

TEXT MINING for PRACTICE

Python을 활용한 비정형 데이터 분석 - WEEK 06

전병진 FINGEREDMAN (fingeredman@gmail.com)

Part 5.

단어빈도분석 & TF-IDF

단어주머니 (Bag of Words)

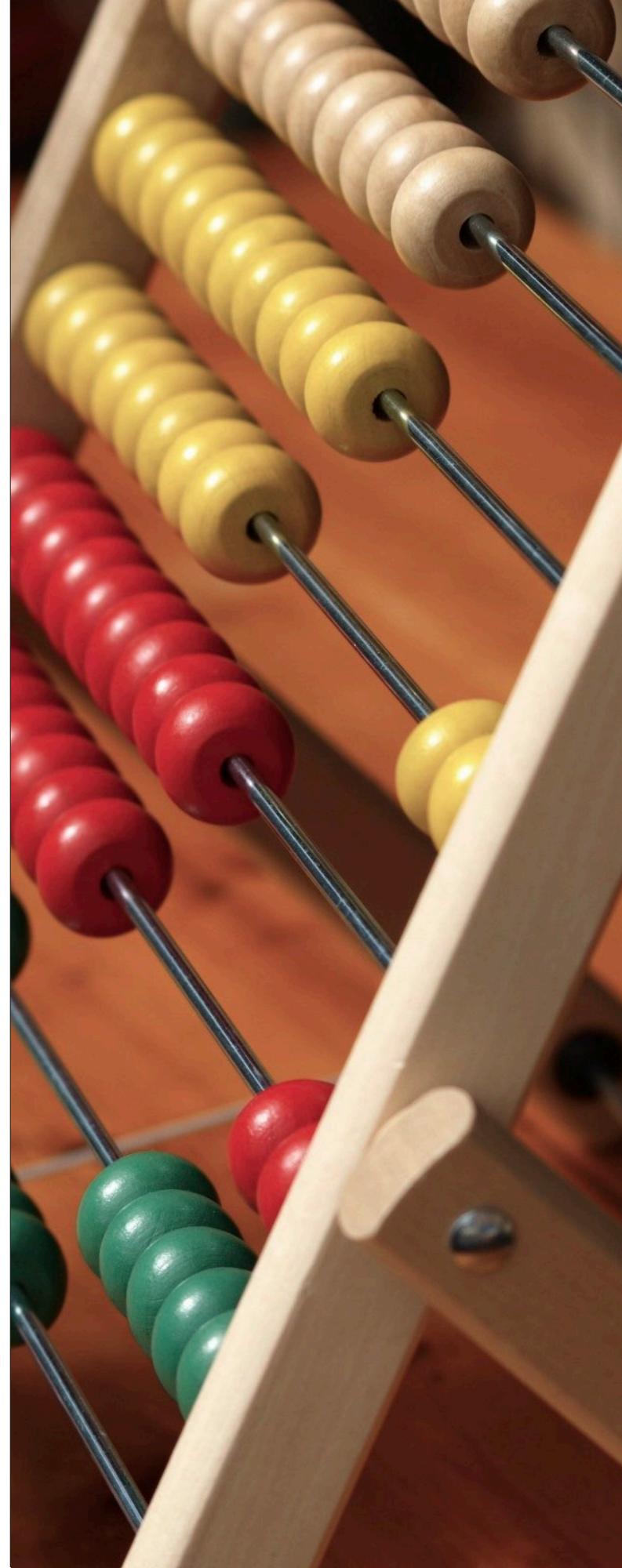
문서를 표현하는 방법 (Document Representation)

- ▶ 비정형 데이터인 문서(텍스트 데이터)를 정형 데이터로 구조화 하여 표현하는 방법
- ▶ 문서를 벡터 (vector) 또는 매트릭스 (matrix) 공간으로 표현하여 벡터기반의 머신러닝 알고리즘 등에 활용할 수 있음

단어주머니

- ▶ 비정형 데이터인 텍스트 데이터(문서)를 정형화된 데이터로 변경하는 가장 대표적인 방법

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



단어주머니 (Bag of Word)

문서를 단어주머니로 표현하는 방법

▶ 문장 단위의 단어주머니

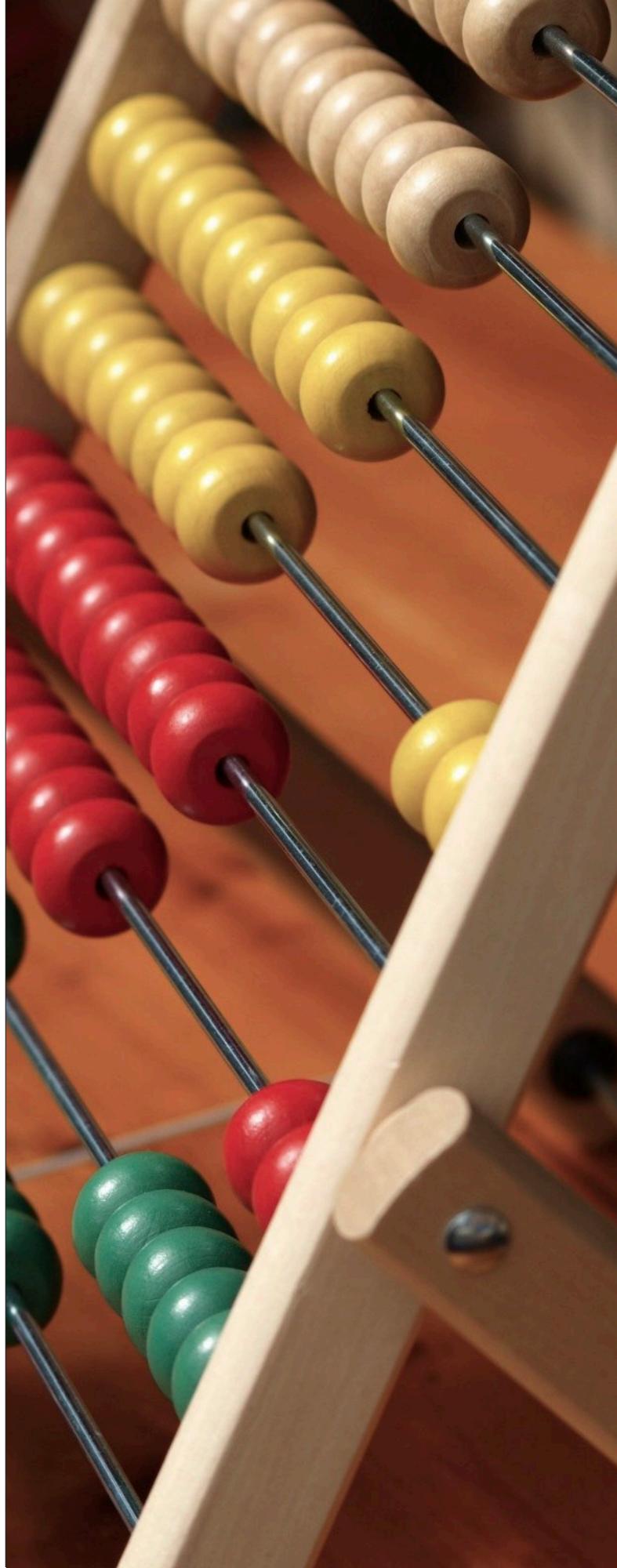
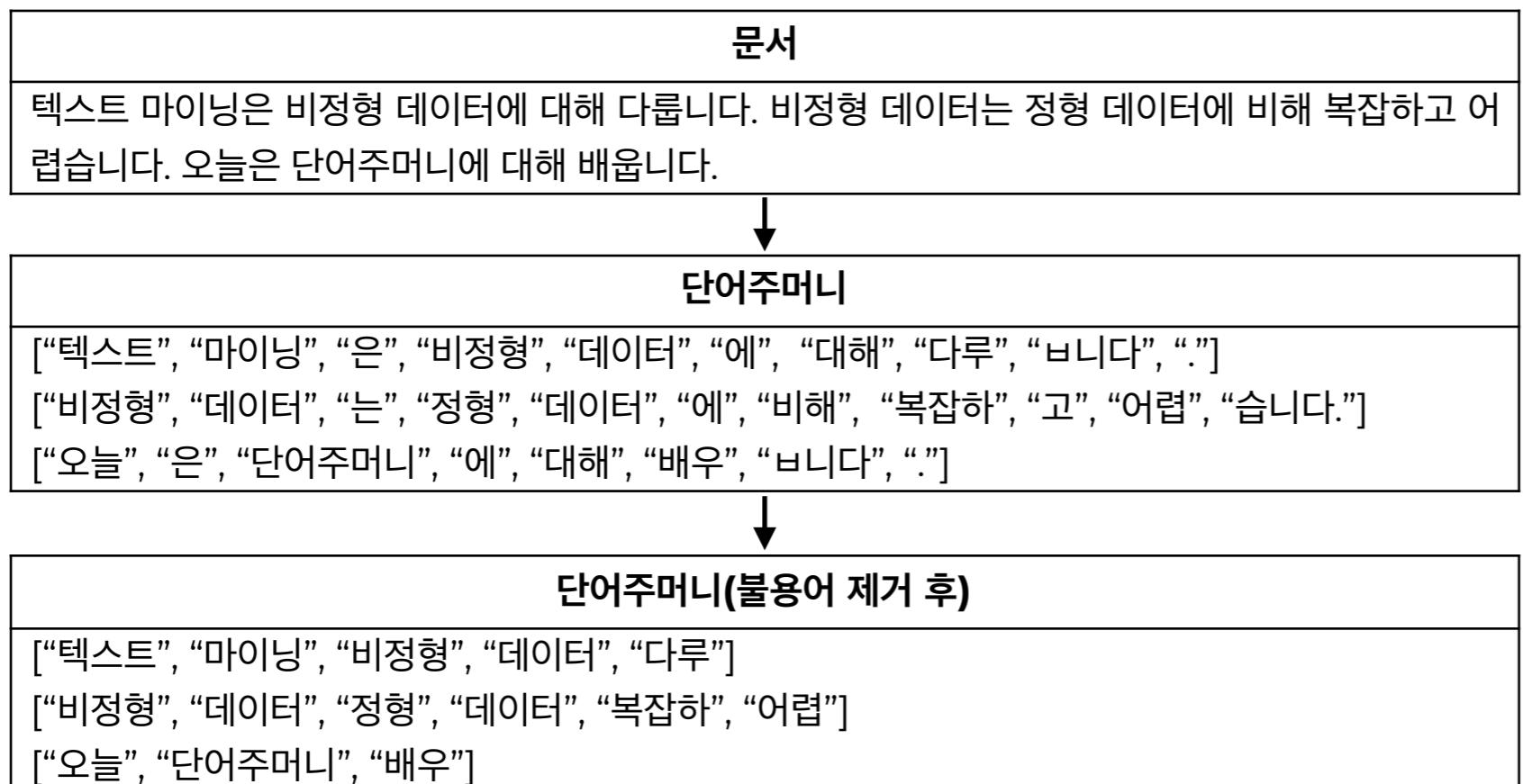
- 나는 밥을 먹는다.

