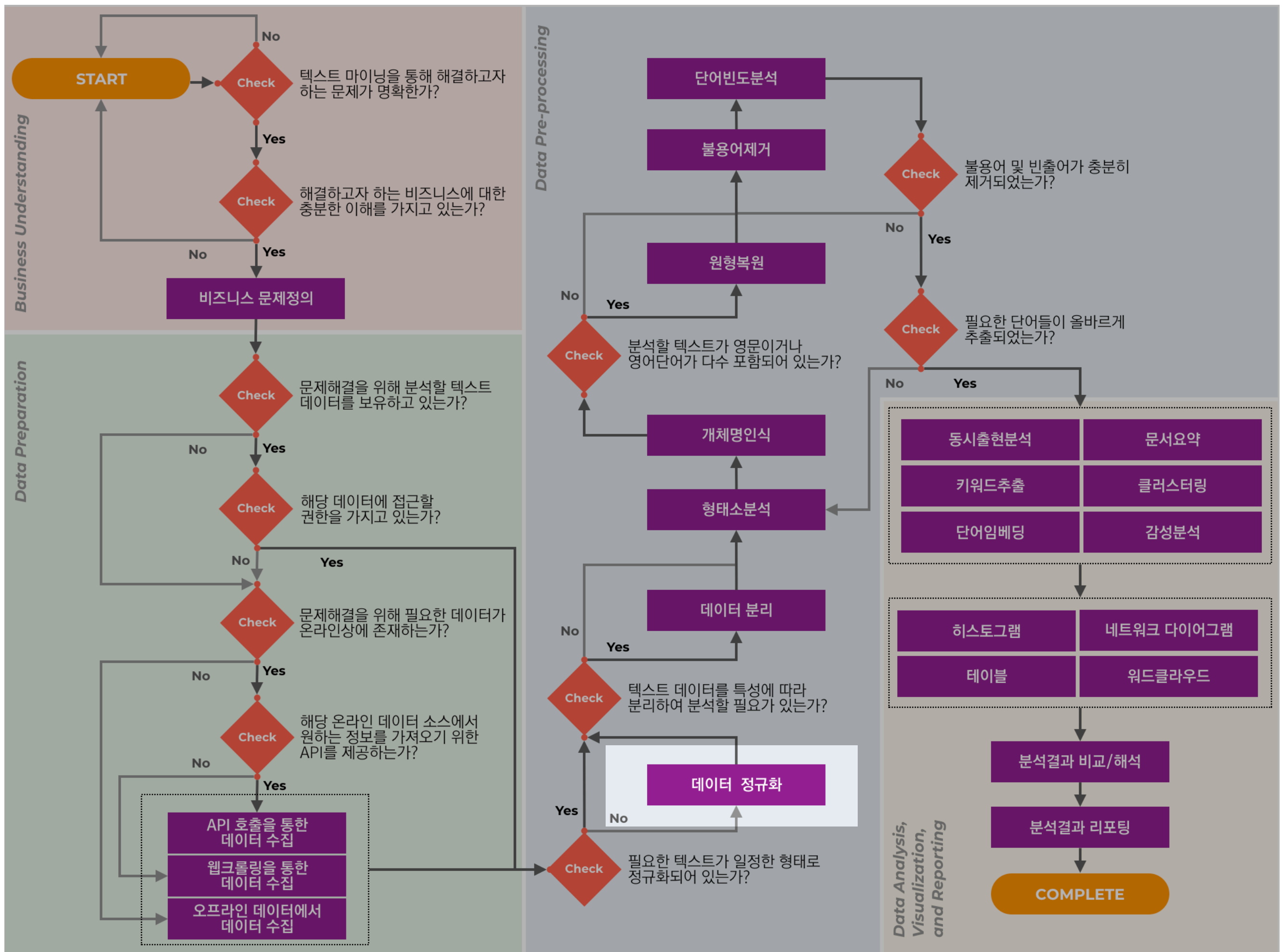
구분	내용
원문	· 여러분 안녕하세요. 재미있는 텍스트 마이닝 수업입니다.
형태소 분석	· (여러분 + 안녕), (안녕 + 하세요), (재미있 + 는), (는 + 텍스트), (텍스트 + 마이닝), (마이닝 + 수업), (수업 + 입니다)

