

TEXT MINING for PRACTICE

Python을 활용한 비정형 데이터 분석 - WEEK 03

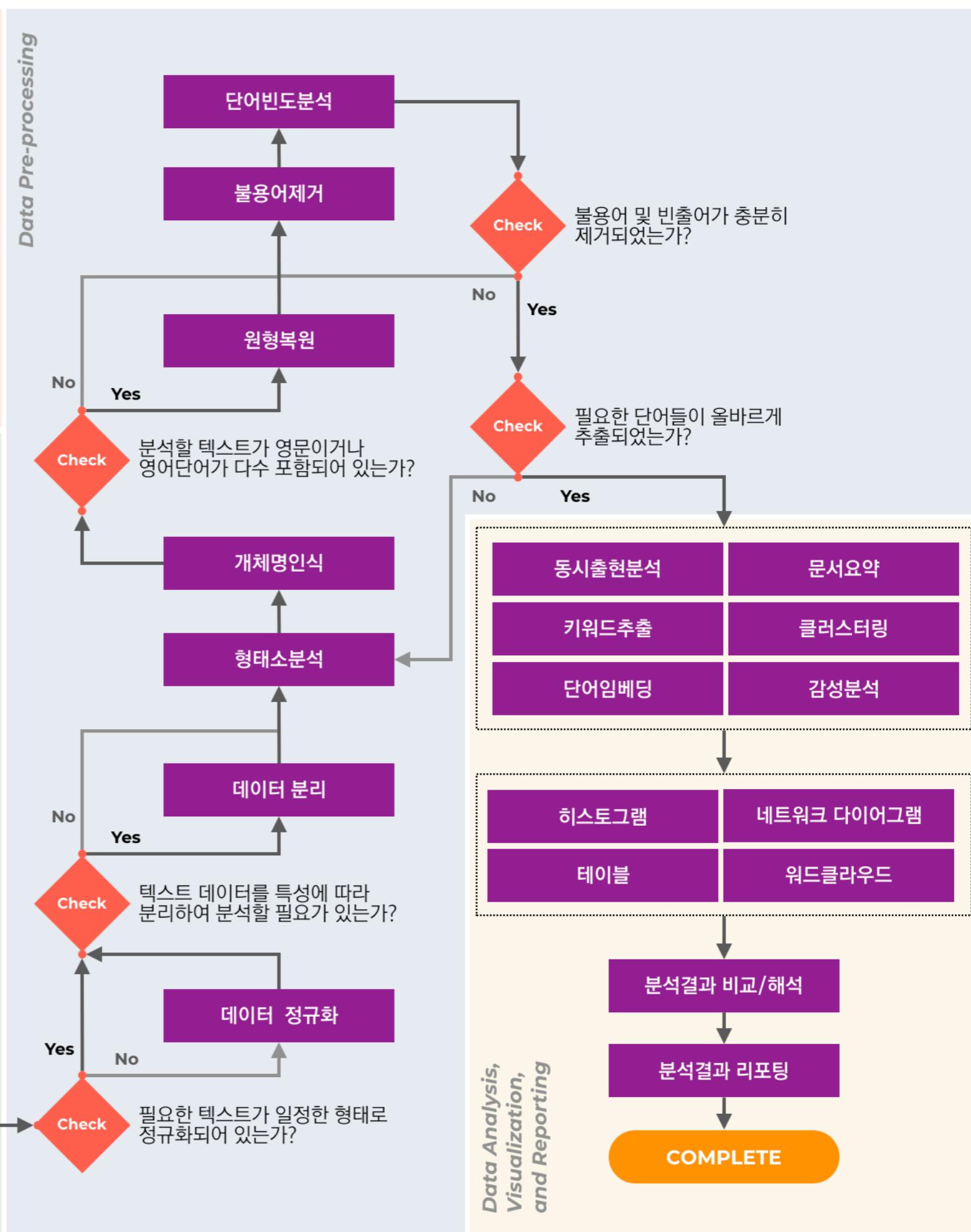
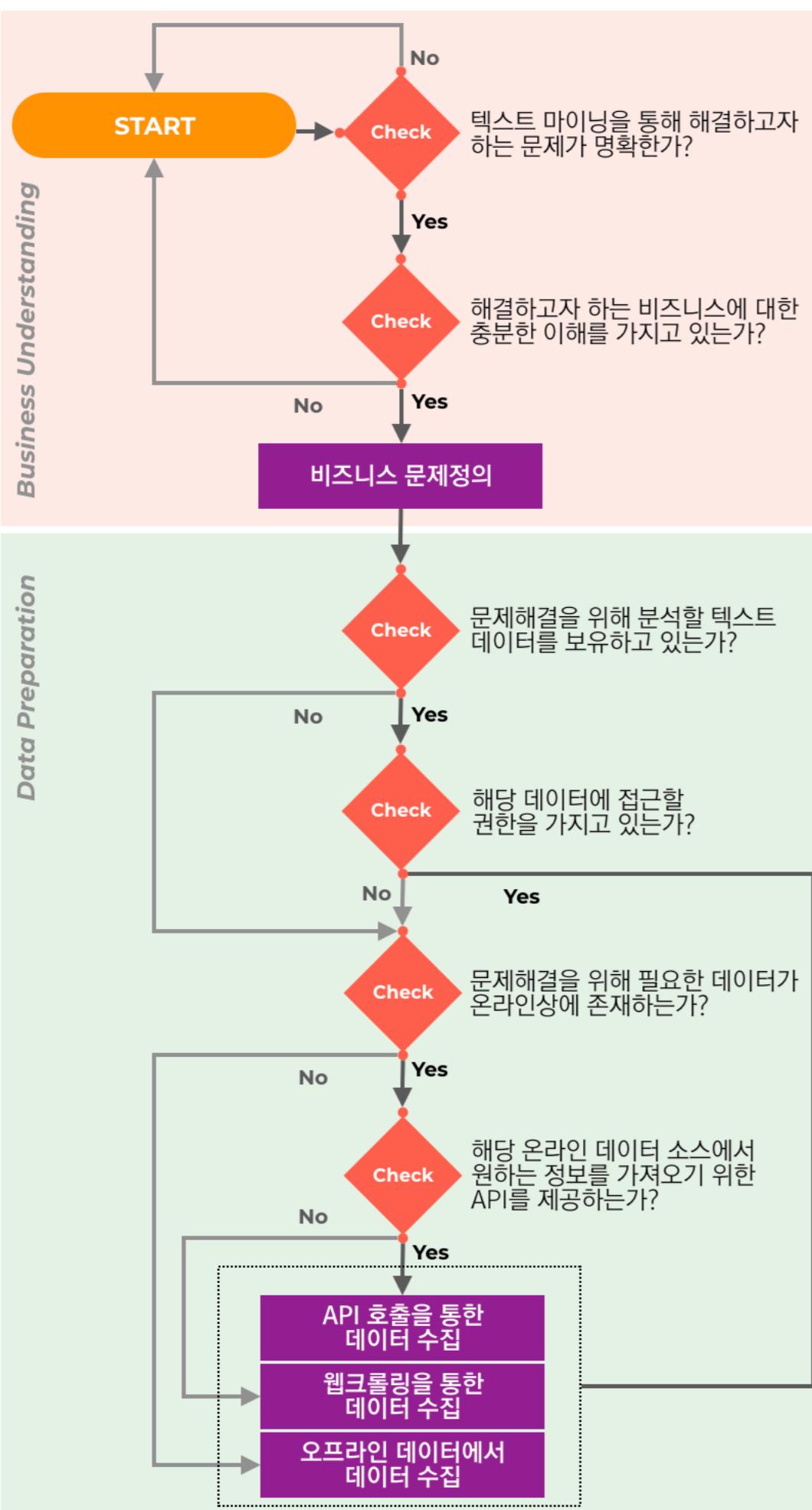
전병진 FINGEREDMAN (fingeredman@gmail.com)