→ [“나”, “는”, “밥”, “을”, “먹”, “는다”, “.”] → [“나”, “밥”, “먹다”]

- 무서운 교수님과 착한 학생들

→ [“무서운”, “교수님”, “과”, “착한”, “학생”, “들”] → [“무서운”, “교수님”, “착한”, “학생”]

▶ 문서 단위의 단어주머니



단어주머니 (Bag of Words)

문서를 단어주머니로 표현하는 방법

- 문서 x 단어 매트릭스 표현

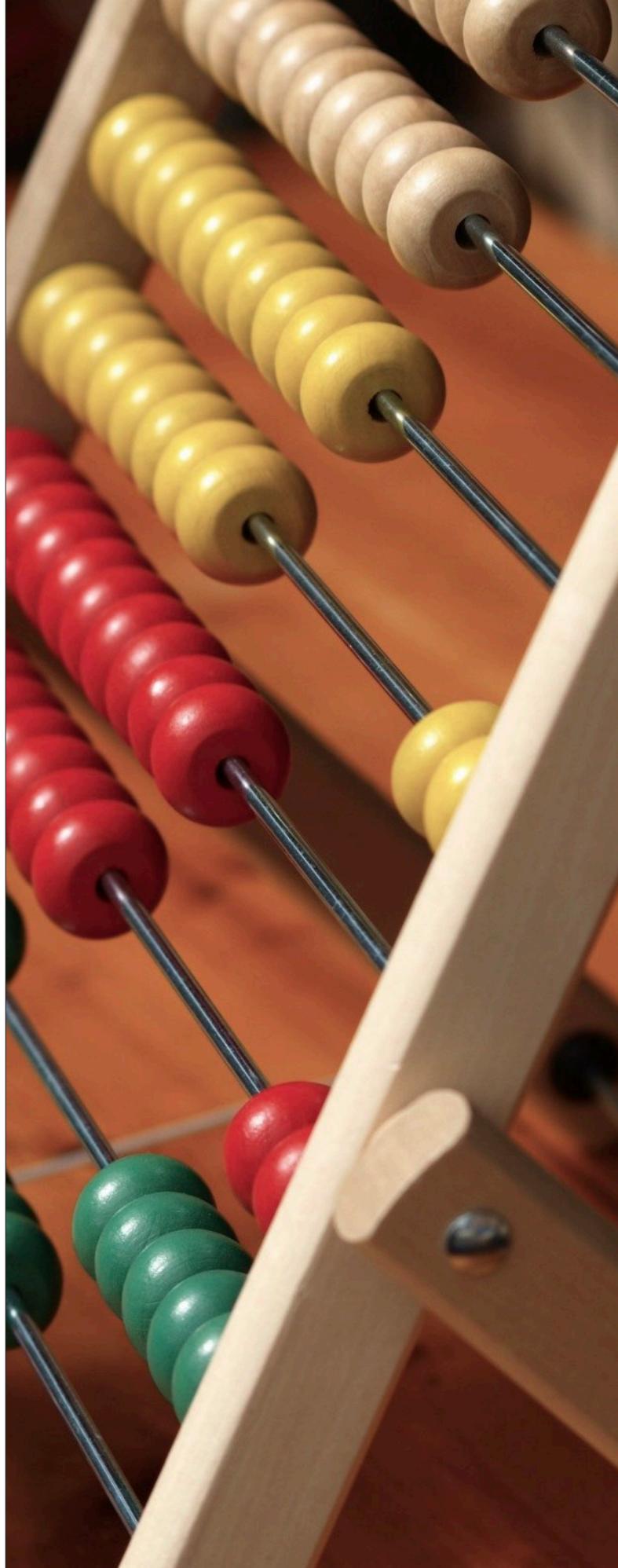
단어주머니(불용어 제거 후)				
[“텍스트”, “마이닝”, “비정형”, “데이터”, “다루”]				
[“비정형”, “데이터”, “정형”, “데이터”, “복잡하”, “어렵”]				
[“오늘”, “단어주머니”, “배우”]				



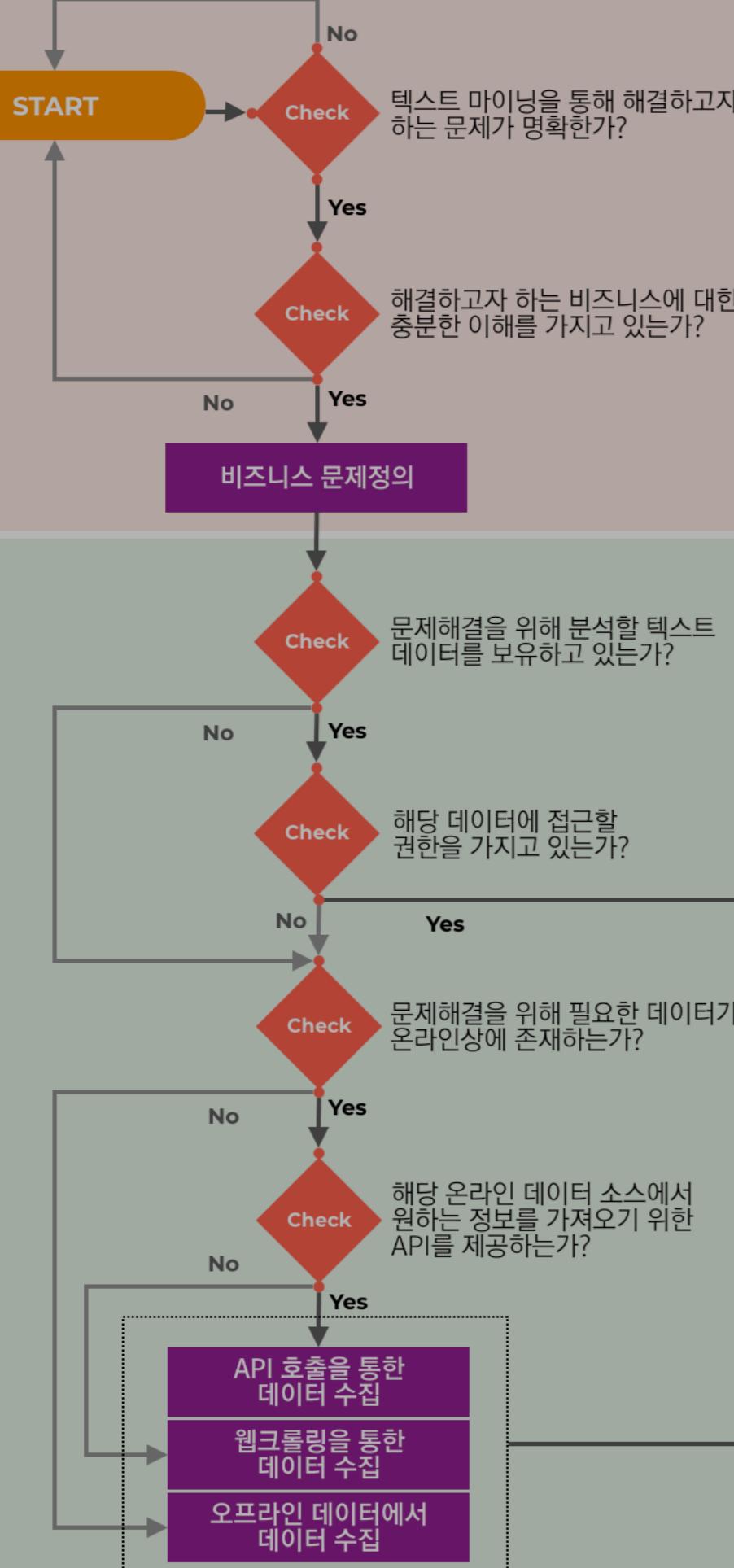
단어	문서 1	문서 2	문서 3	단어빈도
텍스트	1	0	0	1
마이닝	1	0	0	1
비정형	1	1	0	2
데이터	1	2	0	3
다루다	1	0	0	1
정형	0	1	0	1
복잡하다	0	1	0	1
어렵다	0	1	0	1
오늘	0	0	1	1
단어주머니	0	0	1	1
배우다	0	0	1	1



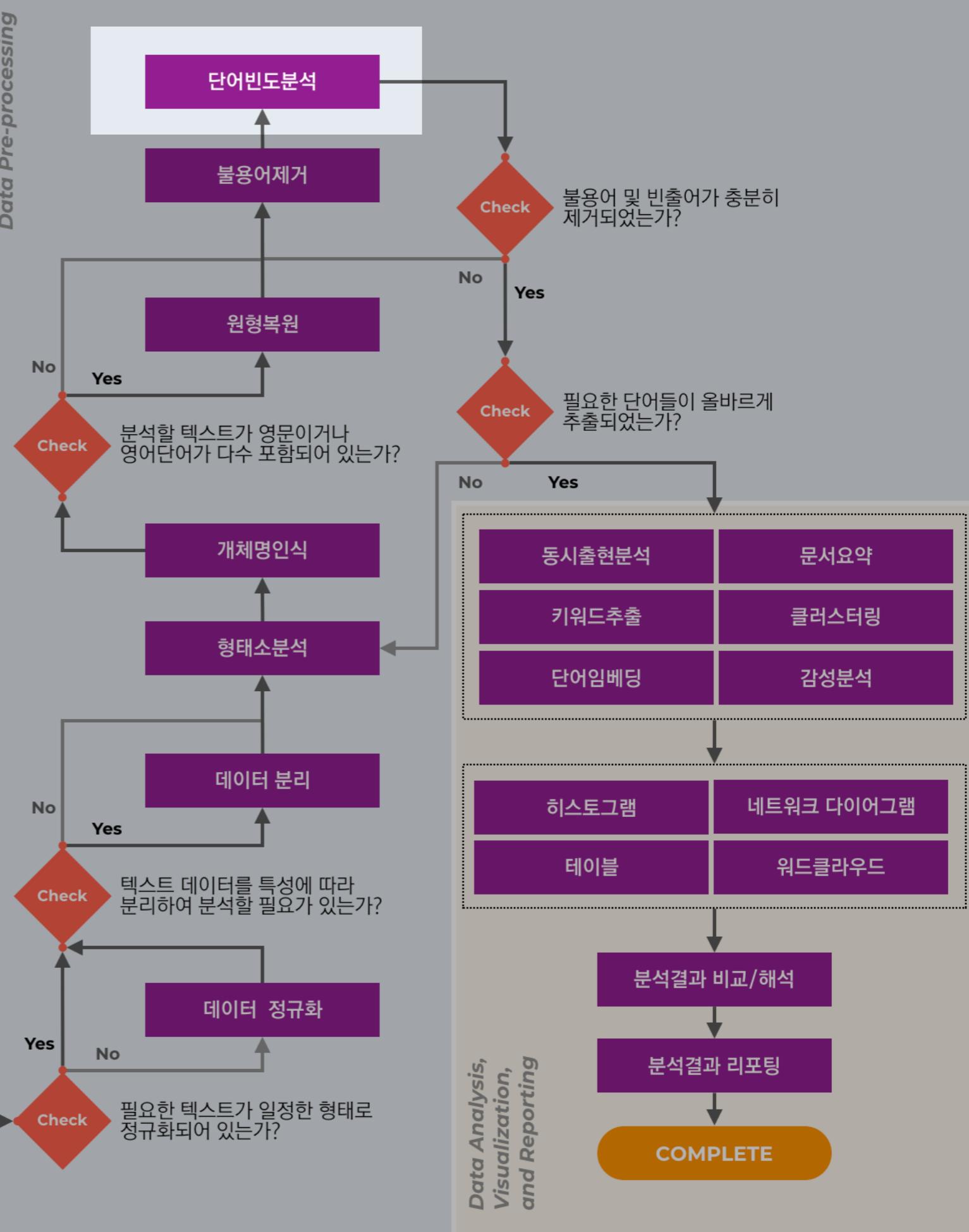
“문서 2”的 벡터 = [0, 0, 1, 2, 0, 1, 1, 1, 0, 0]