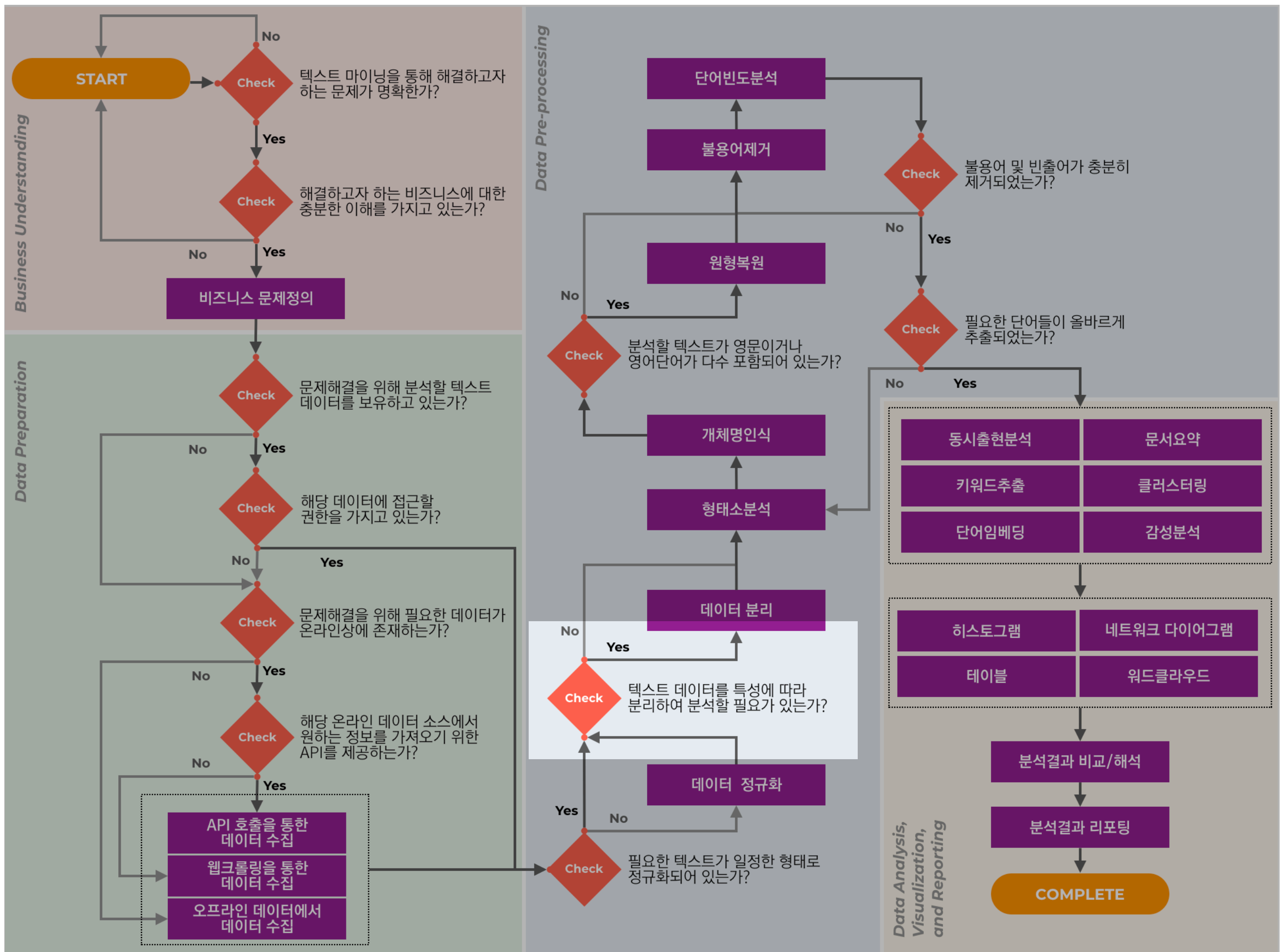
- ▶ 한국어의 문법구조 특성상 (형용사+조사+명사) 조합에서 조사를 생략한 채로 (형용사+명사) 조합을 사용하기도 함

