

05

비정형데이터분석

# 텍스트 데이터 불러오기(1)

통계·데이터과학과 장영재 교수



KOREA NATIONAL OPEN UNIVERSITY

# 학습목차

- 1 비정형 데이터의 현황
- 2 텍스트 데이터의 이해
- 3 텍스트 데이터의 수집 방법
- 4 실습 : 데이터저장소 방문



01

# 비정형 데이터의 현황



## 1. 비정형 데이터의 현황

- 빅데이터 시대에 접어들어 전체 데이터 중 비정형 데이터 비중이 급증
  - 전체 데이터의 80% 이상을 차지(시장 조사 기관 IDC(International Data Corporation))
  - 디지털 데이터\*의 경우로 국한한다면 비정형 데이터가 차지하는 비중이 90%에 달함
    - ▷ 디지털 데이터란 전자적 방법으로 저장되거나 네트워크 및 유선, 무선 통신 등을 통해 전송되는 정보
  - IT 기술의 지속적인 발전은 디지털 데이터의 생산도 크게 늘리면서 소셜 데이터와 같은 비정형 데이터의 증가세도 더욱 가속화시킬 것으로 판단  
→ 트위터나 페이스북, 카카오톡 등 메신저, 인스타그램이나 유튜브 등의 사례



02

## 텍스트 데이터의 이해



## 2. 텍스트 데이터의 이해

- 비정형데이터분석은 텍스트와 이미지 분석이 주류를 이루고 있음
  - 비정형 데이터 중 가장 많은 비중을 차지하고 있는 텍스트 데이터 분석에 초점을 맞추어 개괄