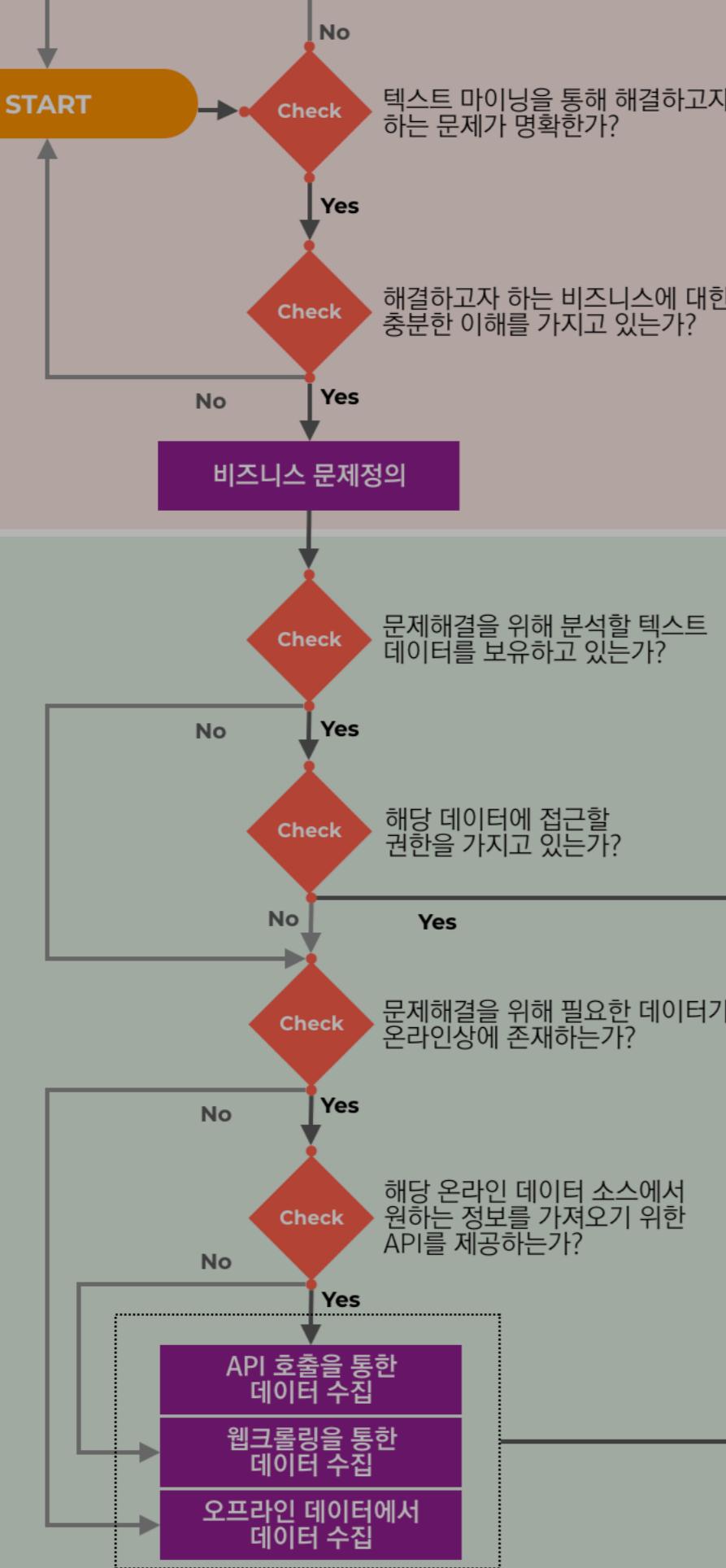
Business Understanding



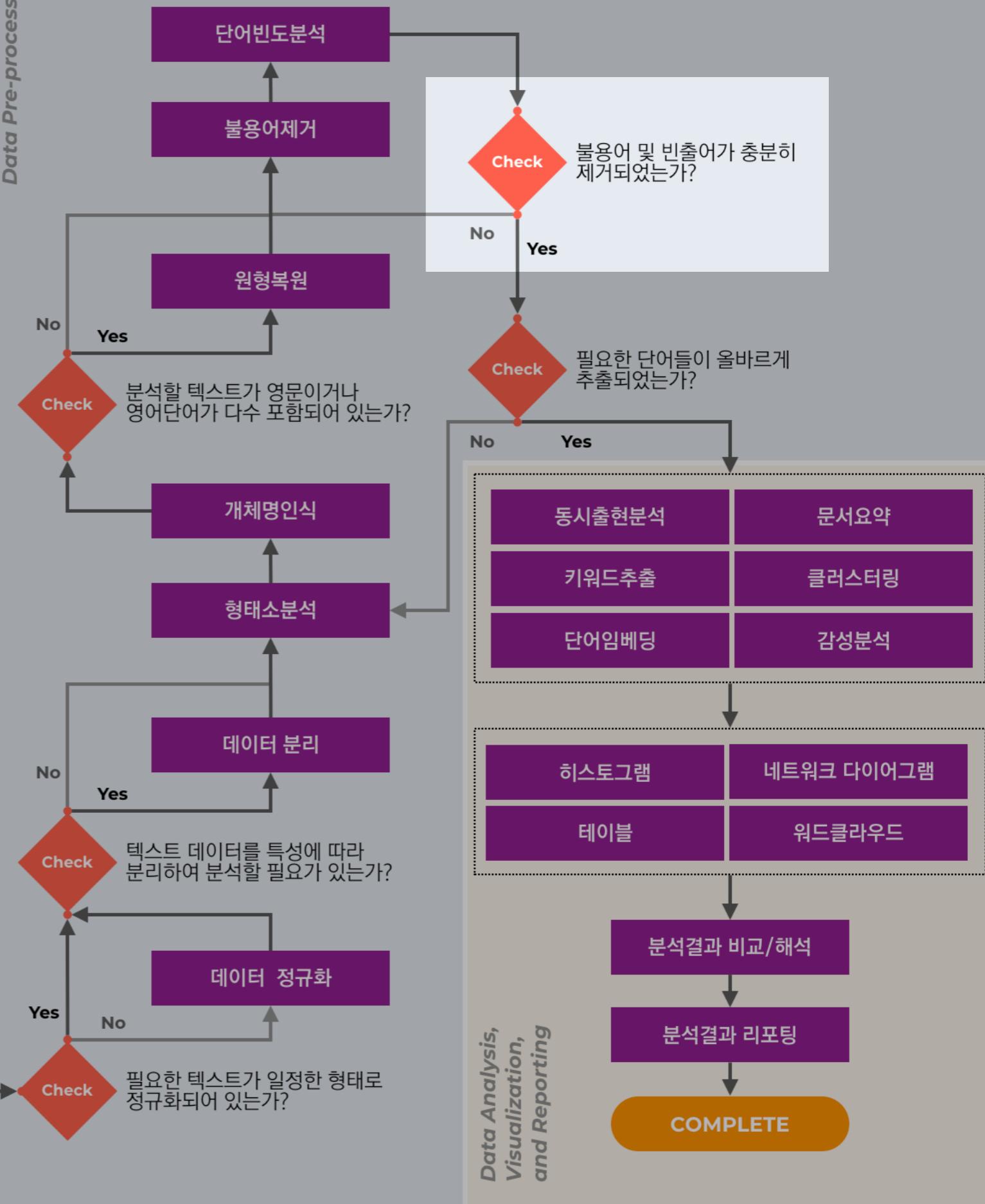
Data Preparation

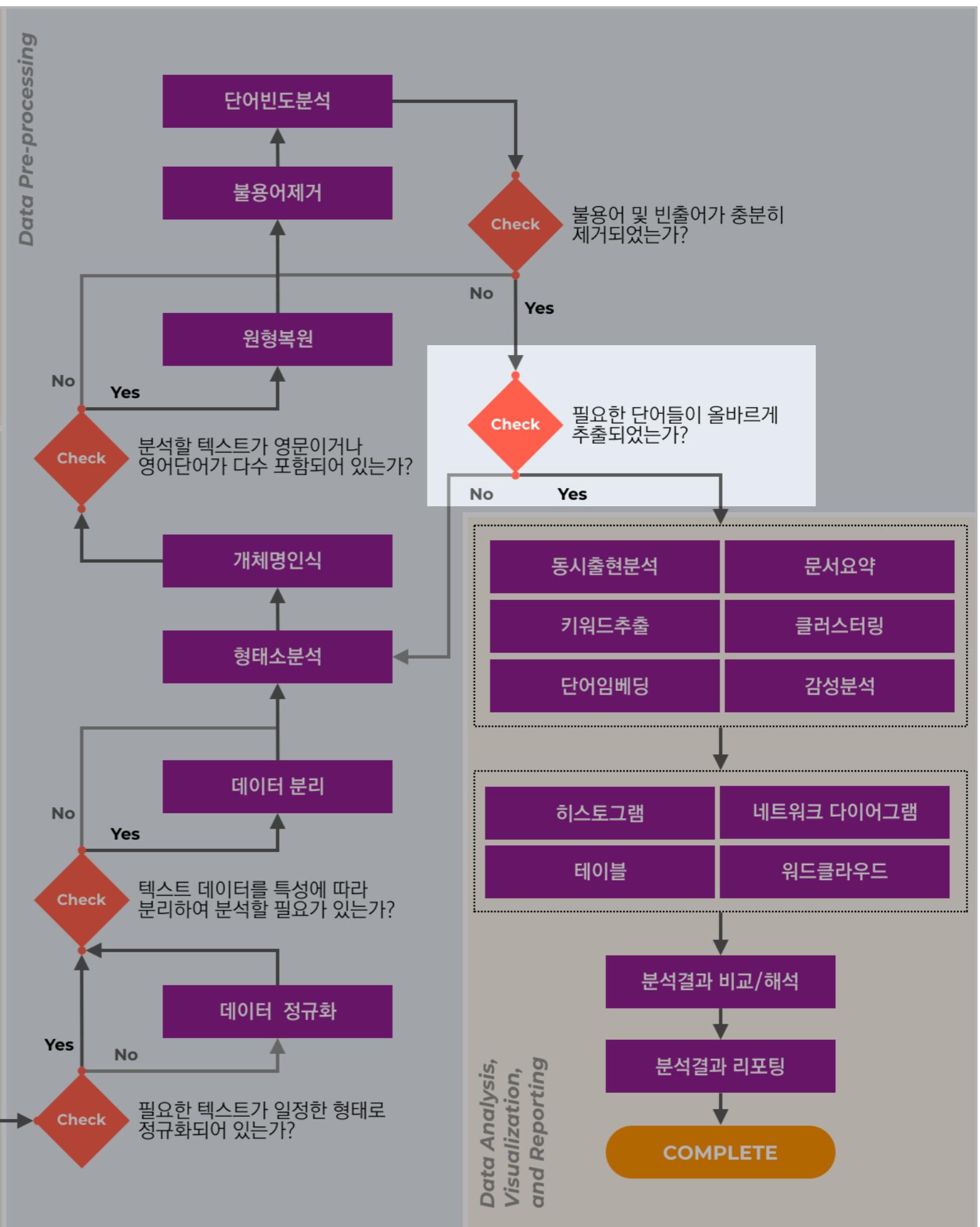
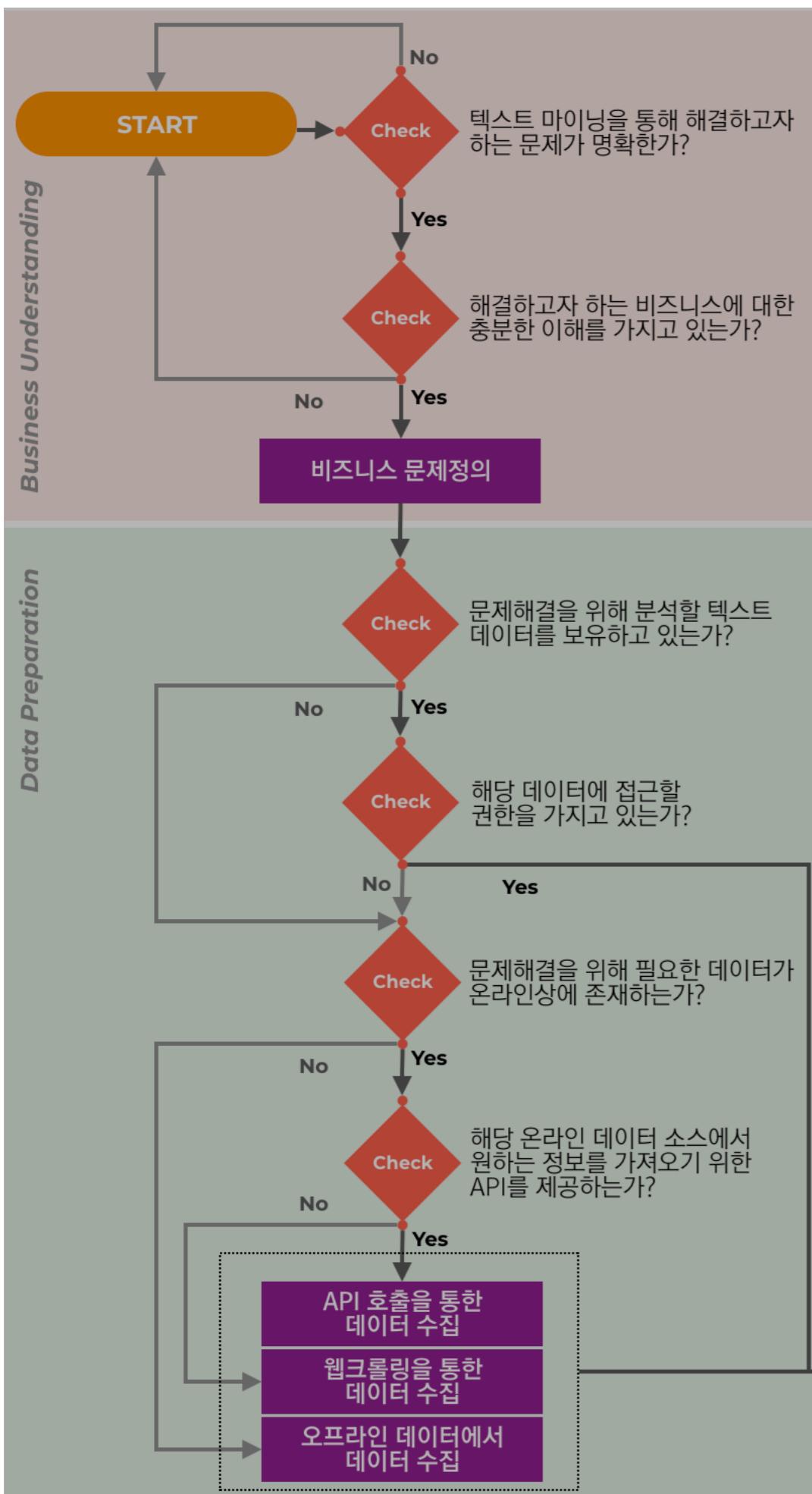


Business Understanding



Data Pre-processing





단어 빈도분석 (Word Frequency)