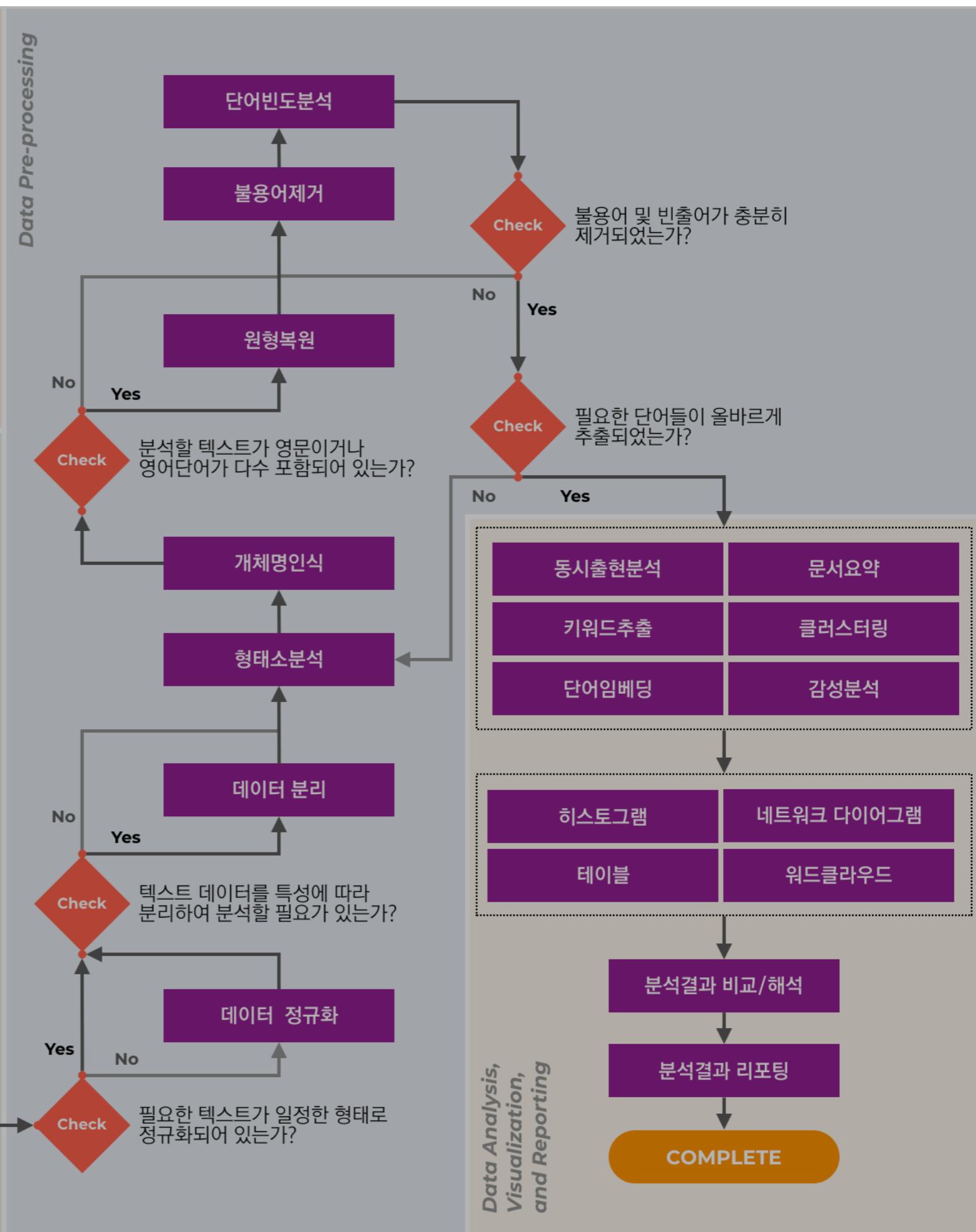
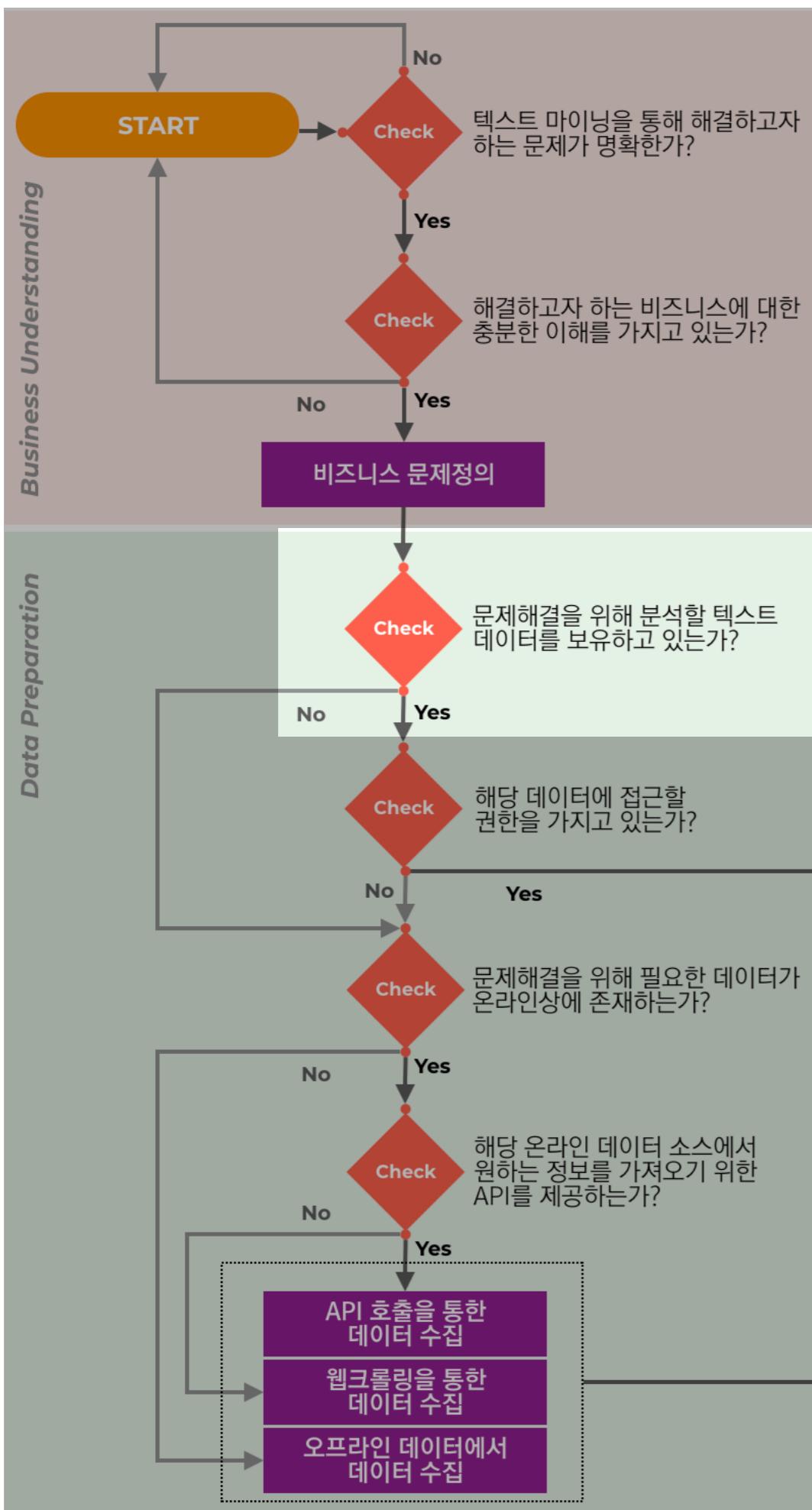
Part 2.