구분	내용
원문	· 여러분 안녕하세요. 재미있는 텍스트 마이닝 수업입니다.
형태소 분석	· (여러분 + 안녕), (안녕 + 하세요), (재미있 + [는] + 텍스트), (텍스트 + 마이닝), (마이닝 + 수업), (수업 + 입니다)









문서 정규화

문서를 하나의 통일된 형식으로 정규화하는 과정

- ▶ 데이터를 수집하면서 정규화를 같이 진행하면 매우 효율적으로 할 수 있음
- ▶ 정규화 유형 : Table, XML, JSON

[Table]

employee	
name	age
James Kirk	40
Jean-Luc Picard	45
Wesley Crusher	27

```
1 name → age
2 James Kirk → 40
3 Jean-Luc Picard → 45
4 Wesley Crusher → 27
```

[XML]

```
<empinfo>
  <employees>
    <employee>
      <name>James Kirk</name>
      <age>40</age>
    </employee>
    <employee>
      <name>Jean-Luc Picard</name>
      <age>45</age>
    </employee>
    <employee>
      <name>Wesley Crusher</name>
      <age>27</age>
    </employee>
  </employees>
</empinfo>
```

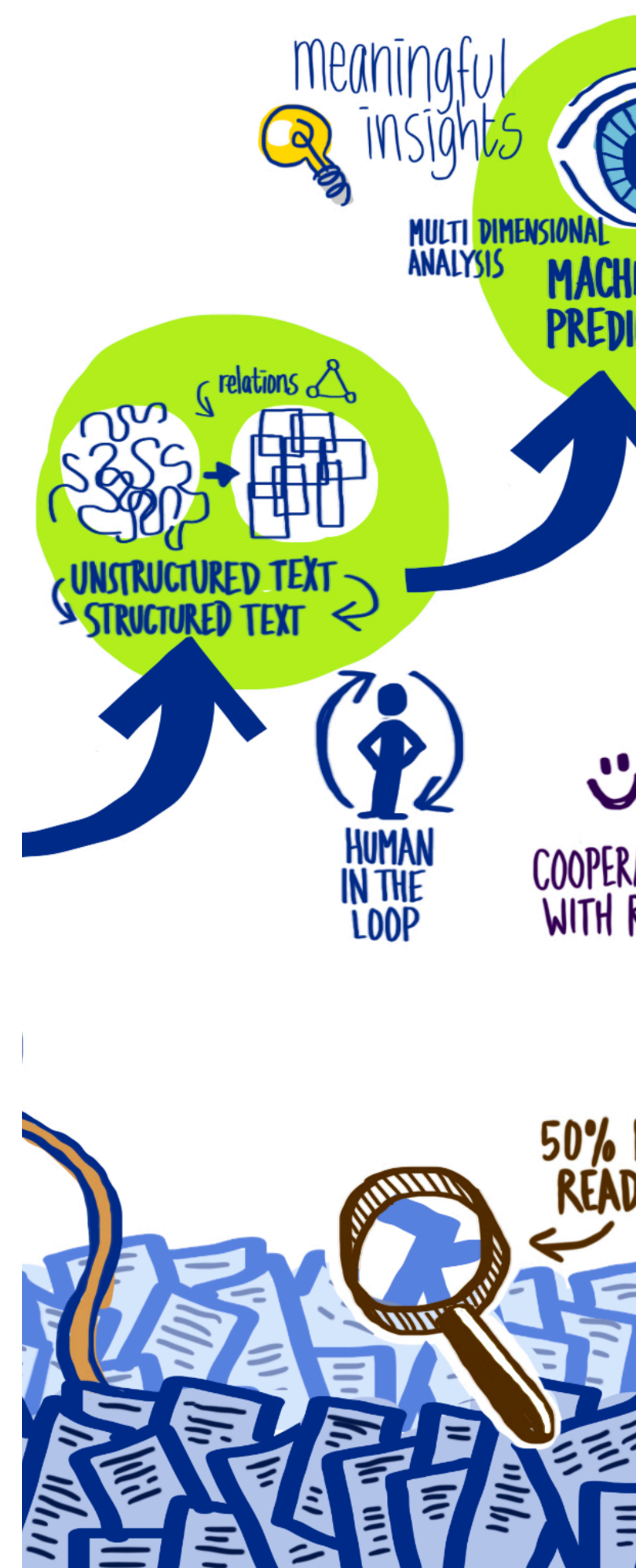
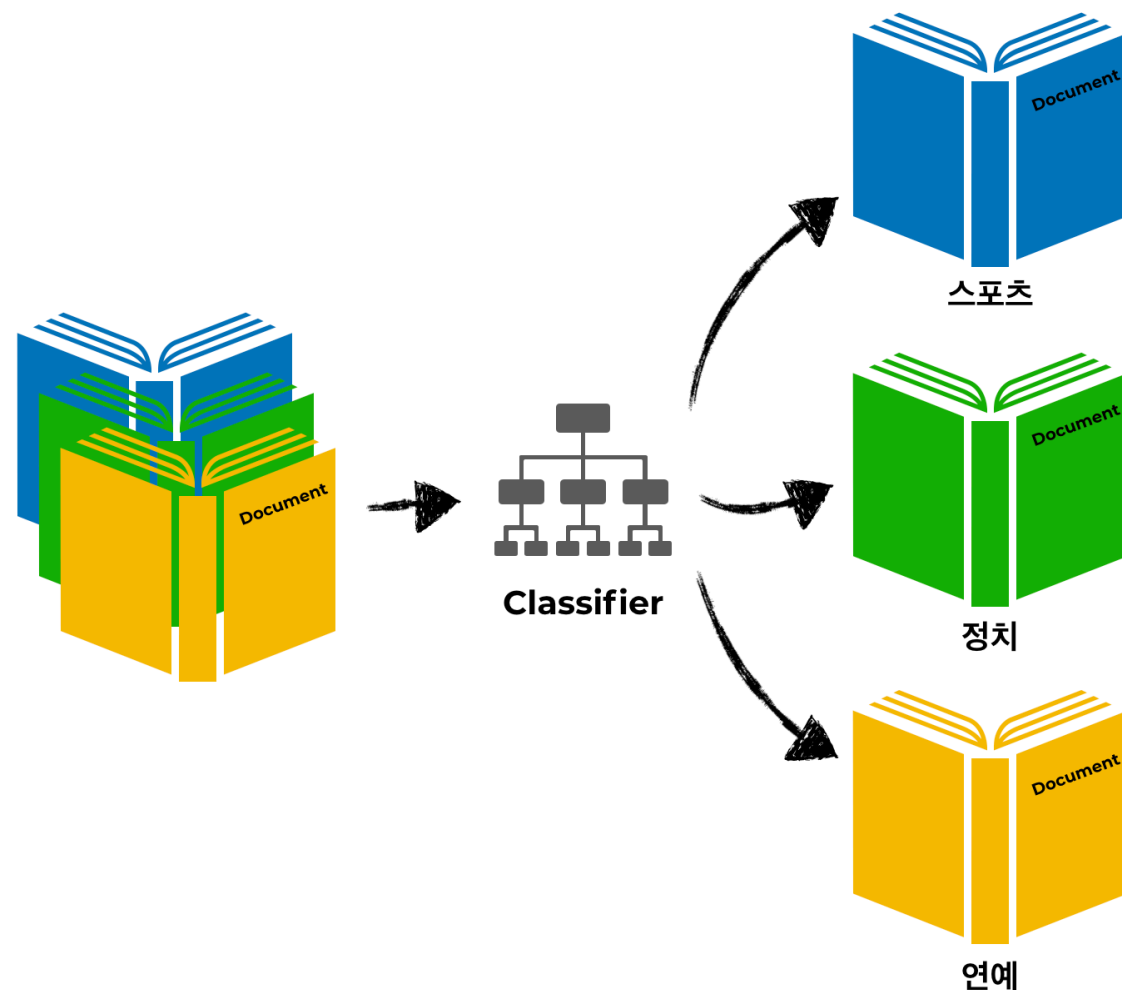
[JSON]

```
{ "empinfo" :
  {
    "employees" : [
      {
        "name" : "James Kirk",
        "age" : 40,
      },
      {
        "name" : "Jean-Luc Picard",
        "age" : 45,
      },
      {
        "name" : "Wesley Crusher",
        "age" : 27,
      }
    ]
  }
}
```