단어의 빈도를 바탕으로 가중치를 계산하는 분석방법

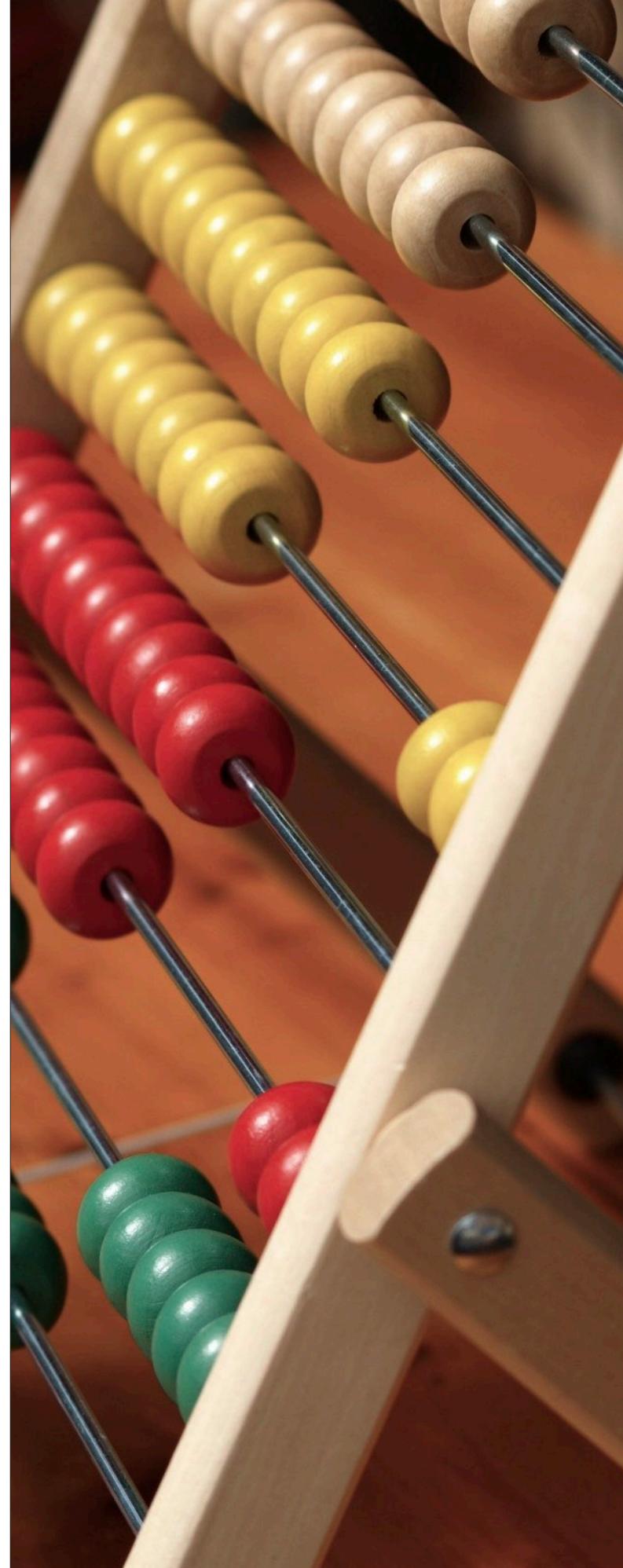
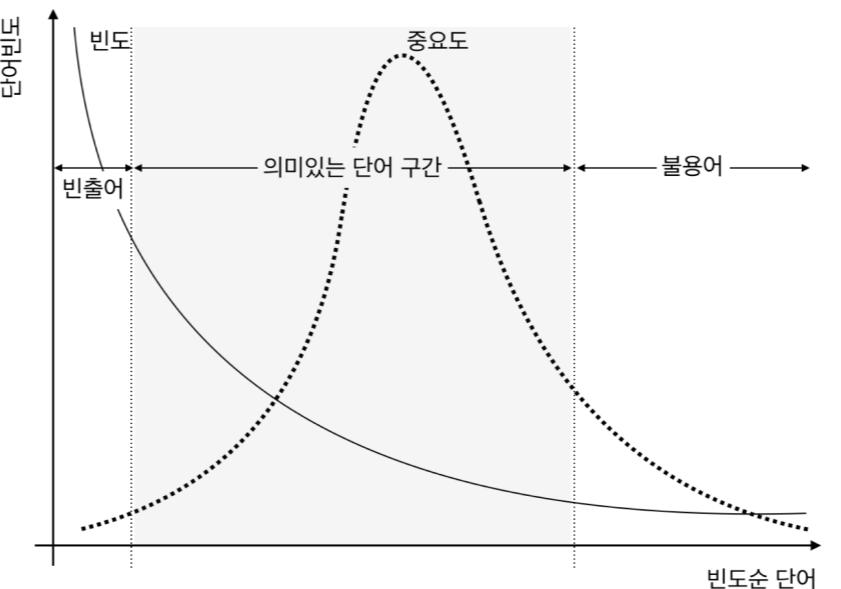
- ▶ 문서에 출현한 단어들의 출현 횟수를 기준으로 단순빈도, TF, TF-IDF 등을 계산
- ▶ 가장 간단한 텍스트 데이터 분석 방법이지만, 가장 빠르게 문서를 파악할 수 있으며 다른 분석방법을 수행하기 전에 반드시 한번이상 거쳐야 하는 과정
- ▶ 활용분야 : 문서요약, 내용 파악, 트렌드 분석, 불용어/빈출어 발견

단순 단어빈도 (Term Frequency, TF)

- ▶ 단어가 전체 문서에서 얼마나 흔하게 출현하는지 고려하는 방법
- ▶ 너무 희귀한 단어인 경우 또는 딱 한 번 나오는 단어는 의미를 부여하기 어려움
- ▶ 단어가 너무 흔한 경우 의미가 과도하게 부여될 가능성이 있음

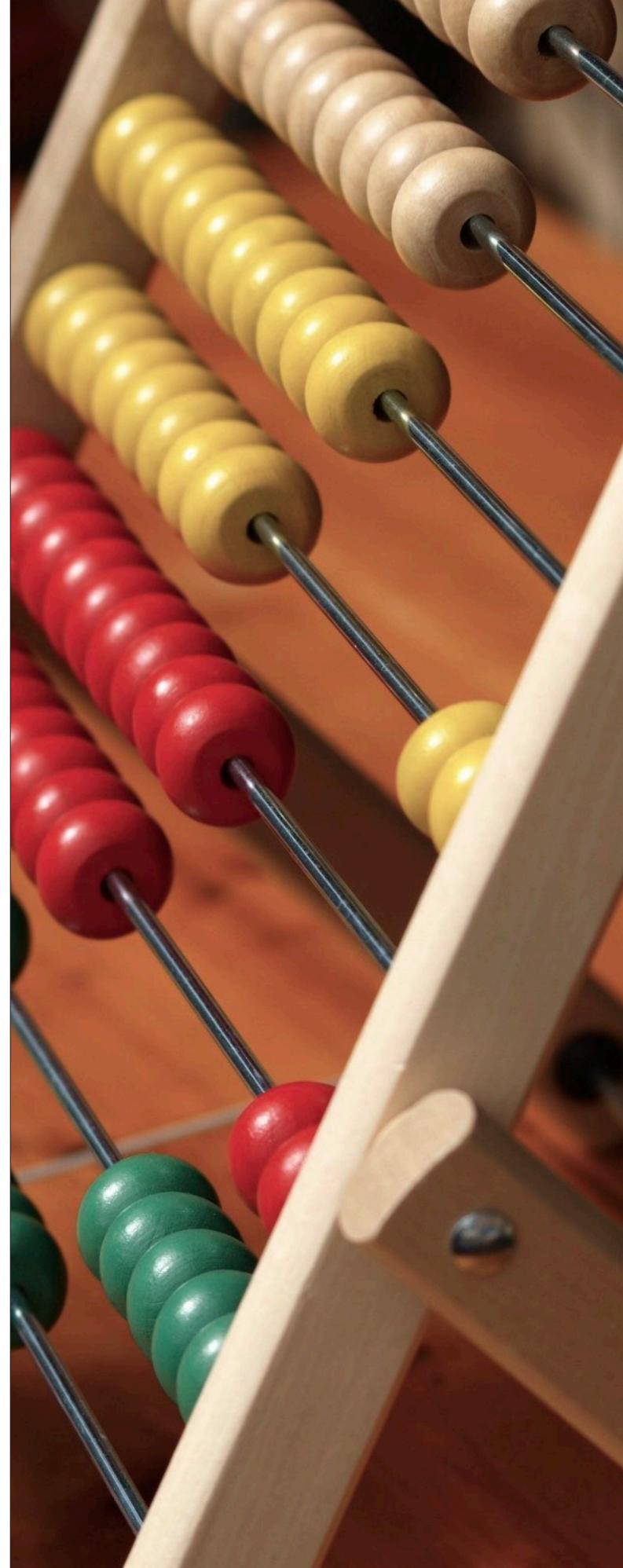
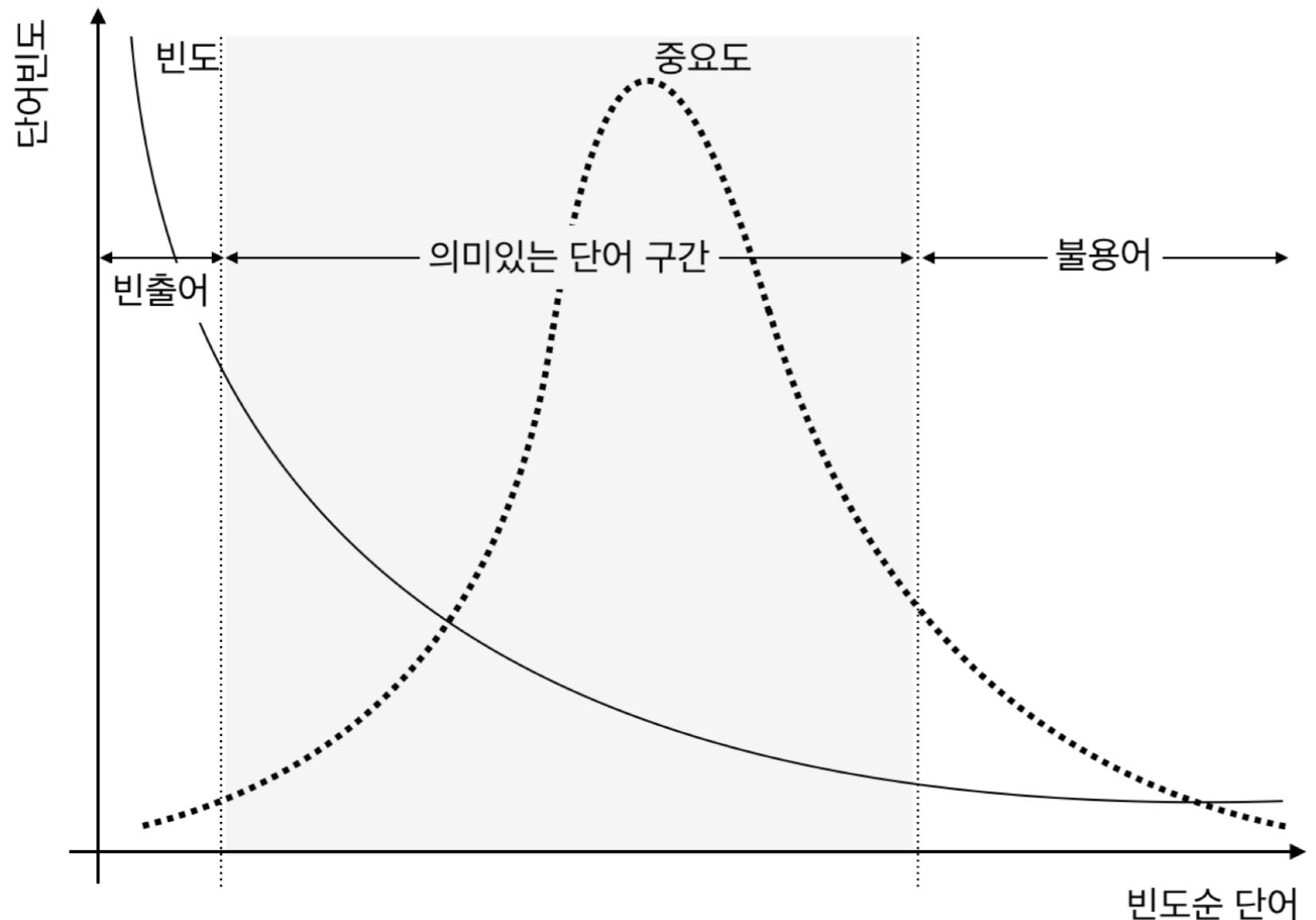
TermFrequency

$$= \text{count}(word | document)$$



단어 빈도분석

(Word Frequency)