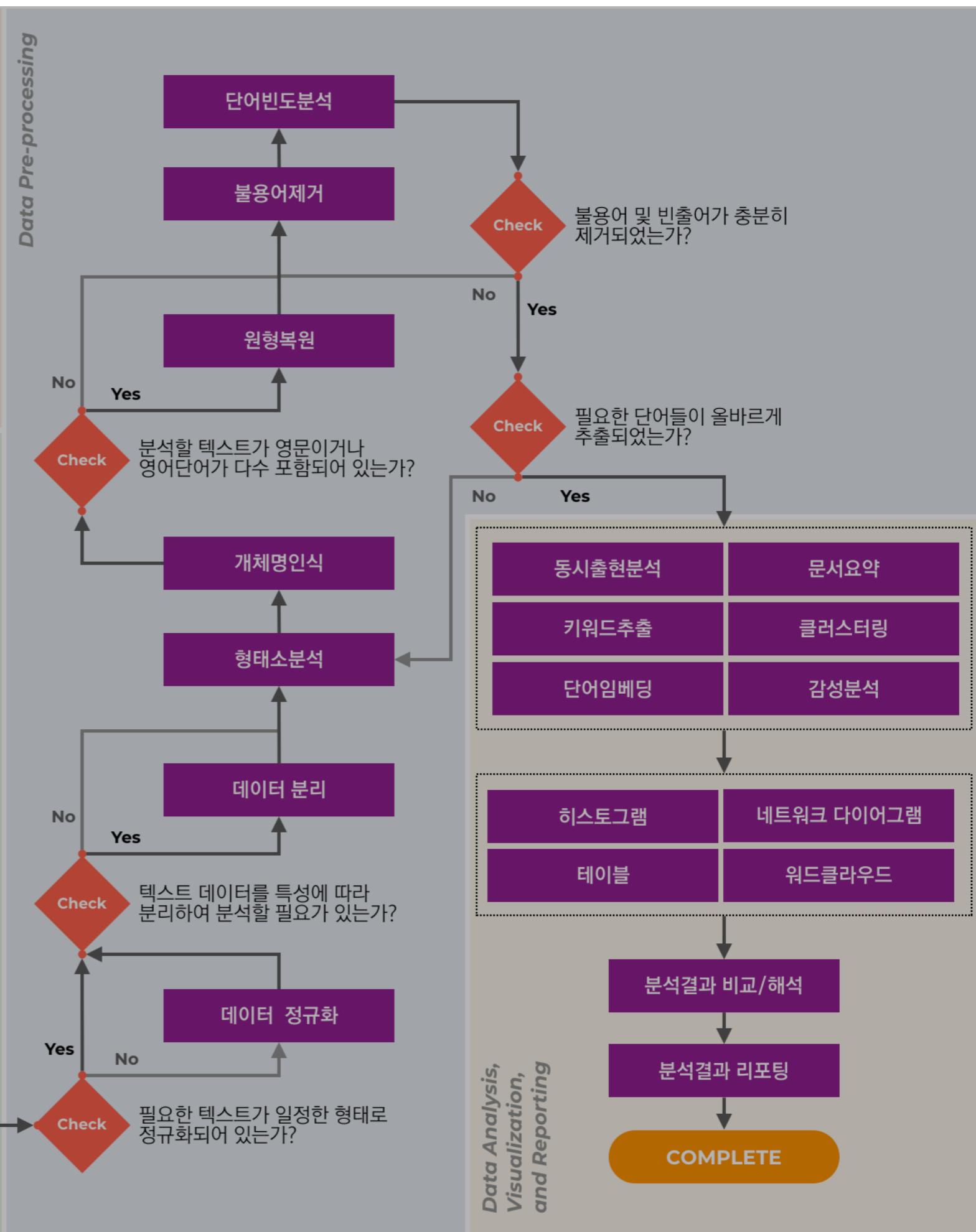
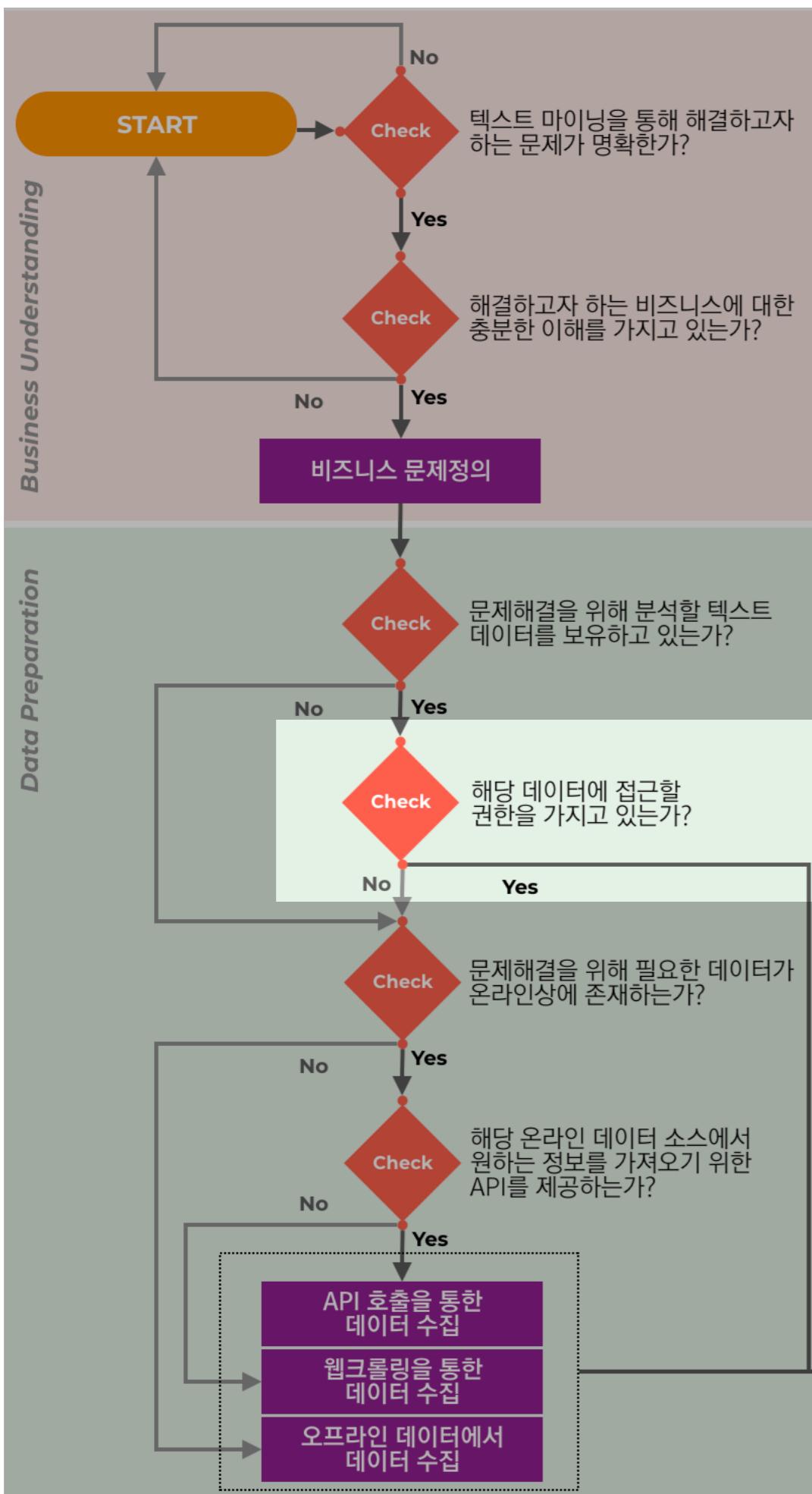
텍스트 데이터 유형 및 수집