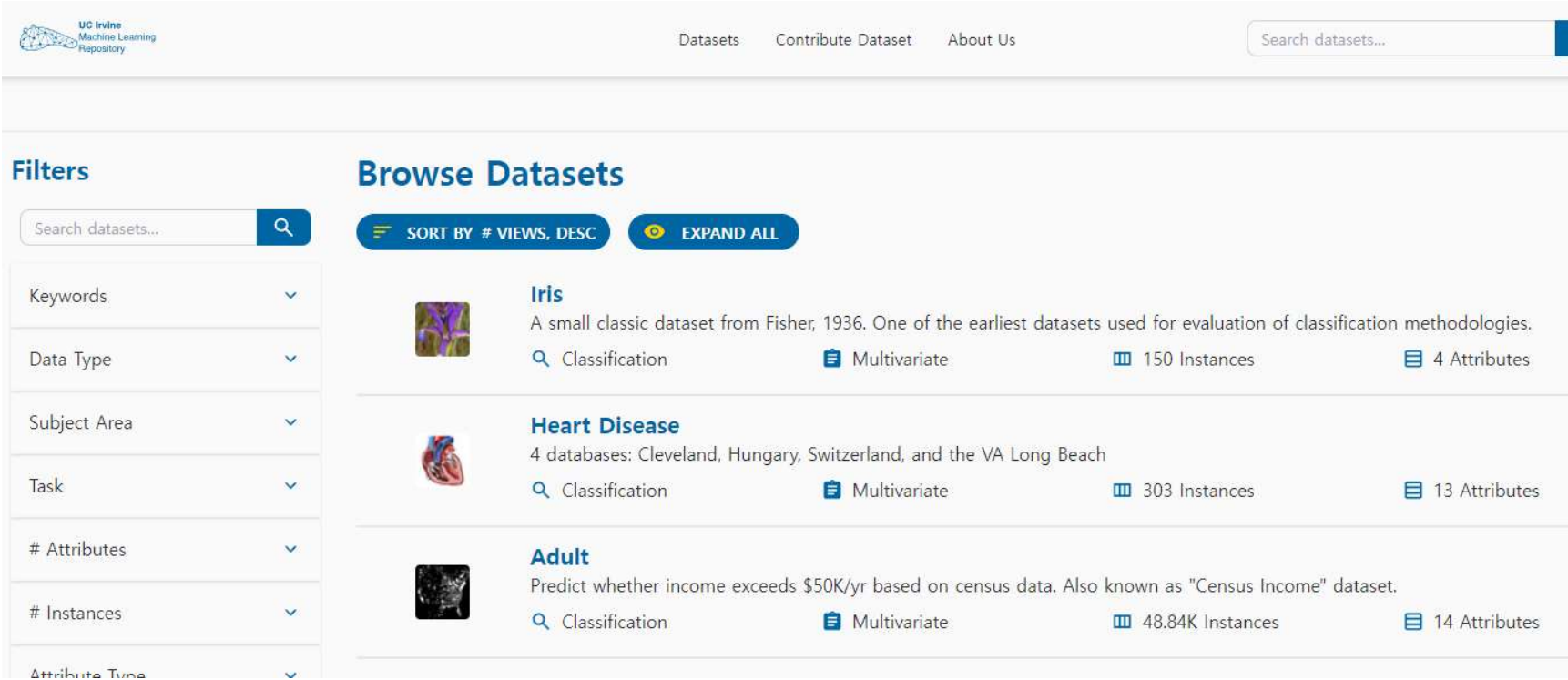
## 2. 텍스트 데이터의 이해 - ① 텍스트 데이터의 사례

- UCI machine learning repository는 연구나 교육 목적의 데이터 집합소

- 첫 화면(<https://archive.ics.uci.edu/datasets>)\*에는 데이터의 이름, 분석 형태, 분석 목적, 변수들의 형태, 관측치 수, 변수의 개수, 생성연도 등이 나타나 있음

\* 주기적인 업데이트 가능성이 있음에 유의





UC Irvine Machine Learning Repository

Datasets Contribute Dataset About Us

Search datasets...




### Filters

Search datasets... 🔍

- Keywords ▾
- Data Type ▾
- Subject Area ▾
- Task ▾
- # Attributes ▾
- # Instances ▾
- Attribute Type ▾

### Browse Datasets

☰ SORT BY # VIEWS, DESC 🔍 EXPAND ALL

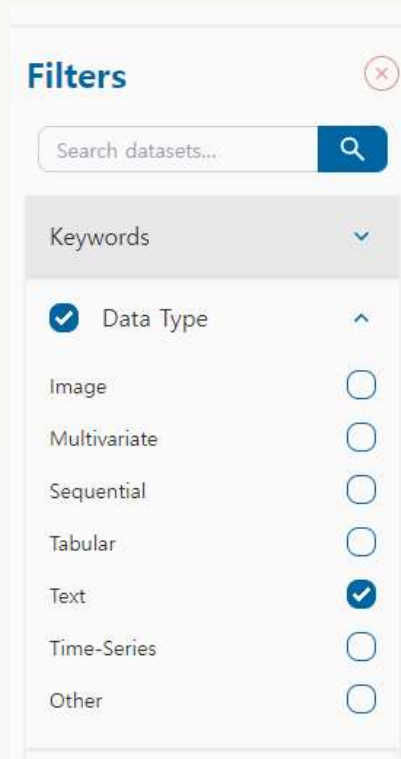
Dataset	Description	Task	Multivariate	Instances	Attributes
 <b>Iris</b>	A small classic dataset from Fisher, 1936. One of the earliest datasets used for evaluation of classification methodologies.	🔍 Classification	📋 Multivariate	📊 150 Instances	📋 4 Attributes
 <b>Heart Disease</b>	4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach	🔍 Classification	📋 Multivariate	📊 303 Instances	📋 13 Attributes
 <b>Adult</b>	Predict whether income exceeds \$50K/yr based on census data. Also known as "Census Income" dataset.	🔍 Classification	📋 Multivariate	📊 48.84K Instances	📋 14 Attributes