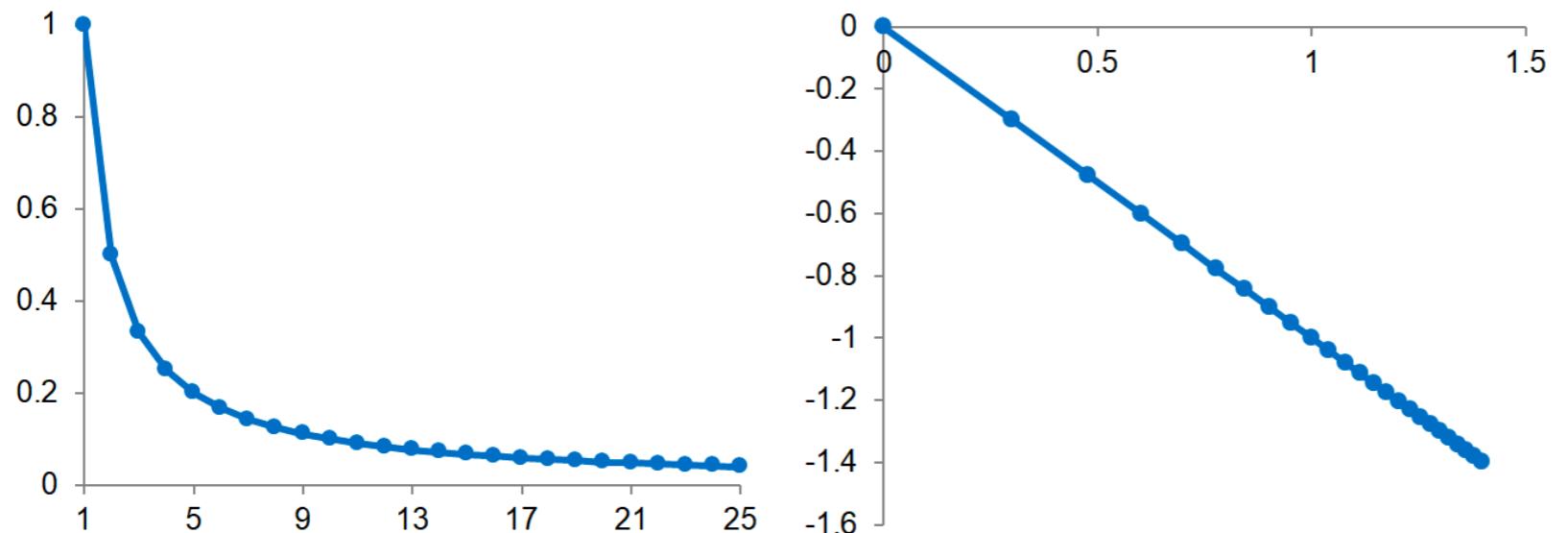
단어 빈도분석 (Word Frequency)

TF-IDF (TF-Inverse Document Frequency)

- ▶ 단어가 나온 문서의 수가 적을수록 단어가 문서에 중요할 가능성이 더 큼
- ▶ 단어의 희박성(sparseness)을 역문서빈도(IDF)로 측정
- ▶ 전체 문서 수가 고정된 체로 단어 t가 출현하는 문서가 많을수록 중요도가 감소

$$TF - IDF = Frequency * IDF$$

$$IDF = \log\left(\frac{N}{n_t} + 1\right)$$



- ▶ 지프의 법칙 (Zipf's law)

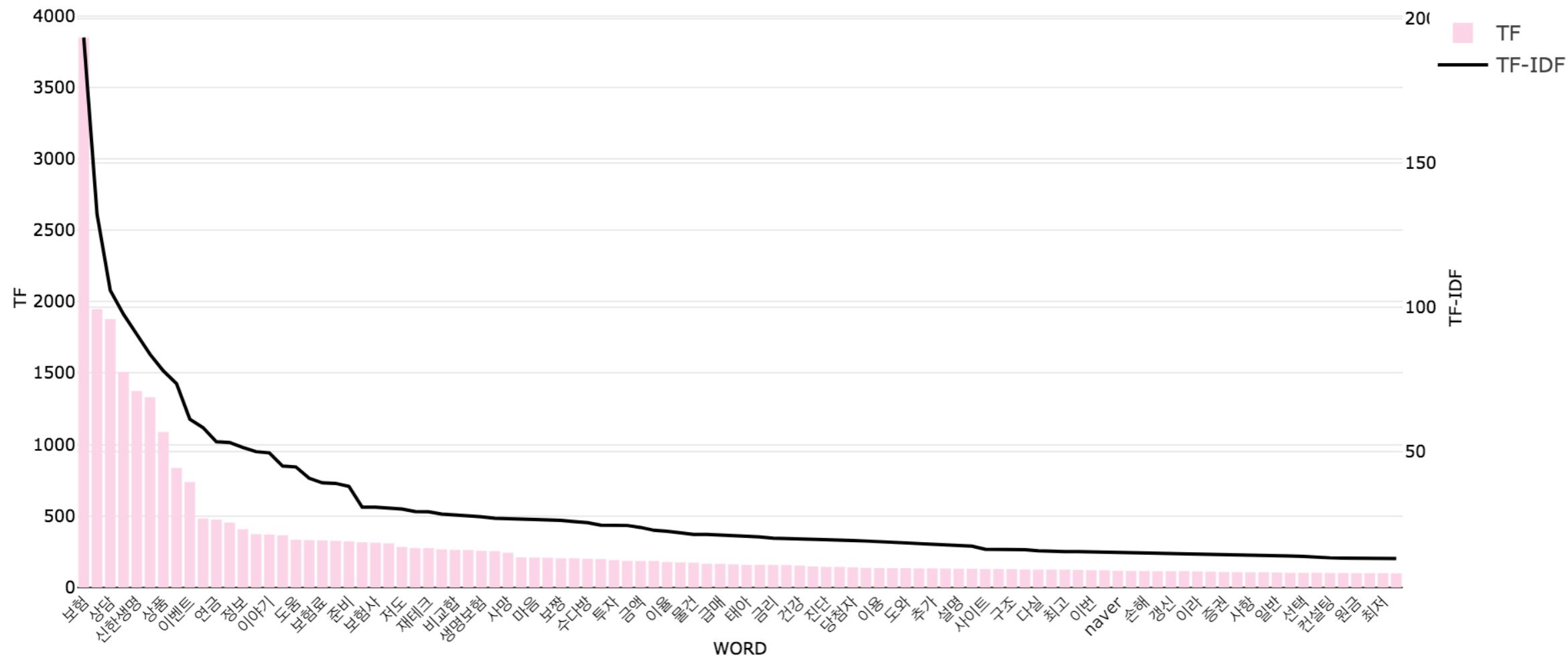
자연어 말뭉치 표현에 나타나는 단어들을 그 사용 빈도가 높은 순서대로 나열하였을 때, 모든 단어의 사용 빈도는 해당 단어의 순위에 반비례한다.

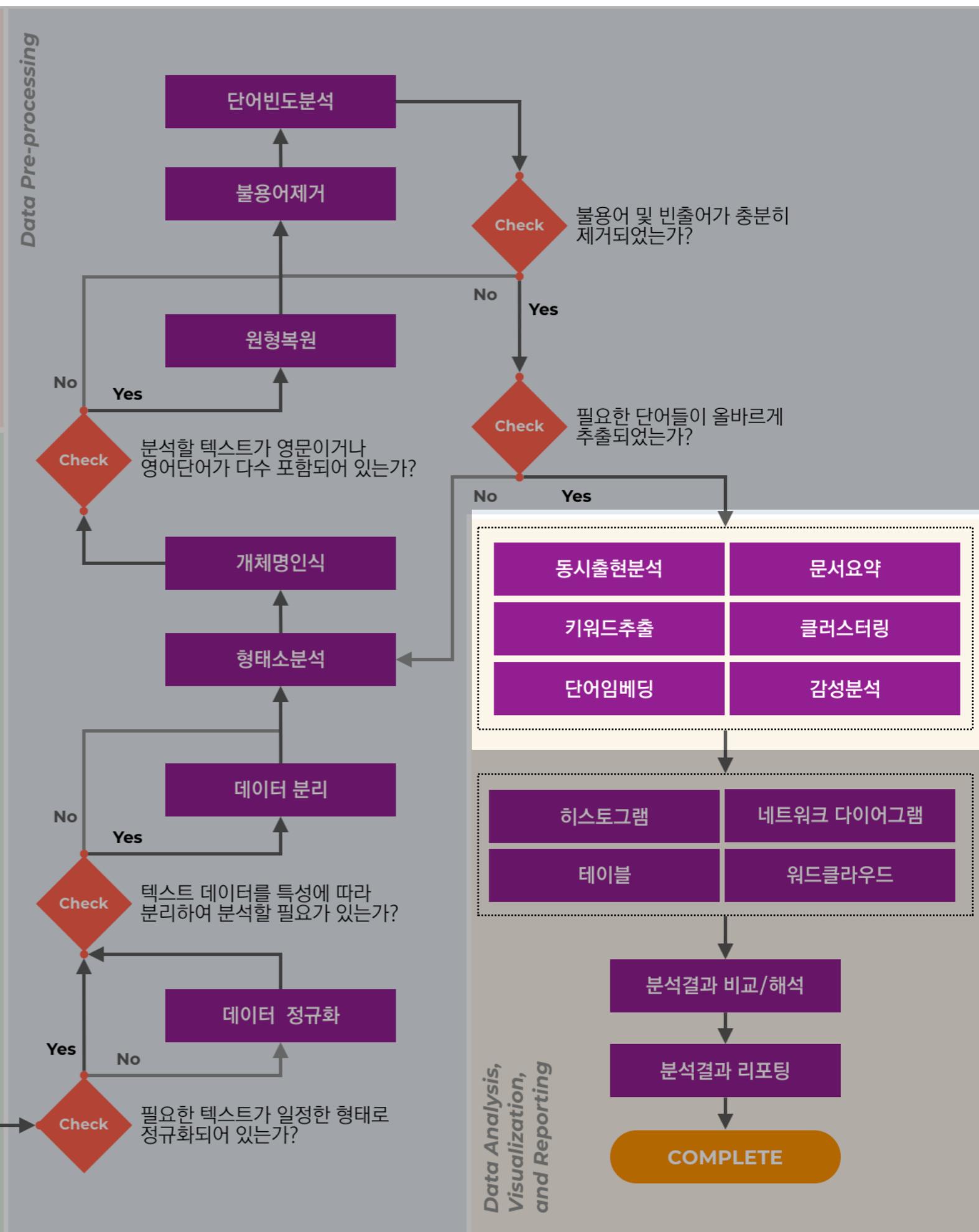
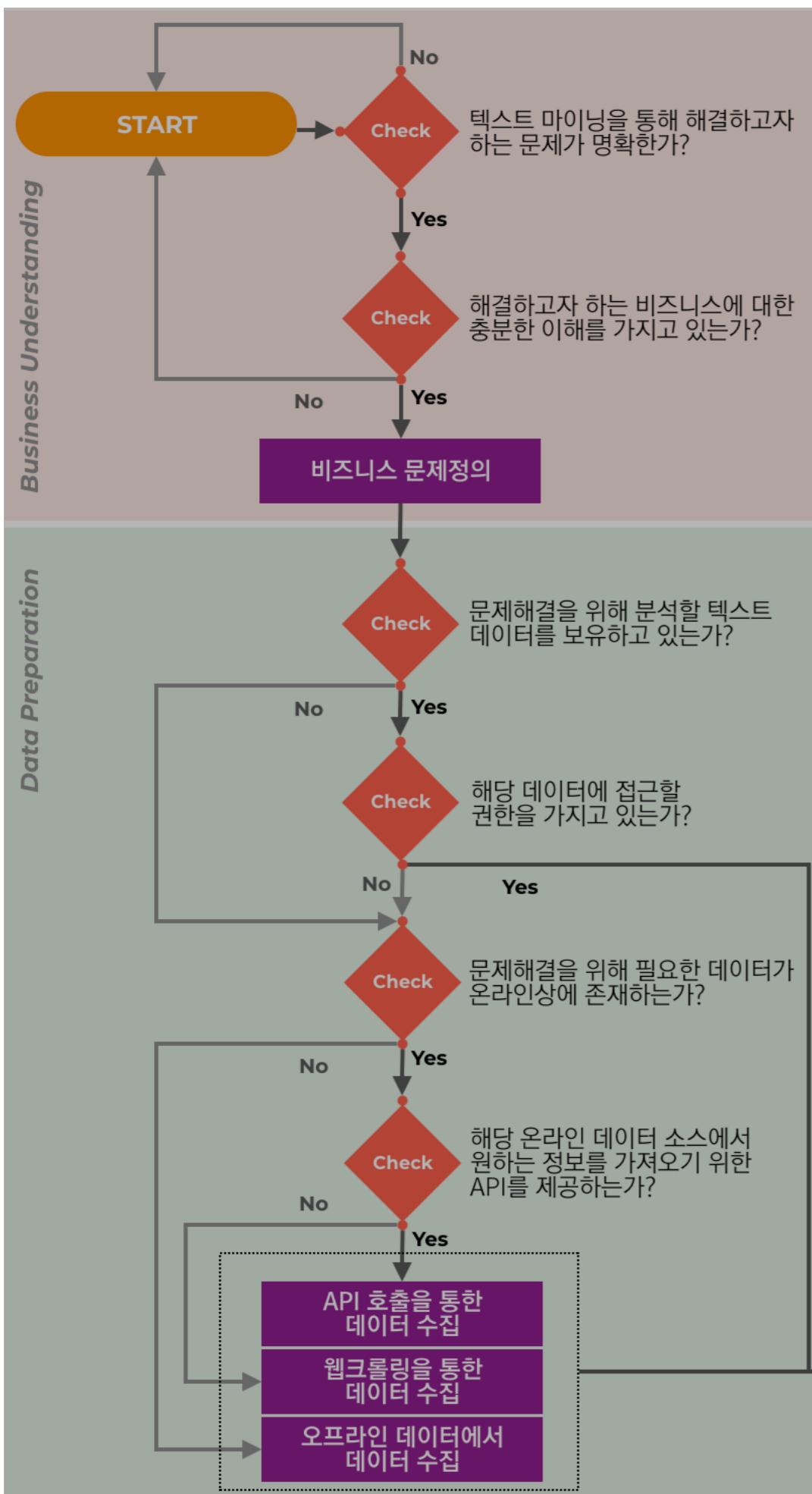
가장 사용 빈도가 높은 단어는 두 번째 단어보다 빈도가 약 두 배 높으며, 세 번째 단어보다는 빈도가 세 배 높다.