문서 분리 (Seperation)

문서를 특정 기준에 의해 분리하거나 통합하는 과정

- ▶ 문서의 작성자, 날짜, 길이, 구분자, 감성 스코어, 랜덤 등을 기준으로 할 수 있음
- ▶ 모델 생성 : 모델 생성을 위한 훈련과 테스트용 데이터를 구분
- ▶ 분석단위 : 데이터에 포함된 특정 값을 기준으로 분리
- ▶ 문장단위 : 텍스트는 문장단위로 구분(언어를 이해하는 최소단위)



원형복원 & 불용어 처리

원형복원

- ▶ 변형된 단어의 원형을 복원하는 과정
- ▶ Stemming : 규칙 기반으로 단어의 변형된 형태를 제거
- ▶ Lemmatizing : 사전 기반으로 품사에 맞는 단어의 원형으로 변환

불용어 처리

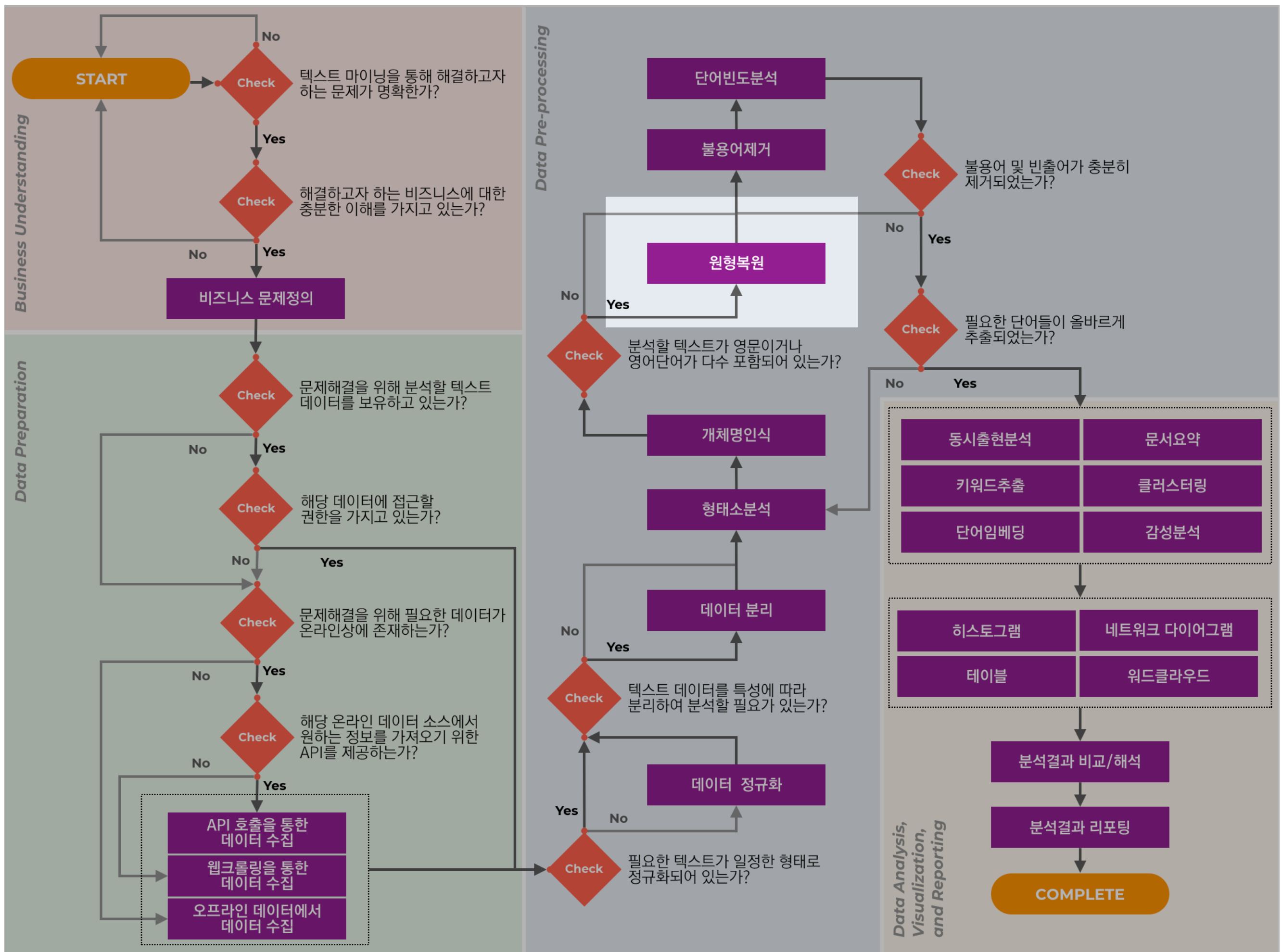
- ▶ 분석에 불필요한 단어나 방해가 되는 단어를 제거하는 과정
- ▶ 주로 불용어 (Stop-word) 또는 최빈어 (Common-word)가 제거됨
- ▶ 하다, 이다, is, 기자, 뉴스