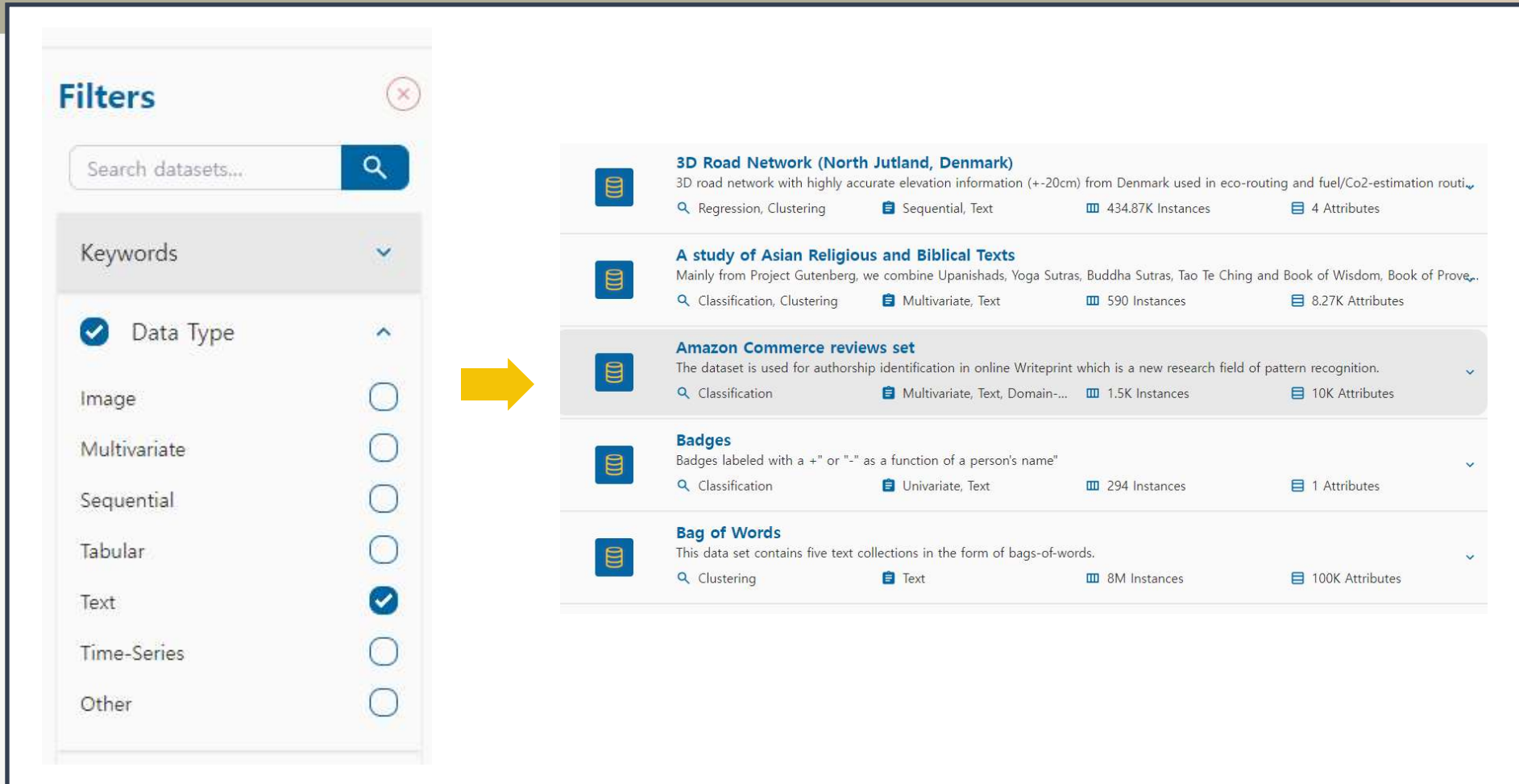
<그림> UCI machine learning repository



## 2. 텍스트 데이터의 이해 - ① 텍스트 데이터의 사례

- 데이터 저장소 첫 화면에서 좌측에 위치한 'Filters' 섹션을 찾아서 'Data Type' 중 'Text' 부분을 클릭





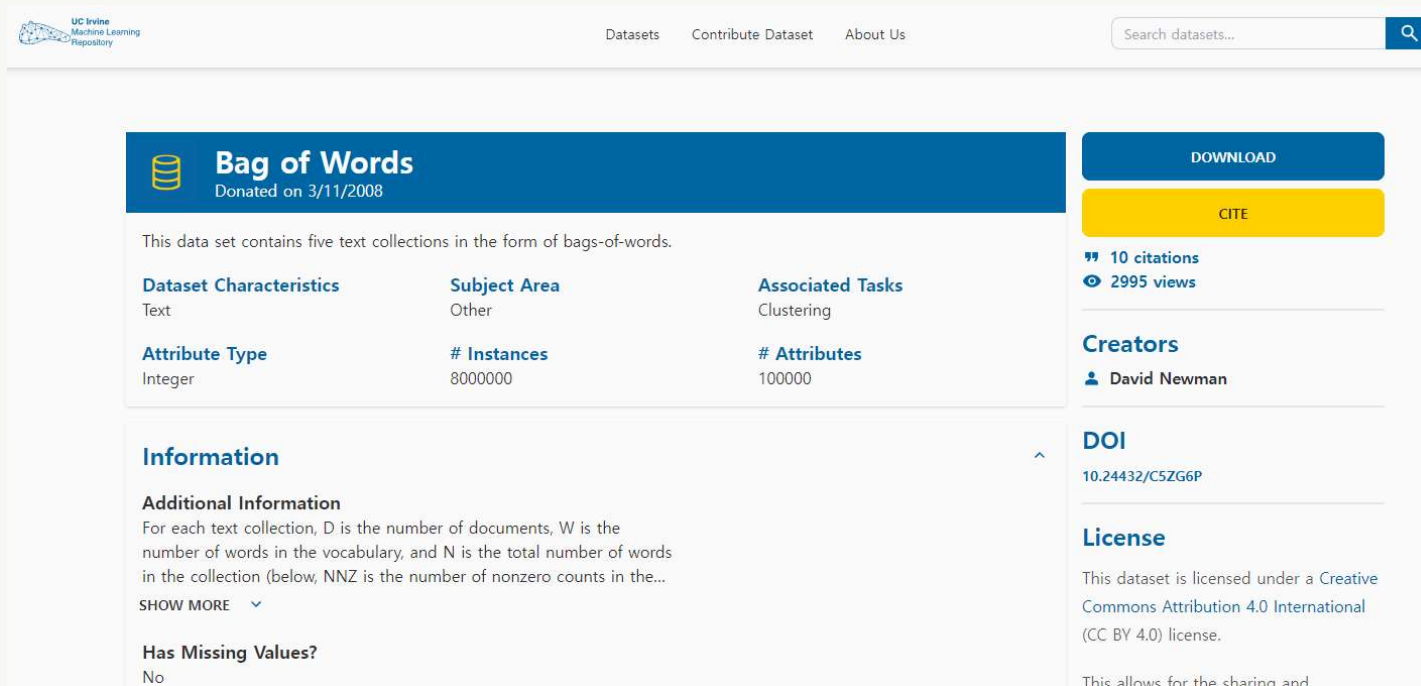
The screenshot shows a web interface for filtering datasets. On the left, a 'Filters' sidebar has a search bar and a 'Keywords' section. Under 'Data Type', the 'Text' option is selected with a blue checkmark. A yellow arrow points from this filter to a list of datasets on the right. The datasets listed are:

Dataset Name	Description	Search	Instances	Attributes
3D Road Network (North Jutland, Denmark)	3D road network with highly accurate elevation information (+/-20cm) from Denmark used in eco-routing and fuel/Co2-estimation routi...	Regression, Clustering	434.87K	4
A study of Asian Religious and Biblical Texts	Mainly from Project Gutenberg, we combine Upanishads, Yoga Sutras, Buddha Sutras, Tao Te Ching and Book of Wisdom, Book of Prove...	Classification, Clustering	590	8.27K
Amazon Commerce reviews set	The dataset is used for authorship identification in online Writprint which is a new research field of pattern recognition.	Classification	1.5K	10K
Badges	Badges labeled with a "+" or "-" as a function of a person's name"	Classification	294	1
Bag of Words	This data set contains five text collections in the form of bags-of-words.	Clustering	8M	100K