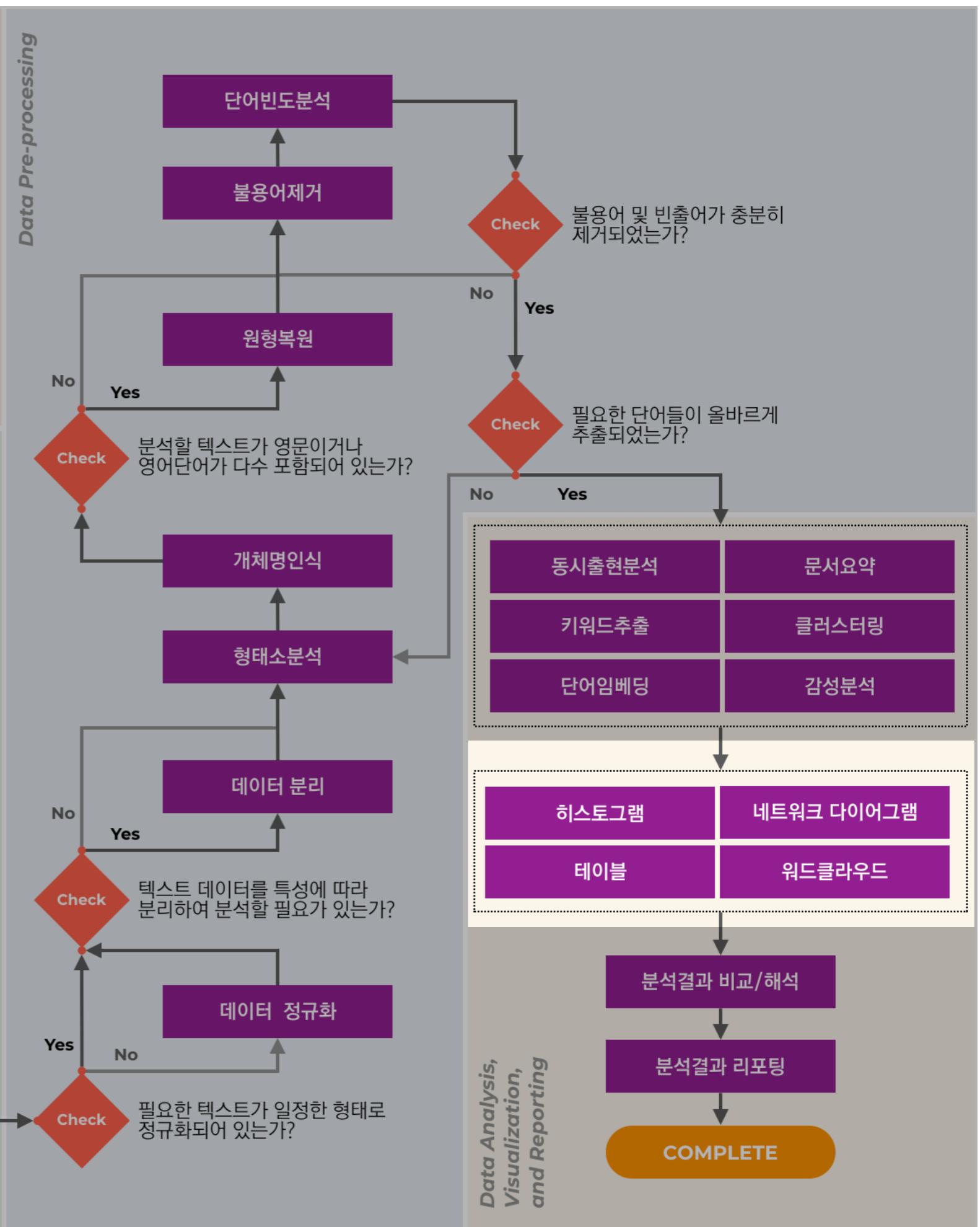
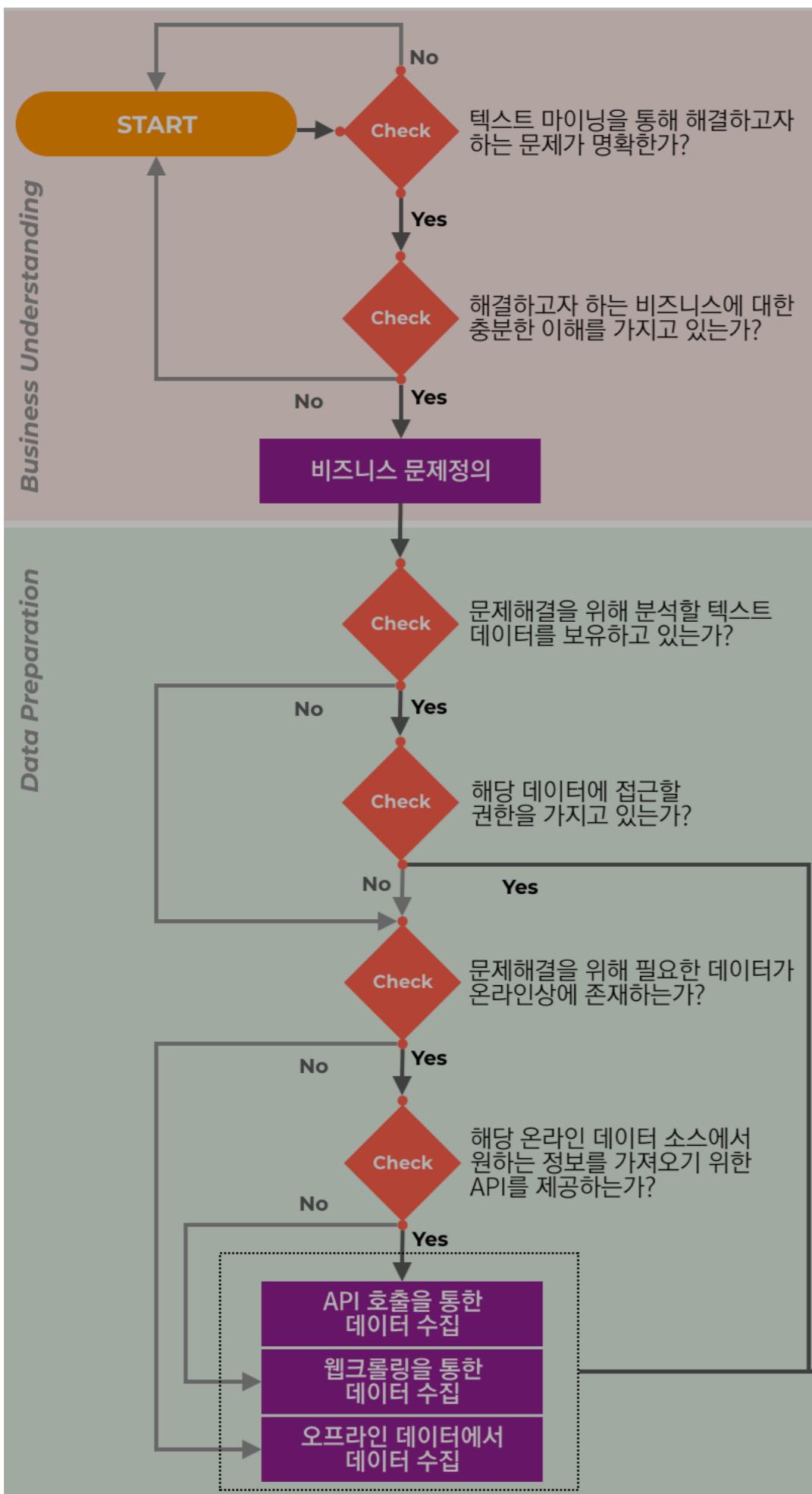


단어 빈도분석 (Word Frequency)