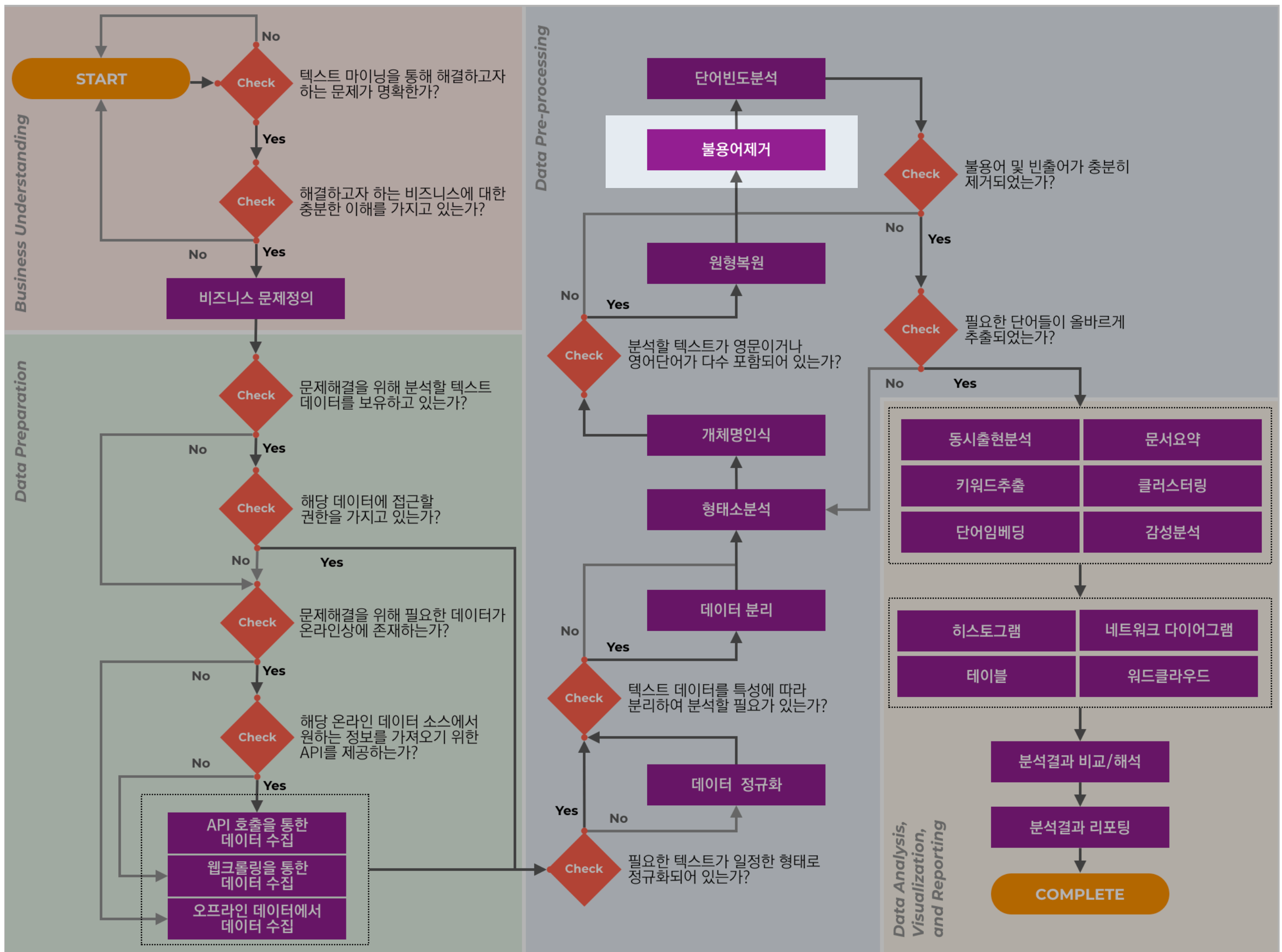
[Stemming & Lemmatizing 비교]

Word	Stemming	Lemmatizing
cooking(v)	cook	cook
cooking(n)	cook	cooking
cookbooks	cookbook	cookbook
believes	believ	believe
using	us	use

Note. 복원 결과는 Stemmer와 Lemmatizer의 종류에 따라서 다를 수 있음

과정	결과
개체명 인식	텍스트 마이닝/NNG/IT + 은/JX + 길/VA + 고/ECE + 지겹/VA + L/ETD + 작업/NNG + 이/VCP + 버니다/EFN + .SF Text mining/NN/IT + is/VBZ + difficult/JJ + but/CC + very/RB + valuable/JJ + ./.
원형 복원	텍스트 마이닝/NNG/IT + 은/JX + 길다/VA + 고/ECE + 지겨운/VA + 작업/NNG + 이/VCP + 버니다/EFN + .SF Text mining/NN/IT + be/VBZ + difficult/JJ + but/CC + very/RB + valuable/JJ + ./.
불용어 제거	텍스트 마이닝/NNG/IT + 길다/VA + 지겨운/VA + 작업/NNG Text mining/NN/IT + difficult/JJ + valuable/JJ





E.O.D