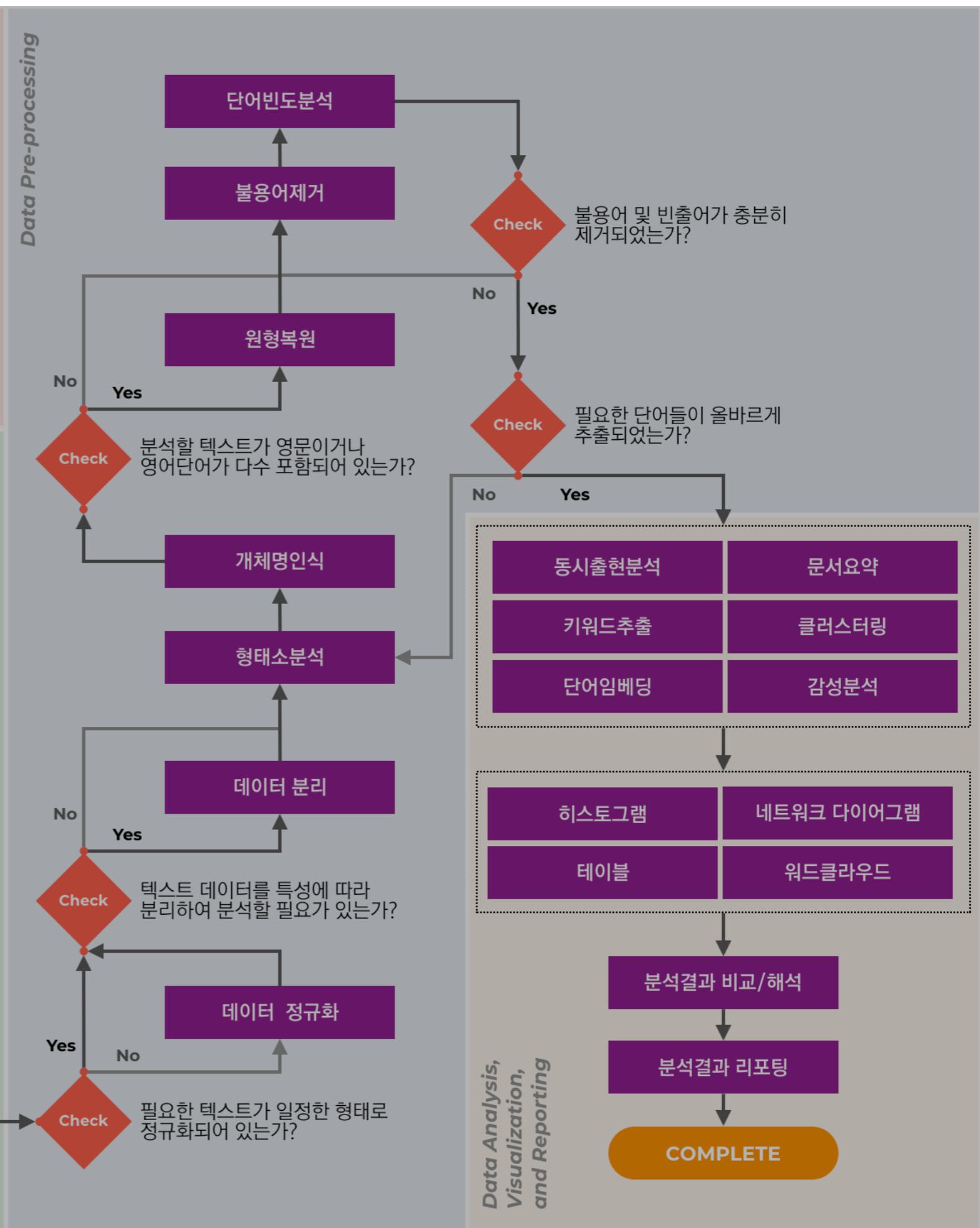
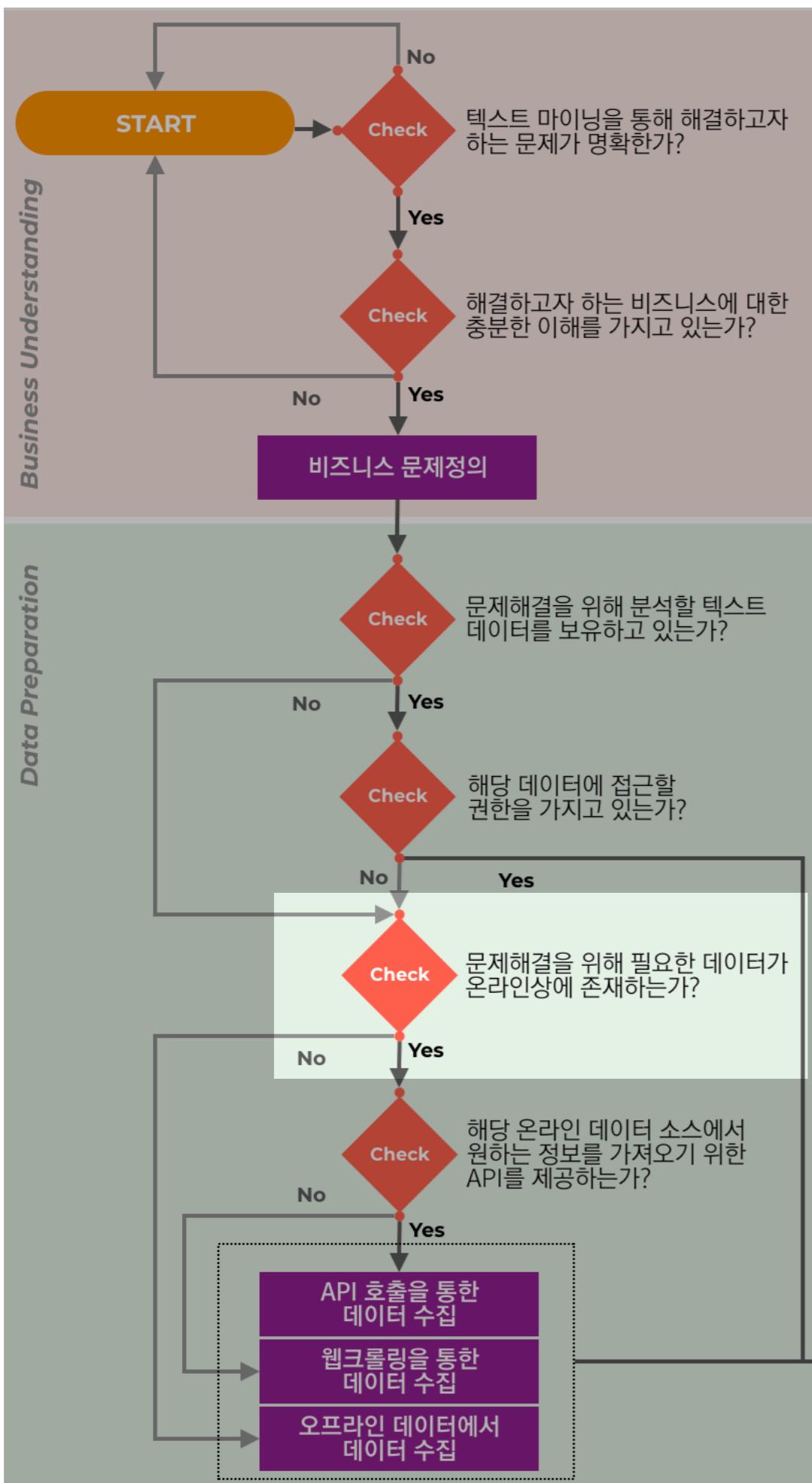
텍스트 데이터 수집