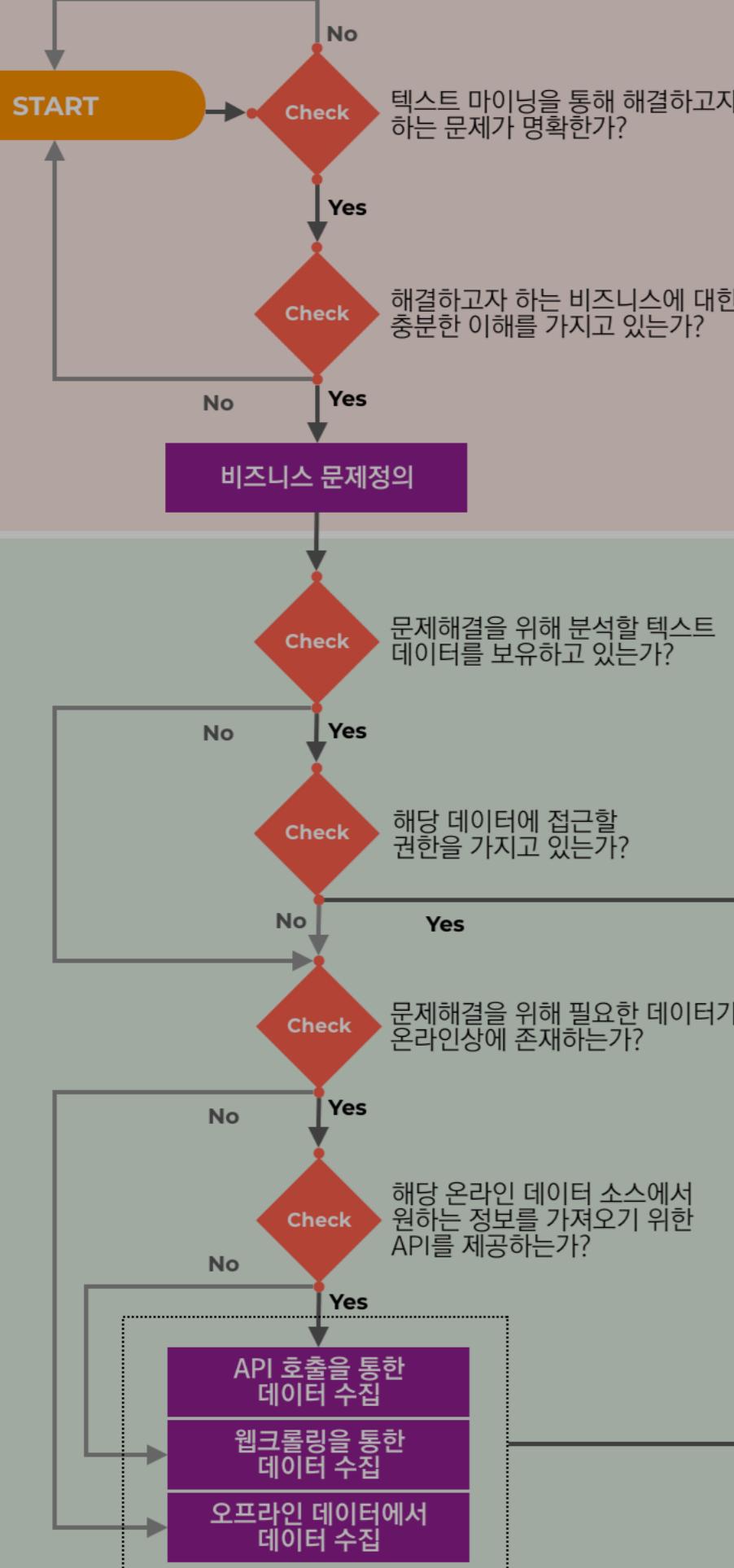
TF & TF-IDF Graph



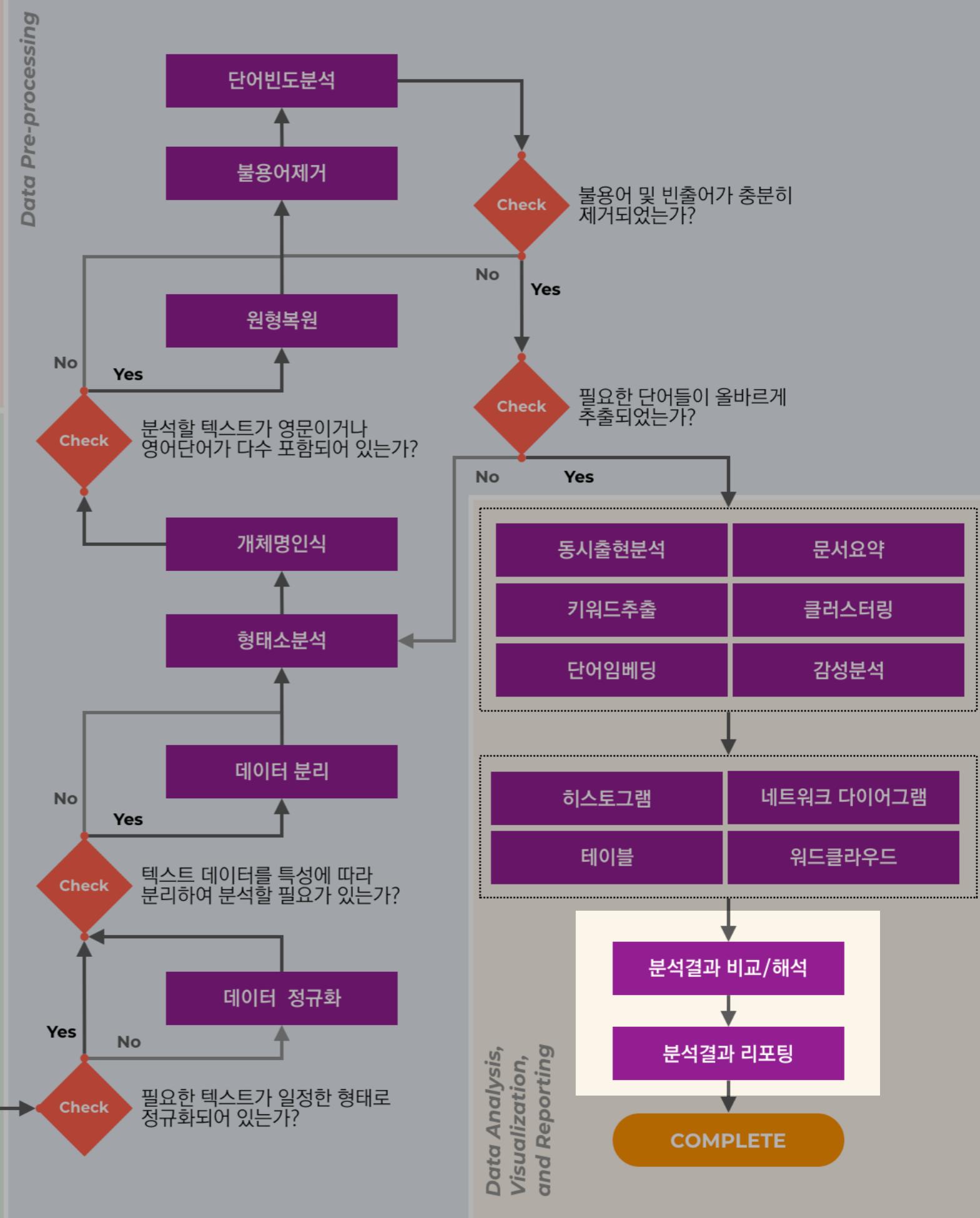


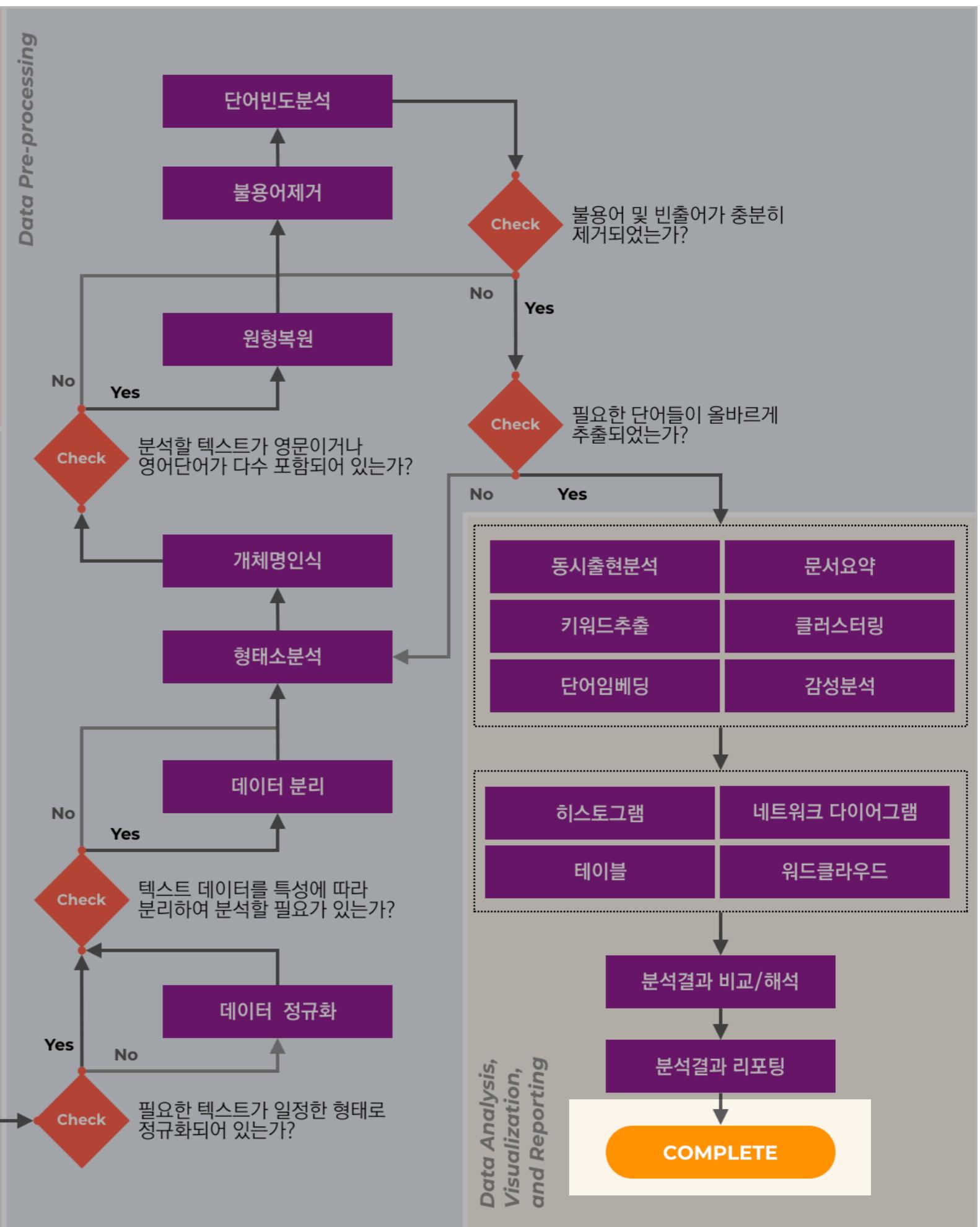
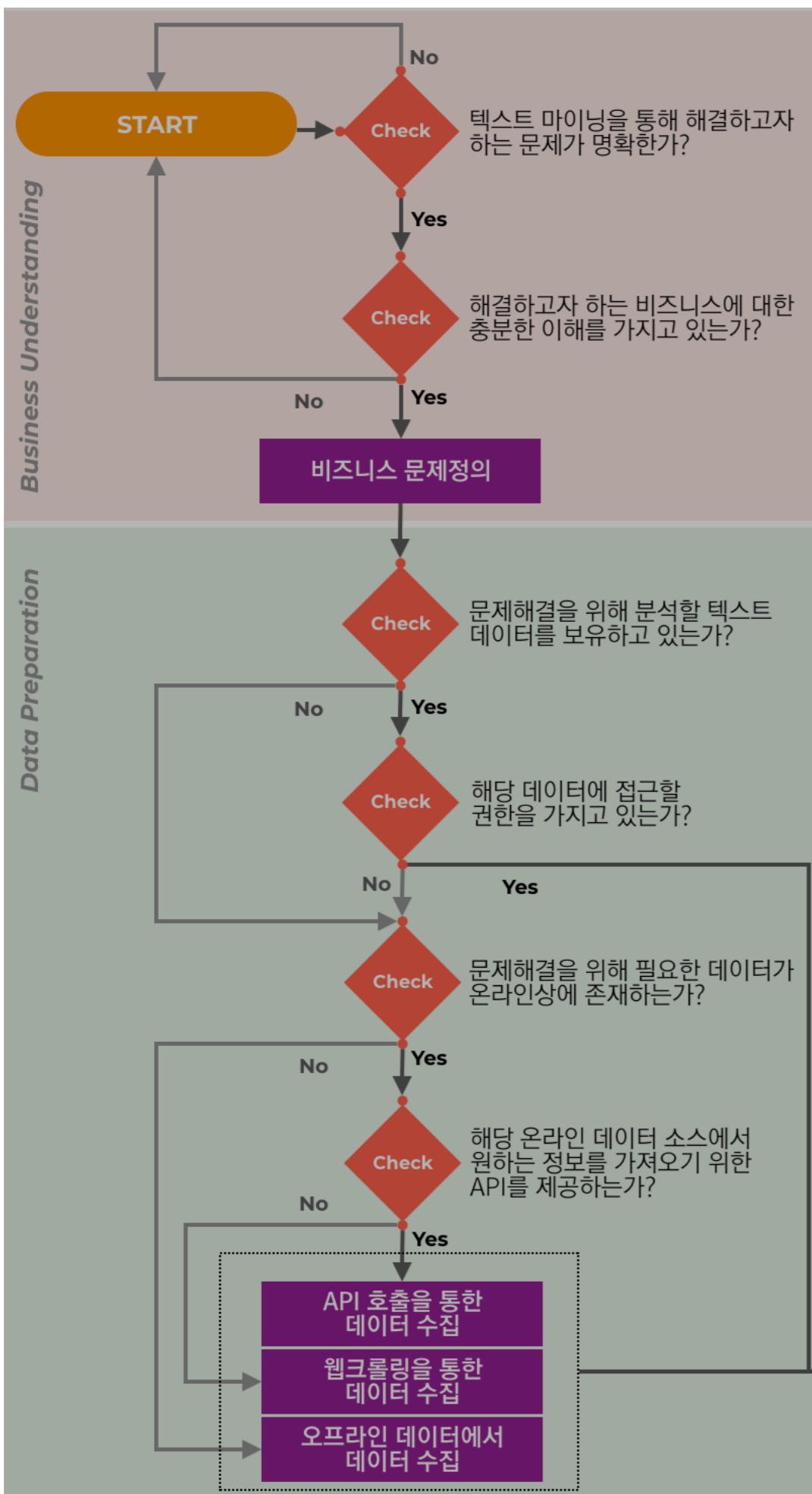


Business Understanding



Data Preparation





테이블 (Table)