<그림> 텍스트 데이터의 조회

## 2. 텍스트 데이터의 이해 - ① 텍스트 데이터의 사례

- 아래 그림은 저장소의 텍스트 데이터 중 'Bag of Words'를 선택하였을 때 나타나는 화면



The screenshot shows the UC Irvine Machine Learning Repository page for the 'Bag of Words' dataset. The page includes a search bar, navigation links, and detailed dataset information.

**UC Irvine Machine Learning Repository**

Datasets   Contribute Dataset   About Us   Search datasets...

**Bag of Words**  
Donated on 3/11/2008

This data set contains five text collections in the form of bags-of-words.

Dataset Characteristics	Subject Area	Associated Tasks
Text	Other	Clustering

Attribute Type	# Instances	# Attributes
Integer	8000000	100000

**Information**

**Additional Information**  
For each text collection, D is the number of documents, W is the number of words in the vocabulary, and N is the total number of words in the collection (below, NNZ is the number of nonzero counts in the...)

SHOW MORE

**Has Missing Values?**  
No

**DOWNLOAD**

**CITE**

10 citations  
2995 views

**Creators**  
David Newman

**DOI**  
10.24432/C5ZG6P

**License**  
This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.  
This allows for the sharing and

## 2. 텍스트 데이터의 이해 - ① 텍스트 데이터의 사례

- 'Attribute Characteristics'가 'Integer', 즉 정수\*라고 되어있다는 점에 유의
  - ▷ 다운로드 받은 압축 파일을 풀면 출현 단어 목록 파일(vocab\*.txt)과 단어 빈도 등의 정보가 수록된 bag of words(docword\*.txt) 파일이 나타남
  - ▷ 이 파일의 관측값 속성이 정수이므로 'integer'로 표시



## 2. 텍스트 데이터의 이해 - ② 텍스트 데이터의 특징

- 텍스트 데이터는 전통적 문헌자료, 정기간행물 및 전문 학술문서, 소셜데이터 등을 포함
  - 텍스트 데이터는 전통적인 의미에서 문헌자료를 지칭하는데 IT 기술로 디지털화(digitalization)되면서 양적인 면에서 급격히 증가
  - 주기적으로 발간되는 신문이나 잡지, 연구자료, 보고서 등도 텍스트 데이터에 속함



## 2. 텍스트 데이터의 이해 - ② 텍스트 데이터의 특징

- 소셜네트워크서비스(Social Network Service)사용의 확대로 텍스트 데이터가 급증
  - 빅데이터의 특징이 가장 뚜렷하게 나타남
  - 커뮤니티의 성장과 함께 기하급수적으로 증가하면서 다양한 형태로 생성







### 텍스트 데이터의 특징

- 전형적인 문서자료+전자자료  
다량의 정성적 자료
- 비정형(Unstructured)  
또는 정형(Structured)
- 실시간(real time) 생성가능  
⇒ 계량분석방법 보완  
객관적 분석