아무데서나 찾은 텍스트는
아무의미없는 정보를 포함한다

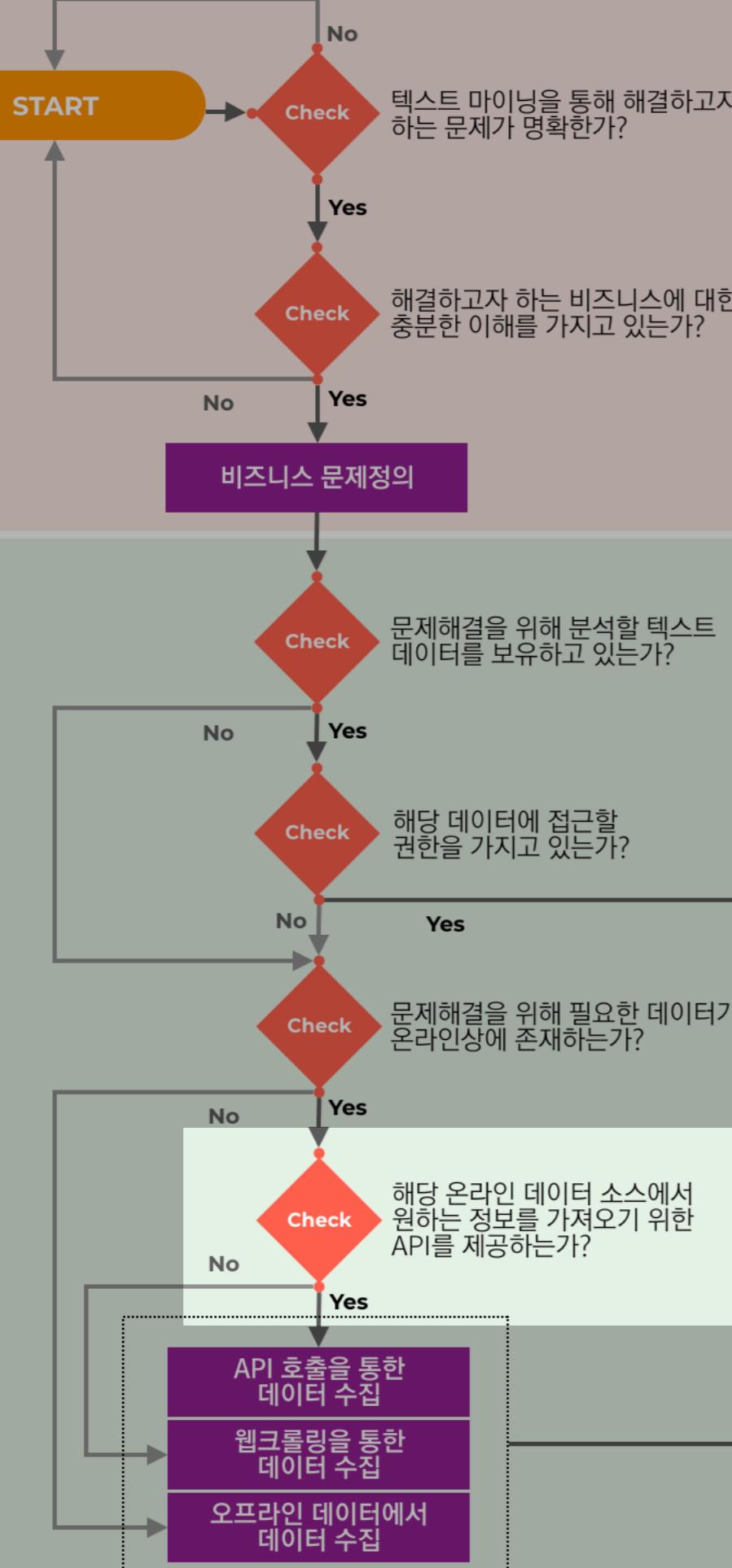




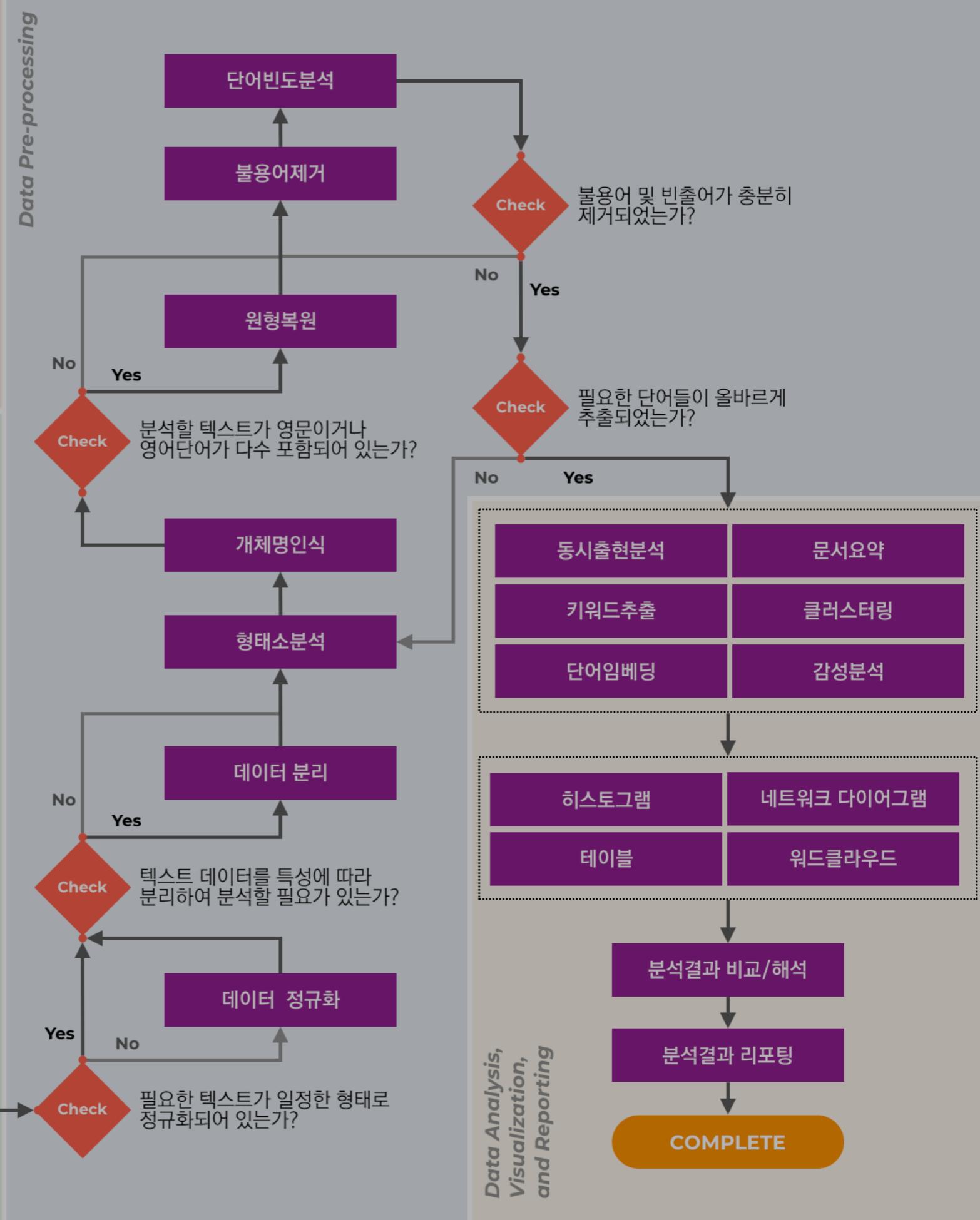


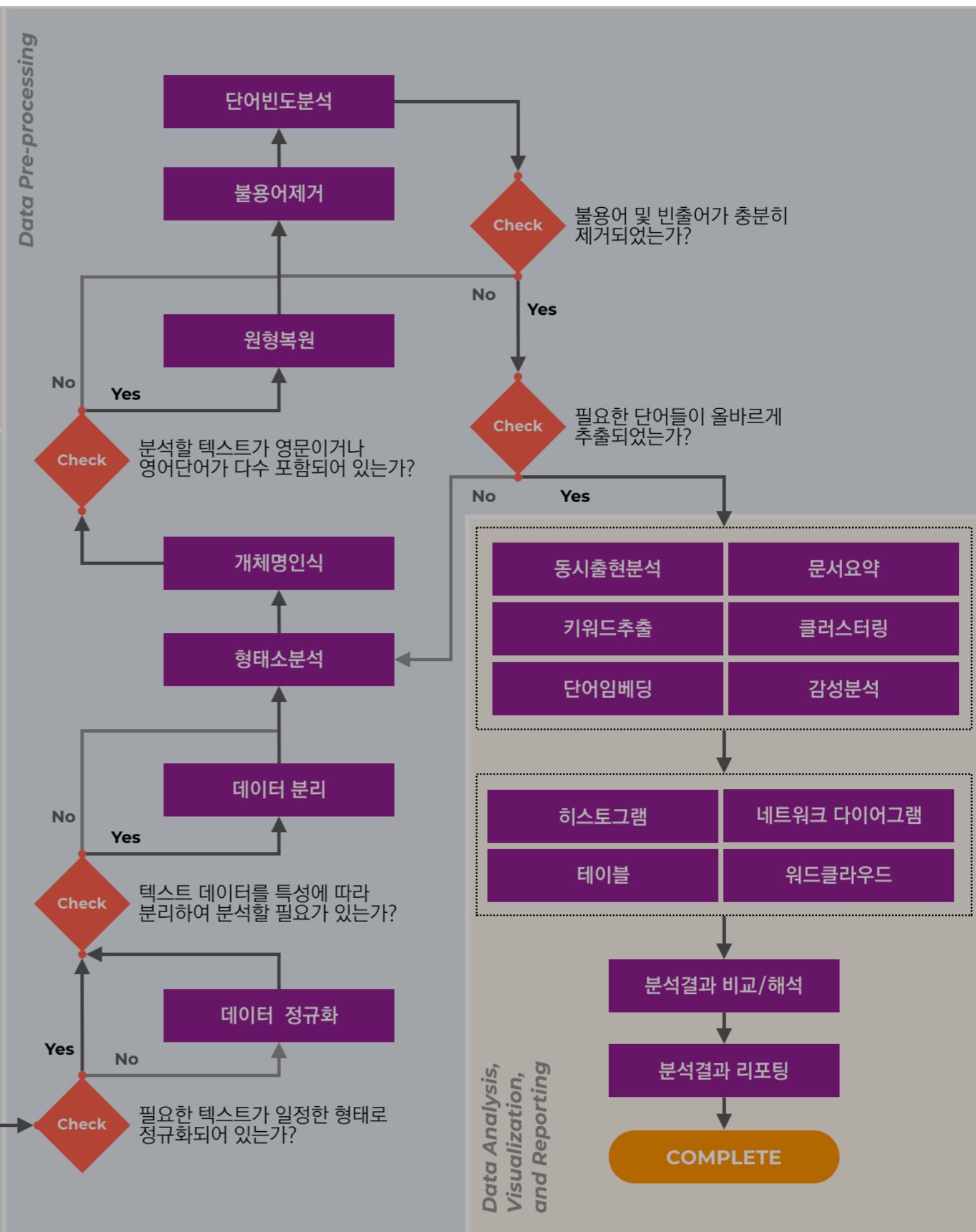
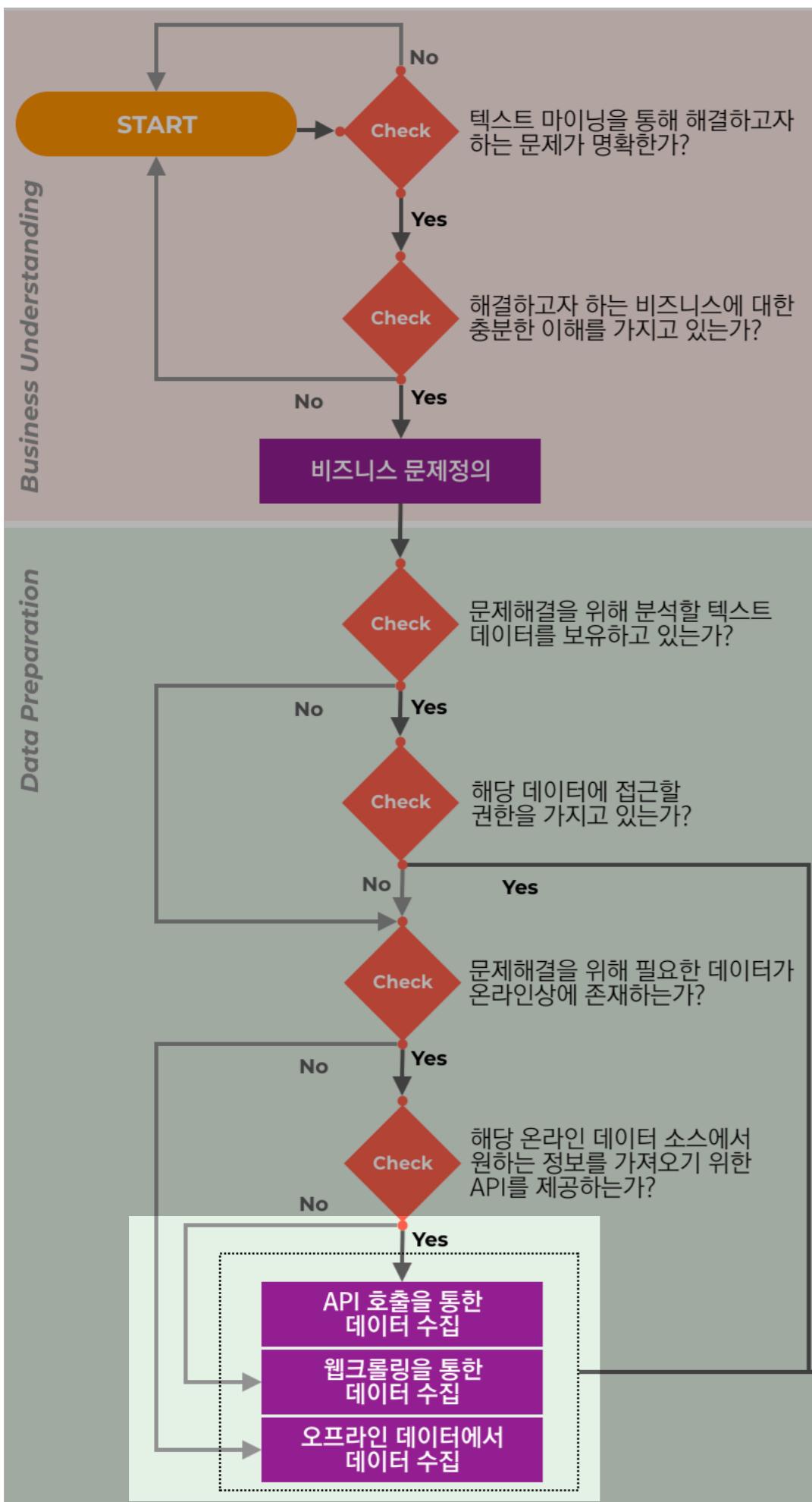