분석결과를 테이블 형태로 구분하여 표현하는 방법

<표 6> 불행요인 세부 토픽 모델링 결과

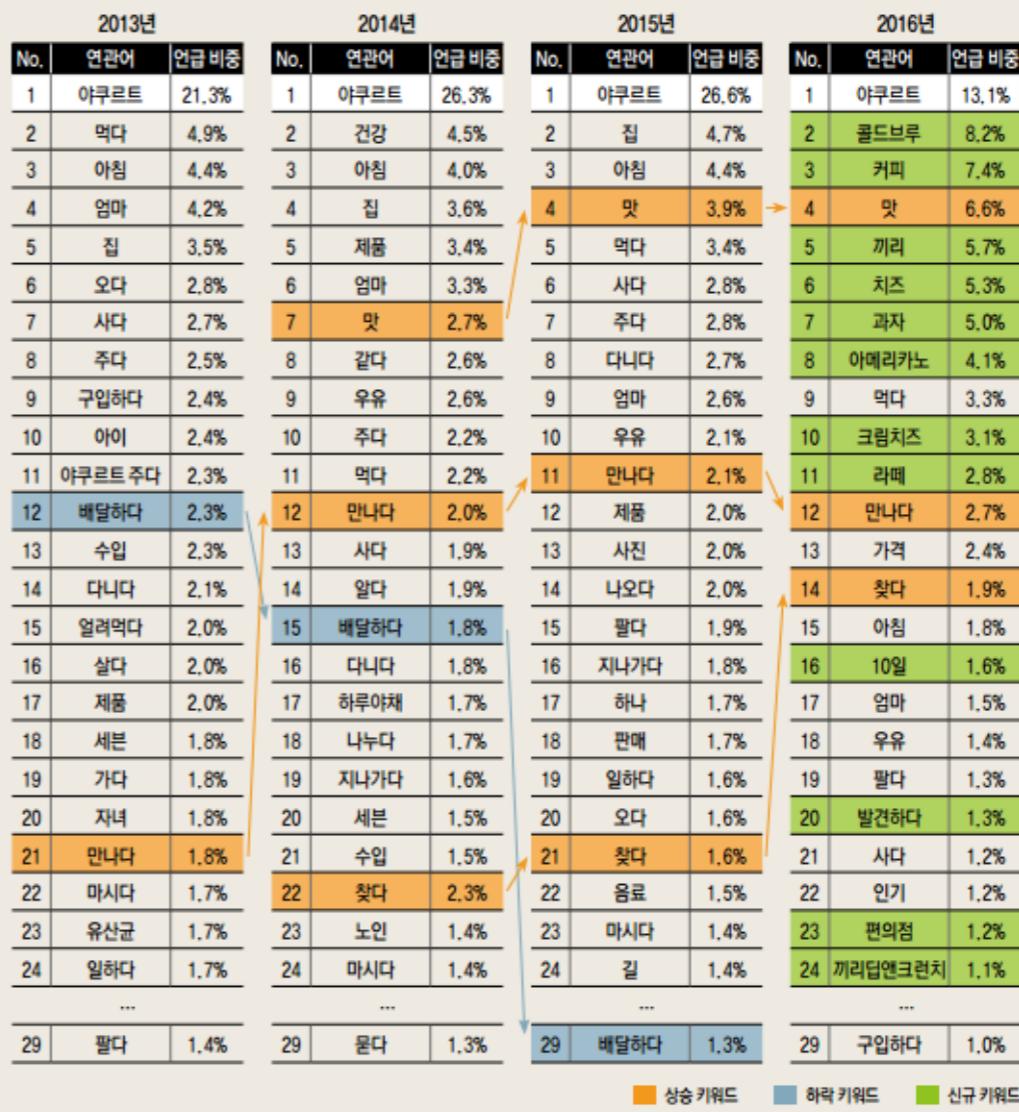
	토 퍽	키 워 드
1	가정 불화	불행, 사랑, 가족, 집, 아버지, 가정, 부모
2	가난	분배, 돈, 소득, 빈곤, 경제, 가난
3	자녀 문제	학교, 위험, 아이, 행동, 상황
4	부정적 인생관	불행, 사람, 인생, 마음, 성공, 공통점
5	인간관계 문제	불행, 자신, 관계, 마음, 생각, 환경, 상황
6	직업 불만족	불행, 사람, 생각, 인생, 직업, 친구
7	건강 문제	불행, 건강, 수명, 질병, 생명, 병, 사고
8	미 취업	오늘, 운세, 불행, 건강, 취업, 뱠띠, 금전
9	부정적 마음가짐	불행, 사람, 마음, 생각, 이기심, 자만심, 피해의식
10	-	예수, 교회, 신앙, 설교, 말씀, 축복

Table 10. Top Seller Characteristics of Rescator

#	Top key words	Interpretation
5	shop, wmz, icq, webmoney, price, dump,	Product: CCs, dumps (valid, verified);
6	валид (valid), чекер (checker), карты (cards), баланс (balance), карт (cards)	Payment: wmz, webmoney, bitcoin, lesspay;
8	shop, good, CCs, bases, update, cards, bitcoin, webmoney, validity, lesspay	Contact: shop, register, deposit, e-mail, icq, jabber
11	dollars, dumps, deposit, payment, sell, online, verified	
16	e-mail, shop, register, icq, account, jabber,	

표1 '아쿠르트 아줌마' 연관어 변화

아쿠르트 아줌마는 여전히 '아쿠르트'와의 연관도가 가장 높지만 2016년 들어 '커피' 및 '크림치즈' 제품 연관어와 '10일'이라는 키워드가 등장. 아쿠르트 아줌마는 '배달하는' 역할에서 맛난 제품을 위해 '만나고' '찾고' '발견하는' 대상으로 변화 중.



*Source : 양승준, 이보연, & 김희웅. (2016). 토픽모델링 기반 행복과 불행 이슈 분석 및 행복 증진 방안 연구. 지식경영연구, 17(2), 165-185.

**Source : Li et al., (2016). Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. Journal of Management Information Systems, 33(4), 1059-1086.

***Source : 백경혜(DBR), “매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다”, 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935/.

워드클라우드 (Wordcloud)

단어의 빈도를 반영해 그 분포를 시각화하는 과정

- ▶ 단어의 크기를 단어의 빈도 수에 비례하도록 아름답게 표현하는 방법
 - ▶ 일반적인 워드클라우드는 빈도 외에 다른 정보를 제공하지 않은나, 단어의 배치에 따라 더 많은 정보를 제공하기도 함



*Source : 몬데이터, [mondata] 남북정상회담 판문점 선언 Text 키워드 분석, 2018.4.28., <https://www.youtube.com/watch?v=ba4EMdzSK-A>.

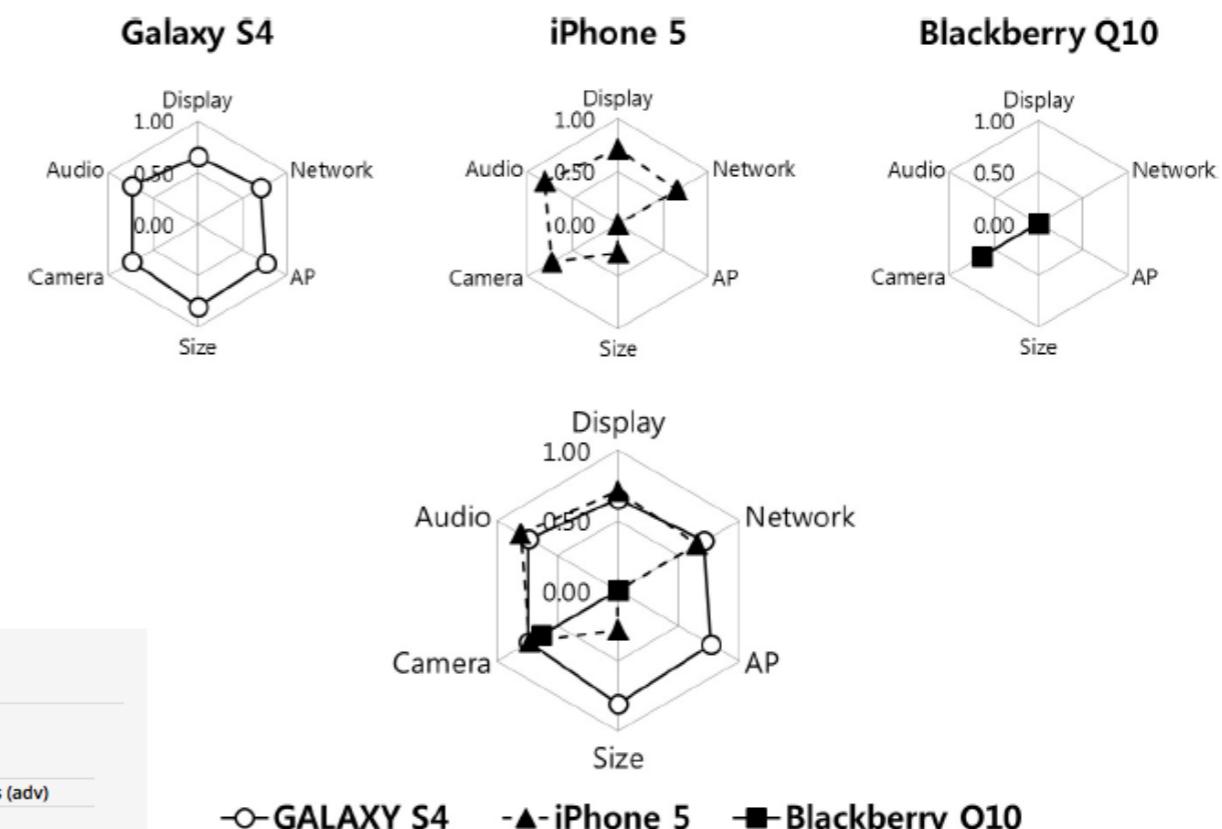
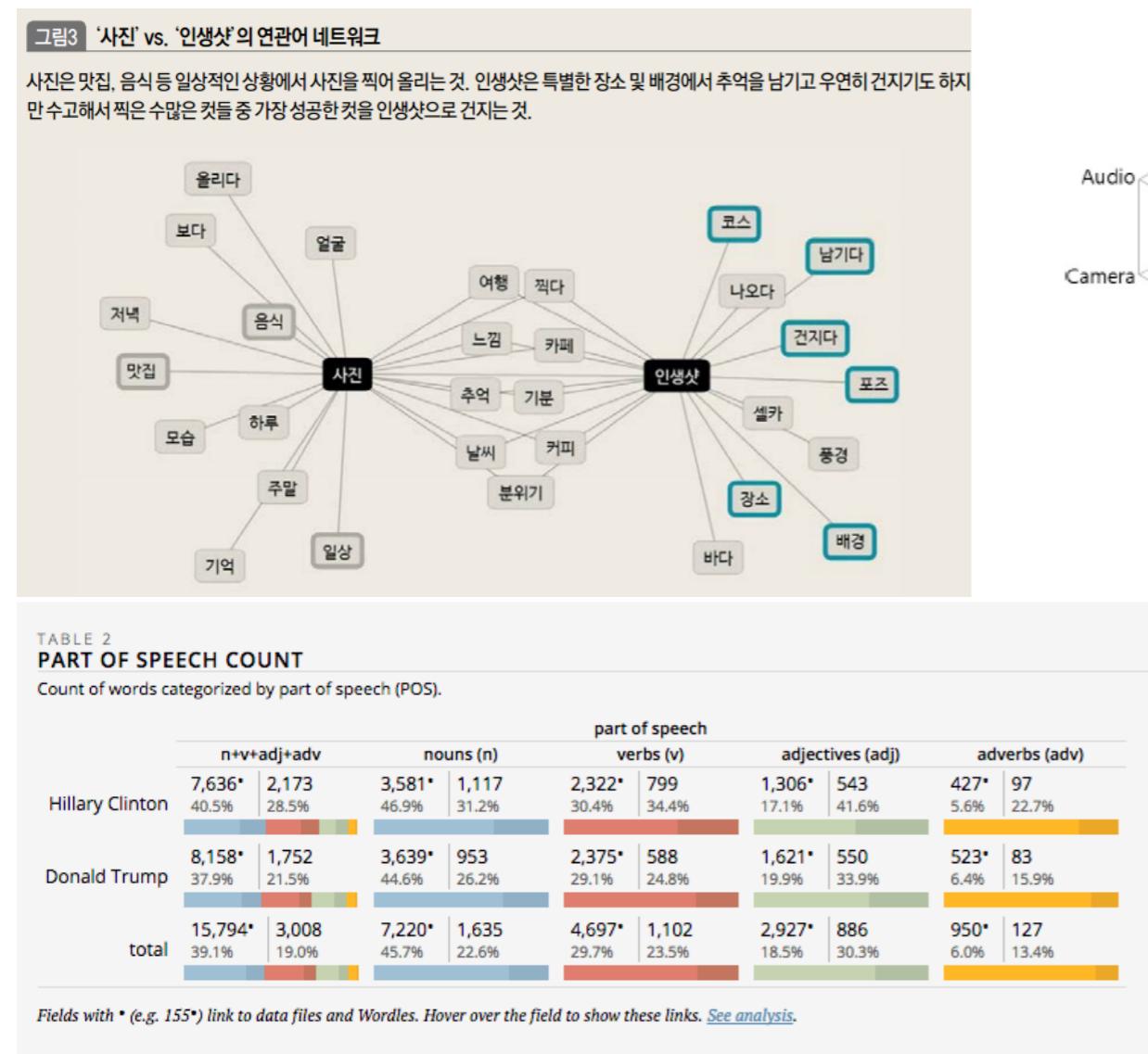
**Source : NÉSTOR CORREA, Cómo implementar el Big Data en tu empresa, 2017., <http://bluelight.tistory.com/298/>

***Source : 전병진, 신한은행 파이썬으로 시작하는 데이터분석: 텍스트 마이닝 기초, 2018.12.12.

그래프 & 네트워크 (Graph & Network)

단어 사이의 관계 강도를 시각화하는 과정

- ▶ 그래프 : 문서 또는 단어의 정량화된 특징을 도표로 표현하는 방법
- ▶ 네트워크 : 단어를 노드, 단어들 사이의 관계를 엣지로 취급하여 네트워크를 표현하는 방법



*Source : 백혜(DBR), “매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다”, 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935/.

**Source : 최홍규(슬로우뉴스), 2016 미국 대선을 보여주는 텍스트 마이닝 분석방법들, 2017.1.9., <http://slownews.kr/60919/>.

***Source : Kim et al. (2014). Analysis on smartphone related twitter reviews by using opinion mining techniques. In Advanced Approaches to Intelligent Information and Database Systems (pp. 205-212).

E.O.D