### 빅데이터의 요소

- Volume
- Variety
- Velocity

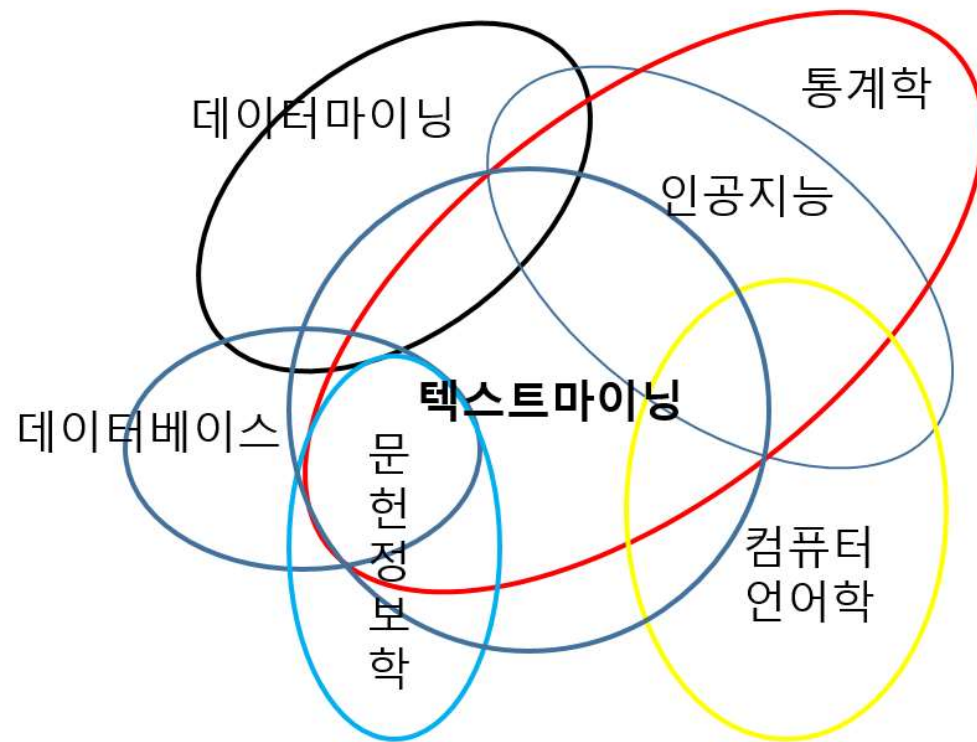
<그림> 텍스트 데이터의 특징



## 2. 텍스트 데이터의 이해 - ② 텍스트 데이터의 특징

- 텍스트 데이터 분석은 텍스트 데이터의 특징을 고려한 접근이 필요
  - 다양한 인접 학문들의 종합적인 시각에서 이루어져야 함
  - 도구적인 면에서도 여러 분야의 기법들이 적절히 조화를 이루어 적용되어야 함





<그림> 텍스트 마이닝 관련 분야



03

## 텍스트 데이터의 수집 방법



### 3. 텍스트 데이터의 수집 방법 - ① 데이터 저장소를 통한 텍스트 데이터 수집

- 텍스트 분석을 위한 데이터를 수집하기 위해서는 기술적인 준비가 필요
  - 가장 손쉬운 방법은 데이터 저장소에서 데이터를 수집하는 것
    - 간단한 실습용 데이터를 쉽게 얻을 수 있는 방법