Business Understanding



Data Preparation





텍스트 데이터 수집

텍스트 데이터 수집 유형

- ▶ 수집대상 : 웹페이지, SNS, 댓글, 음성, 비디오 등 텍스트 형태로 변환 가능한 모든 데이터
- ▶ 저장유형 : Plain Text, PDF, Table, XML, JSON
- ▶ 수집방법 : Web Crawling, API 호출, DB Query, Online Survey

유형	수집방법	장/단점
오프라인 데이터	<ul style="list-style-type: none">온/오프라인 설문지음성 녹음, 비디오 촬영	<ul style="list-style-type: none">타겟팅 대상에 대한 데이터 수집 가능사람이 직접 수집해야함수집에 시간적, 공간적 제약이 큼
자체 시스템	<ul style="list-style-type: none">서비스 또는 사내 데이터베이스 활용사내 게시판, 유저 댓글, 업무 보고서, 내부문서	<ul style="list-style-type: none">기 수집된 데이터를 빠르게 활용 가능소속기관/업체/서비스 내부 관련자가 아니면 접근이 어려움정보 유출에 대한 위험이 큼
웹크롤링	<ul style="list-style-type: none">프로그래밍 언어를 활용해 웹페이지에 존재하는 대량의 정보를 반복 수집	<ul style="list-style-type: none">대량의 정보를 빠르게 수집할 수 있음데이터 수집과 함께 정규화된 데이터셋 구성 가능프로그래밍 언어 활용이 필요함개인정보 문제에 취약함
API 호출	<ul style="list-style-type: none">프로그래밍 언어를 활용해 서비스에서 정식으로 제공하는 데이터를 수집네이버 API, 카카오 API, Reddit API, News API, SNS (Twitter, Facebook, Instagram)	<ul style="list-style-type: none">바로 활용할 수 있는 양질의 데이터를 얻을 수 있음수집할 수 있는 소스가 제한적임프로그래밍 언어 활용이 필요함

텍스트 데이터 수집

텍스트 데이터를 제공하는 API

서비스 유형	서비스	제공 유형	비고
SNS	Facebook	<ul style="list-style-type: none">Graph APIPost, Comments	<ul style="list-style-type: none">https://developers.facebook.com 참고
	Instagram	<ul style="list-style-type: none">Media, Comments, Tags	<ul style="list-style-type: none">https://www.instagram.com/developer 참고
	Twitter	<ul style="list-style-type: none">Search APIStreaming API	<ul style="list-style-type: none">https://developer.twitter.com 참고
포털	네이버	<ul style="list-style-type: none">블로그 API뉴스 API백과사전 API웹문서 API	<ul style="list-style-type: none">https://developers.naver.com 참고본문만 제공
	다음(카카오)	<ul style="list-style-type: none">웹문서 검색 API블로그 검색 API카페 검색 API	<ul style="list-style-type: none">https://developers.kakao.com 참고본문만 제공
커뮤니티	Reddit	<ul style="list-style-type: none">Thread, Comment	<ul style="list-style-type: none">https://www.reddit.com/dev/api 참고
뉴스	News API	<ul style="list-style-type: none">Article Summery	<ul style="list-style-type: none">https://newsapi.org 참고
	American Broadcasting Company	<ul style="list-style-type: none">Resources API	<ul style="list-style-type: none">Resources API retrieves content produced by ABC businesses, including national and local news, entertainment videos, and more
	BBC	<ul style="list-style-type: none">Platform API	<ul style="list-style-type: none">Platform API power all the BBC's product areas
	New York Times	<ul style="list-style-type: none">Article Search APICommunity APIMovie Reviews APITimesTags APITop Stories API	<ul style="list-style-type: none">Headlines, abstracts and links to associated multimediaComments by NYTimes.com usersLinks to reviews and NYT Critics' Picks, and search movie reviewsTerms that match search queries, and filters by Times dictionariesLists of home page articles and associated images

텍스트 데이터 수집

타겟팅 유저 텍스트 수집을 위한 소스

소스	유형	플랫폼	주요토픽	회원수	사용연령대	성별
디시인사이드	커뮤니티	자체플랫폼	공통		10~30	공통
루리웹	커뮤니티	자체플랫폼	공통		20~30	공통
뽐뿌	커뮤니티	자체플랫폼	공통		10~30	공통
일베저장소	커뮤니티	자체플랫폼	공통		10~40	공통
스레딕	커뮤니티	자체플랫폼	공통		20~30	여성
도탁스	카페	다음	공통	511,049	-	공통
이토랜드	토렌트	자체플랫폼	공통		-	공통
네이트판	커뮤니티	자체플랫폼	고민, 이슈		10~30	공통
오늘의유머	커뮤니티	자체플랫폼	유머		10~30	공통
웃긴대학	커뮤니티	자체플랫폼	유머		10~30	공통
엽기혹은진실	카페	다음	유머	247,754	-	공통
유머나라	카페	다음	유머	114,626	-	공통
와이고수	커뮤니티	자체플랫폼	유머, 스포츠, 게임		10~40	남성
쭉빵카페	카페	다음	연예, 뷰티	1,731,956	20~30	여성
뉴빵카페	카페	다음	연예, 뷰티	1,101,596	20~30	여성
여성시대	카페	다음	연예, 뷰티	729,142	20~30	여성
파우더룸	카페	네이버	뷰티	1,856,696	20~30	여성
인스티즈	커뮤니티	커뮤니티	연예, 오락		10~30	여성
theqoo	커뮤니티	자체플랫폼	연예		10~20	여성
해연갤	커뮤니티	자체플랫폼	해외 연예		20~30	여성
가생이	커뮤니티	자체플랫폼	연예, 한류		20~40	-
베스티즈	커뮤니티	자체플랫폼	연예		20~30	여성
디젤매니아	카페	네이버	패션	882,132	20~30	남성
외방커뮤니티	커뮤니티	자체플랫폼	미용, 패션		20~30	여성

텍스트 데이터 수집

타겟팅 유저 텍스트 수집을 위한 소스

소스	유형	플랫폼	주요토픽	회원수	사용연령대	성별
레몬테라스	카페	네이버	육아, 인테리어	3,020,341	30~40	여성
맘스홀릭 베이비	카페	네이버	육아	2,684,457	20~30	여성
개드립	커뮤니티	자체플랫폼	유머, 게임		10~20	공통
인벤	커뮤니티	자체플랫폼	게임		10~20	남성
에펨코리아	커뮤니티	자체플랫폼	축구		10~40	남성
아이러브사커	카페	다음	축구	167,706	10~40	남성
MLB파크	커뮤니티	커뮤니티	야구		20~40	남성
이종격투기	카페	다음	격투기	1,023,757	10~30	남성
클리앙	커뮤니티	자체플랫폼	테크, 통신, 앱		20~40	남성
쿨엔조이	커뮤니티	자체플랫폼	테크, 하드웨어		20~40	남성
Seeko	커뮤니티	자체플랫폼	전자기기		30~40	남성
아사모 - 애플	카페	네이버	애플 아이폰	1,635,061	-	공통
중고나라	카페	네이버	중고거래	16,477,444	10~50	공통
중고카페 그린유즈	카페	네이버	중고거래	2,543,783	-	공통
보배드림	커뮤니티, 쇼핑몰	자체플랫폼	중고거래		30~50	공통
취업뽀개기	카페	다음	취업, 학생	1,399,394	20~30	공통
독취사 - 취업	카페	네이버	취업	2,393,699	20~30	공통
오르비	커뮤니티	자체플랫폼	수험생, 입시		10~20	공통
수만휘	카페	네이버	수험생	2,515,951	10~20	공통
82쿡	커뮤니티, 쇼핑몰		요리		20~50	여성
SLR클럽	커뮤니티	자체플랫폼	사진		30~50	공통
유랑 - 유럽여행	카페	네이버	여행	1,880,443	20~30	공통
네일동 - 일본여행	카페	네이버	여행	1,205,047	20~30	공통

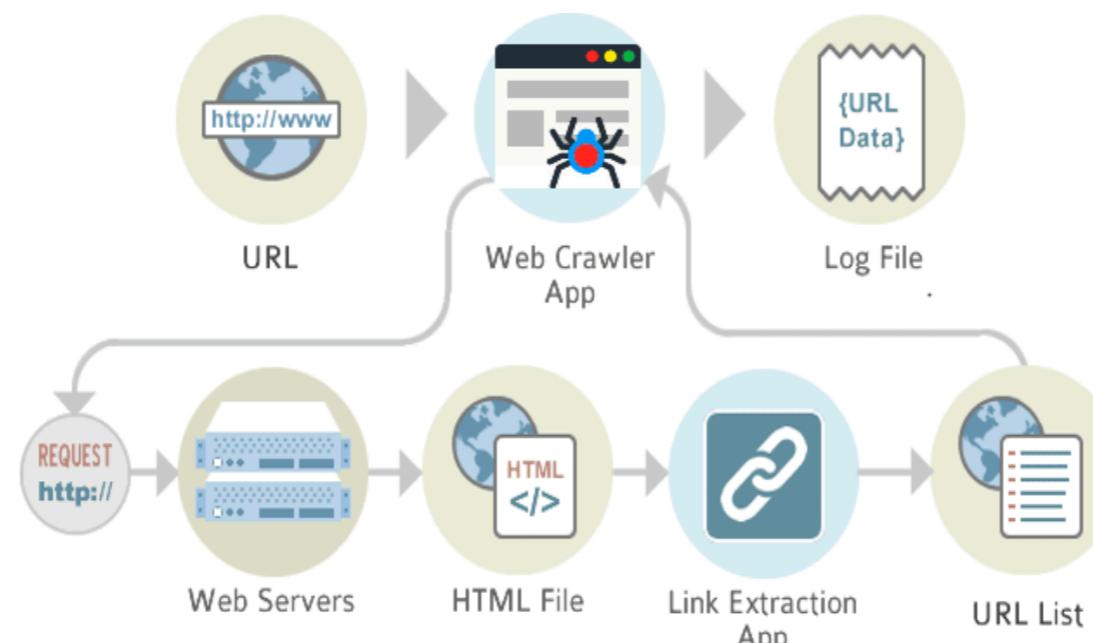
텍스트 데이터 수집

Crawling 이란?

- ▶ 정제되지 않은 웹페이지에서 필요한 데이터를 추출하는 행위
- ▶ 활용할 수 있는 데이터가 정리되어 올라가 있는 데이터(API, 파일형태)를 제외하고, 웹페이지에 게시된 자료를 가져오는 기술
- ▶ API를 통한 데이터 제공이 활성화 되고 있는 추세이나, 국내에서는 소극적인 API 제공으로 웹크롤링을 통한 수집이 반드시 필요함

Data Crawling & Scrapping 차이점

- ▶ Data Crawling : 웹페이지의 정보를 추출하는 방법(=spider, bot), 단순히 하이퍼링크를 돌아다니며 웹페이지를 다운로드
- ▶ Data Scrapping : 웹페이지에서 필요한 정보만을 추출하는 방법, 다운로드한 웹페이지에서 필요한 부분만을 추출하고 저장
- ▶ 단순히 데이터 수집을 넘어, 데이터를 수집하고 활용하기 위한 사전작업



*Source : Rodrigo, I Will Code A Small Web Crawler, <https://www.fiverr.com/rodrigo/code-a-small-web-crawler/>.

크롤링의 도덕적 문제

데이터 무단수집과 저작권 침해

- ▶ 웹크롤링은 원래 검색엔진 등의 인터넷 사이트에서 데이터를 최신 상태로 유지하기 위해 사용
- ▶ 웹크롤링을 활용하여 타사 컨텐츠를 무단 활용하는 것은 불법행위에 해당함
- ▶ 웹크롤링을 통한 과도한 요청은 대상 서비스 서버 운영과 서비스 관리에 안 좋은 영향을 끼침
- ▶ 경쟁사간의 상도덕 문제 또는 개인 양심상의 문제

채용정보 무단복제 '사람인HR', 잡코리아에 120억 지급

양보다 질 중요한 취업포탈 업계…접근 쉬운 채용공고 속성 악용한 편취사례

이준영 기자 | 승인 2018.02.09 12:31 | 댓글 0

댓글부대 의혹 야놀자, 무단 DB 크롤링 의혹 여기어때

숙박 O2O 시장 논란 언제까지

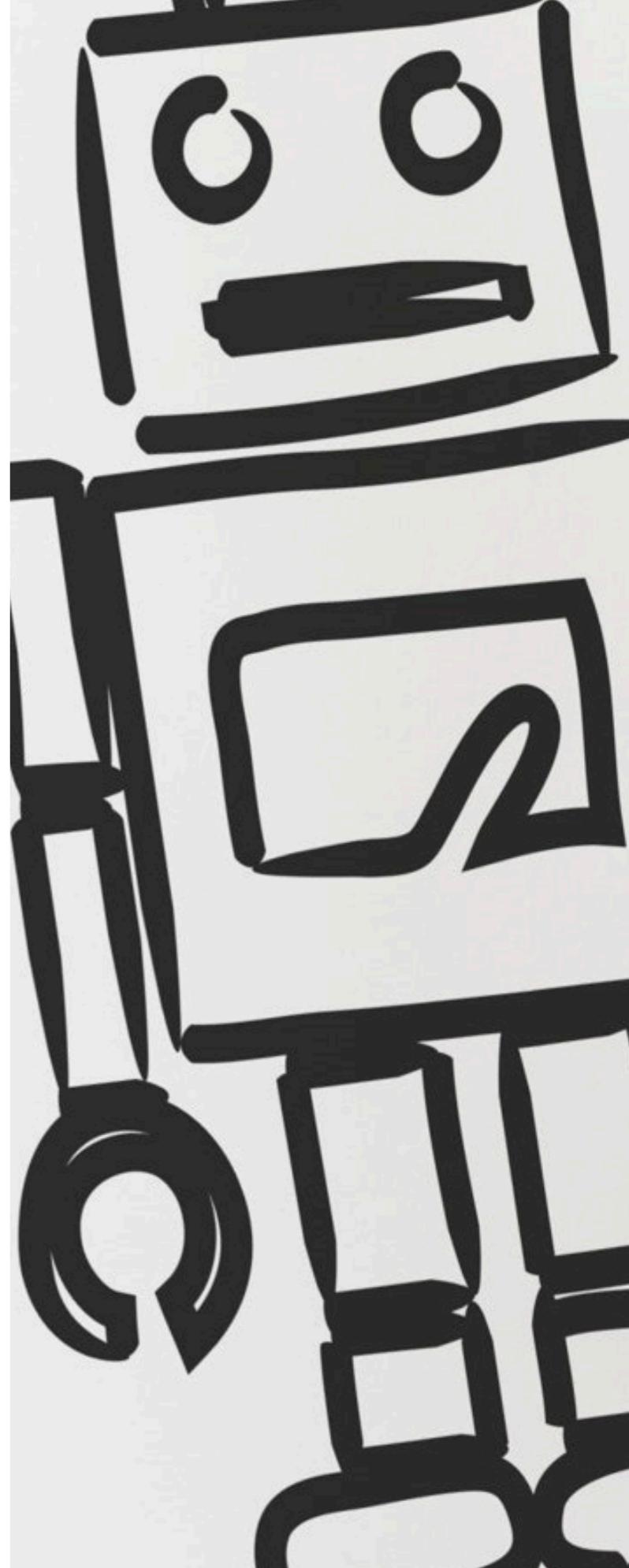
최진홍 기자 | rgdsz@econovill.com | 승인 2017.11.03 16:23:50



사진=각 사

제 및 게재 행위를 하지 않고 공정한 경쟁질서

모바일 시대가 도래하며 O2O 스타트업의 존재감이 날카로워지고 있지만 잡음 또한 높아지고 있다. 이들은 온라인 경쟁력을 키우면서도 오프라인 거점도 확보, 이를 통한 다양한 파생 서비스에 나선다는 목표도 세워놓고 있다. 그러나 숙박 O2O 업체들이 용인할 수 있는 수준을 넘어설 정도로 구설에 오르고 있는 것은 여간 심각한 문제가 아니다. 최근 나름 적절한 수위를 찾아간다는 평가가 나오고 있지만 배달의민족, 요기요, 배달통 등이 포진한 배달앱 업계도 마찬가지고 다방과 직방 등 부동산 O2O 시장도 사정이 비슷하다. 그 중에서 숙박 O2O 시장을 둘러싼 논란은 상상 이상이다.



*Source : 이준영(시장경제), 채용정보 무단복제 '사람인HR', 잡코리아에 120억 지급, 2018.2.9., <http://www.mconomynews.com/news/articleView.html?idxno=11088/>.

**Source : 최진홍(이코노미리뷰), 댓글부대 의혹 야놀자, 무단 DB 크롤링 의혹 여기어때, 2017.11.3., <http://www.econovill.com/news/articleView.html?idxno=325820/>.