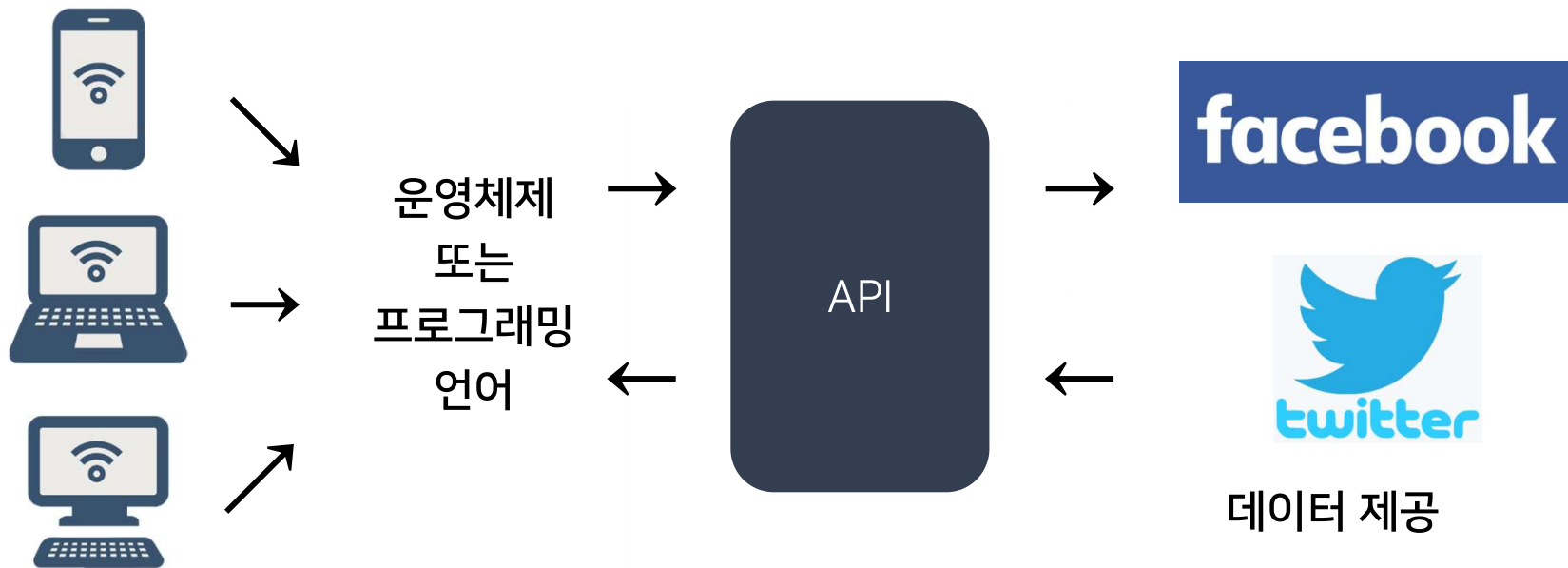


### 3. 텍스트 데이터의 수집 방법 - ② API를 통한 텍스트 데이터 수집

- API(Application Programming Interface)방식을 이용하면 소셜데이터의 수집이 가능
  - API는 운영체제나 프로그래밍 언어가 제공하는 기능을 제어하는 인터페이스를 의미(프로그램과 프로그램 간의 연결고리)
  - 데이터가 공개되어 있더라도 이 연결고리가 규격화되어 있지 않다면, 즉 호환성이 결여되어 있다면 연결고리로서의 의미도 없음
    - 데이터 제공을 위해 규격화된 인터페이스가 API임
    - API는 웹 기반 시스템, 운영 체제, 데이터베이스 시스템, 컴퓨터 하드웨어 또는 소프트웨어 라이브러리 등에 응용되고 있음







<그림> API의 동작 과정

### 3. 텍스트 데이터의 수집 방법 - ② API를 통한 텍스트 데이터 수집

- 소셜네트워크서비스업체를 비롯한 IT 관련 기업들은 API를 통하여 일부 데이터를 수집할 수 있도록 공개 API를 제공
  - 기본적인 검색의 기능부터 검색어 트렌드, 쇼핑 트렌드 등 추세 및 이미지나 지도에 관한 서비스도 제공
- 공공분야에서도 API를 통해 다양한 유형의 데이터를 수집할 수 있음



### 3. 텍스트 데이터의 수집 방법 - ③ 웹문서 데이터의 수집

- 웹문서는 웹 페이지 내에 나타나는 문서로서 적절한 검색의 규격에 맞도록 구조를 잘 갖추어야 함
  - 최대한 규격화된 구조를 잘 지키면서도 최적화를 통해 적절한 검색이 이루어지도록 작성
  - 최적화는 크롤링(crawling)과 인덱싱(indexing) 두 가지 측면을 고려
    - 크롤링 최적화란 문서에 접근하는 검색로봇이 해당 웹문서의 내용을 최대한 많이 긁어갈 수 있도록 하는 것을 의미
    - 인덱싱 최적화는 검색로봇에 의해 수집된 문서가 원하는 검색어에 대해 최대한 상위에 노출되도록 하는 것을 의미