크롤링의 도덕적 문제

Robots.txt

- ▶ 웹사이트에 배치된 텍스트 파일로, 크롤링 접근권한에 대해 명시해 놓은 문서
- ▶ 웹크롤링은 Robots.txt 파일에서 허용하는 항목에 대해서만 수집 가능하며 그 외의 수집에 대한 책임은 모두 본인에게 있음
- ▶ 수집이 허용되 있더라도 대상 서비스 운영에 피해를 주지 않는 선에서 필요한 만큼만 수집
- ▶ Robots.txt 파일이 없는 경우 서비스 관리자에 직접 허락을 구한 후 수집

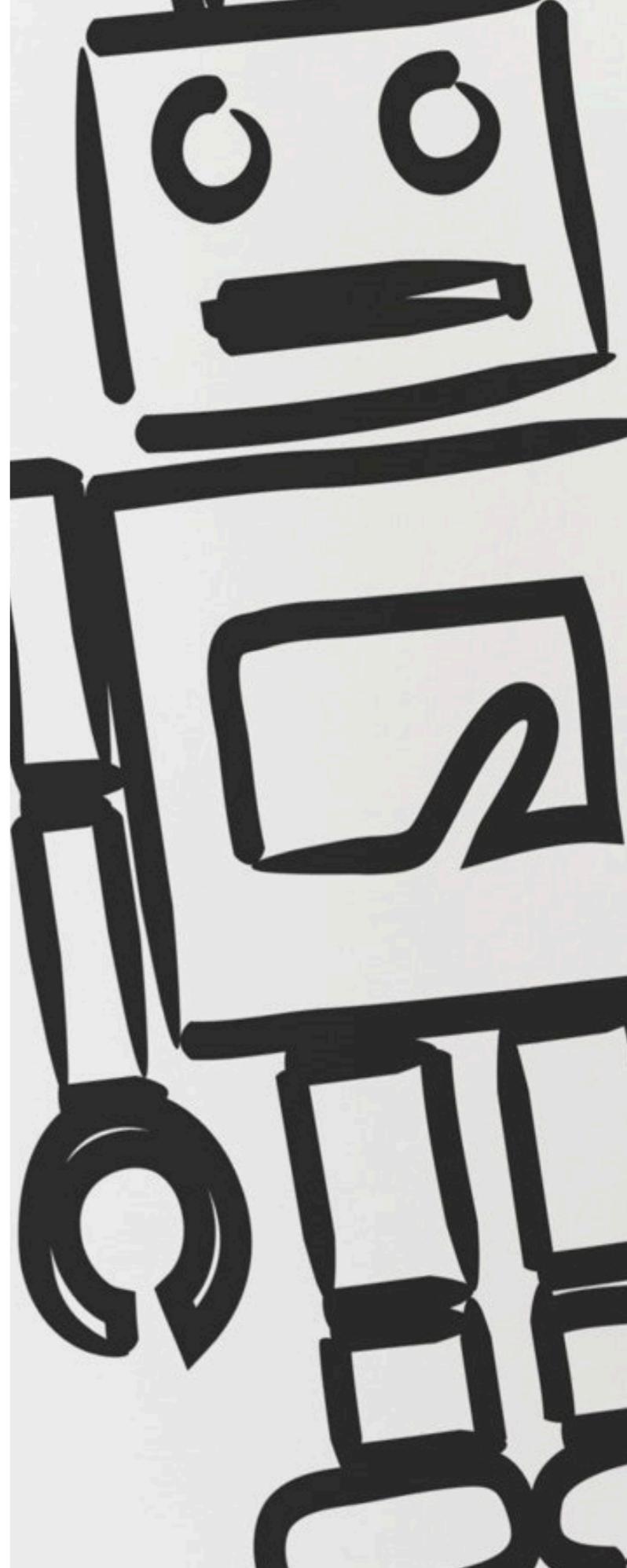


```
User-agent: Bingbot
Allow: /ajax/pagelet/generic.php/PagePostsSectionPagelet
Allow: /safetycheck/
```

```
User-agent: Googlebot
Allow: /ajax/pagelet/generic.php/PagePostsSectionPagelet
Allow: /safetycheck/
```



```
User-agent: *
Allow:/service/board/
Disallow:/service/group/
Disallow:/service/board/sold/
Disallow:/service/mypage/
Disallow:/service/message/
Disallow:/service/popup/
Disallow:/service/search/
Disallow:/service/cs/
```



Python 웹크롤링

Python을 활용하는 이유

- ▶ 가장 배우기 쉽고 읽기 쉬는 프로그래밍 언어
- ▶ 풍부한 외부 라이브러리
- ▶ 데이터 수집 뿐만아니라 데이터 전처리, 분석과의 친화성

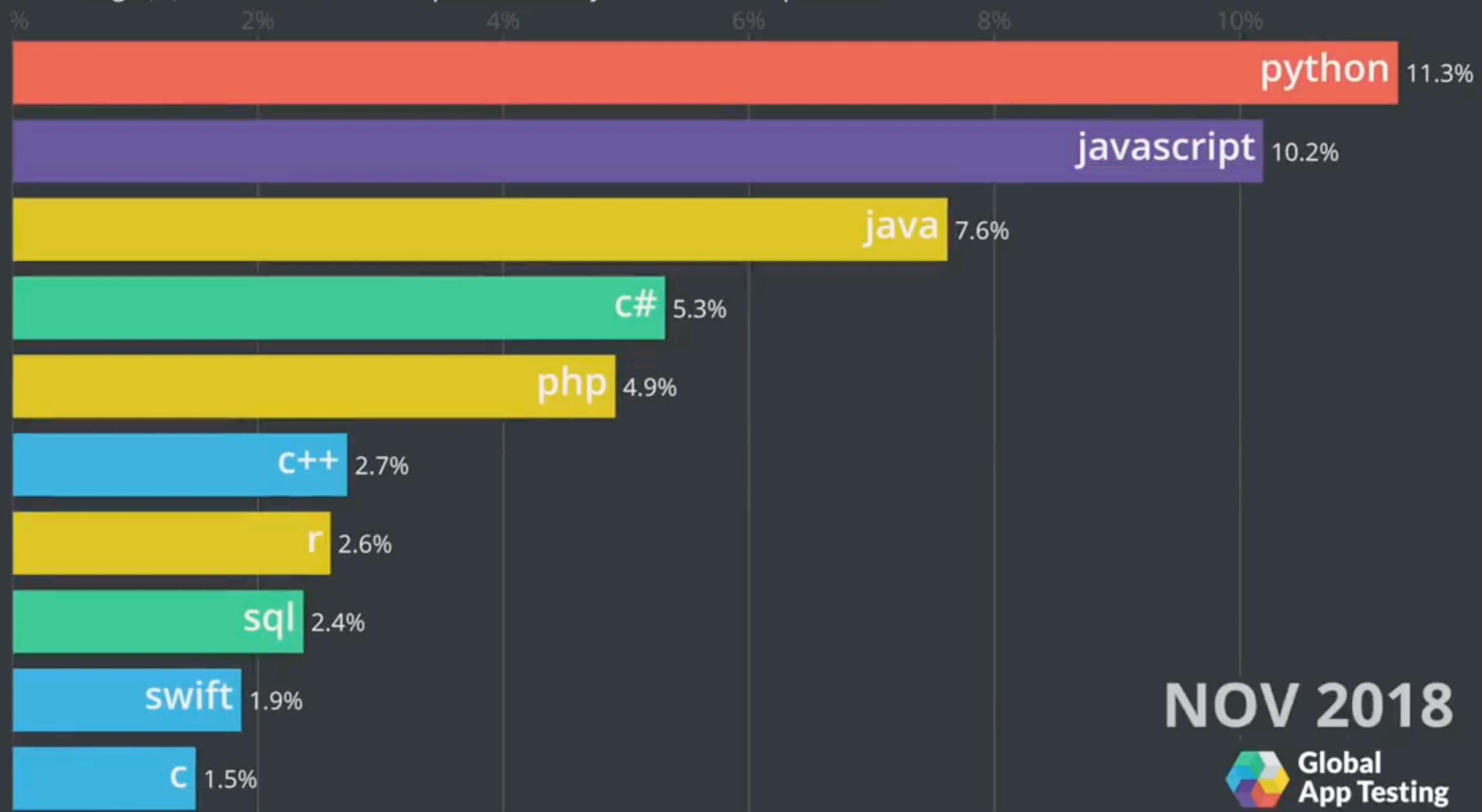
크롤링하는 방법 선택

구분	라이브러리	역할 및 장점/단점
브라우저 운영	Requests (urllib)	장점 : Python에서 동작하는 가장의 작고 빠른 브라우저 거의 모든 플랫폼에서 구동 가능 단점 : 웹서버로부터 초기 HTML만 받아 추가적인 CSS/JavaScript 처리결과가 반영되지 않음
	Selenium	장점 : 웹브라우저를 통합된 방식으로 원격 컨트롤하는 라이브러리 (Chrome, Firefox, IE, PhantomJS 등) 단점 : 직접 브라우저를 사용하여 추가로 CSS/JavaScript 처리를 위한 리소스가 많이 필요함
파서 (Parser)	BeautifulSoup4	HTML로부터 원하는 위치/형식의 문자열을 획득



Most popular programming languages on Stack Overflow.

Percentage (%) of all Stack Overflow questions every month since September 2008.



E.O.D