### 3. 텍스트 데이터의 수집 방법 - ③ 웹문서 데이터의 수집

- 웹문서를 수집하는 방법으로 웹스크래핑과 웹크롤링을 고려
  - 웹스크래핑(web scraping)은 웹문서에서 데이터를 추출하는 기술
    - 화면에 표시되는 다양한 정보 중 사용자가 지정하거나 필요한 정보만을 추출하여 가공하고 저장하며 사용자에게 제공
  - 웹크롤링(web crawling)은 기초가 되는 URL(uniform resource locator) seed들을 저장한 뒤 웹페이지의 하이퍼링크를 인식하여 URL을 갱신하며 반복적으로 웹 링크(web link)를 찾는 과정
    - 웹페이지에 이르기 위해 웹링크를 따라가는 과정(following links to reach numerous pages)을 의미



### 3. 텍스트 데이터의 수집 방법 - ③ 웹문서 데이터의 수집

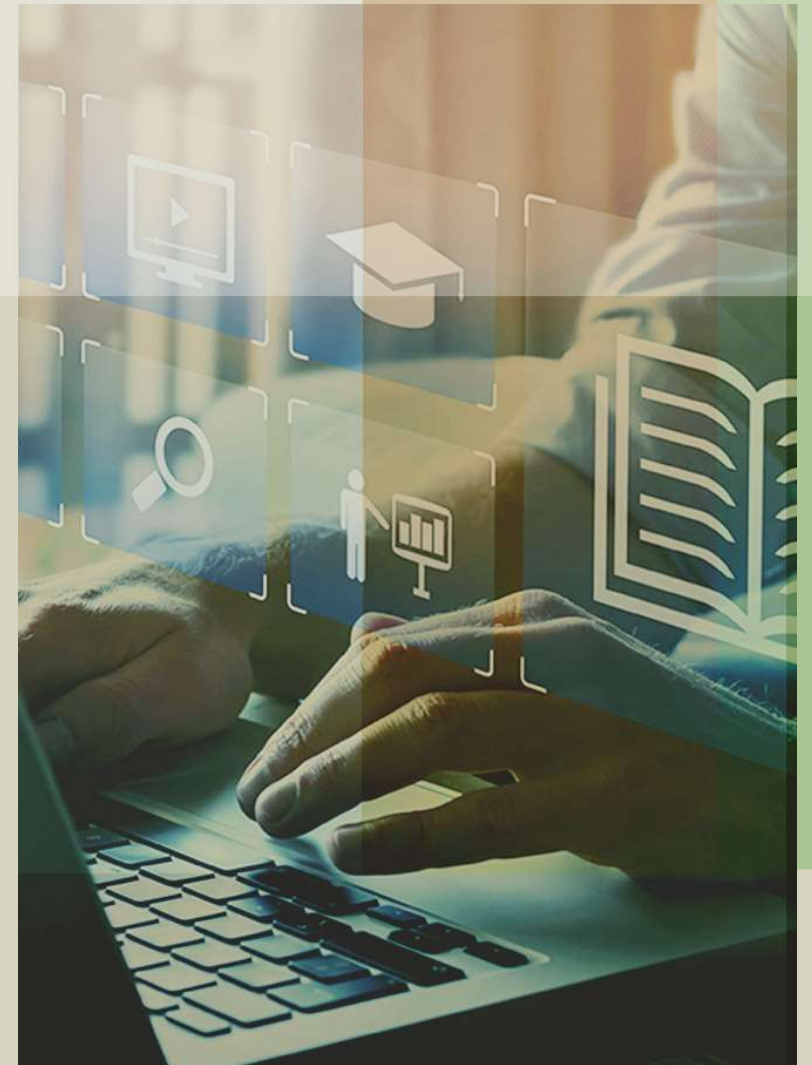
- 웹스크래핑이나 웹크롤링의 도구를 웹크롤러(web crawler)\*라고 하며 조직적, 자동화된 방법으로 월드 와이드 웹을 탐색하는 기능
  - \* 앤트(ants), 자동 인덱서(automatic indexers), 봇(bots), 웜(worms), 웹 스파이더(web spider), 웹 로봇(web robot) 등으로 불림



04

## 실습: 데이터 저장소 방문

(<https://archive.ics.uci.edu/datasets>)





다음시간안내

06

## 텍스트 데이터 불러오기(2)

