

데이터언어의 이해와 활용

경영정보학과

김 근 형

목 차

- 빅데이터의 출현배경, 특징, 유형, 분석절차
- 데이터 분석도구 유형과 특징
- 관련 자격증 소개
- 파이썬 활용 분석사례

데이터와 정보

◆ 데이터

- 관찰된 사실(**fact**)
- 예
 - ✓ 학생들의 시험점수

◆ 정보

- 의사결정에 사용되는 (가공)데이터
- 원데이터 vs. 가공데이터
 - ✓ 원데이터 예: 개인시험점수 => 용돈 상향조정 여부
 - ✓ 가공 데이터 예 : 시험점수 순위 => 장학생 결정

◆ 지식

- 정보에 대한 추가적인 데이터를 가공처리한 고급 정보
- 예:
 - ✓ 성적 우수학생은 수면시간이 적다

빅데이터 출현 배경(1)

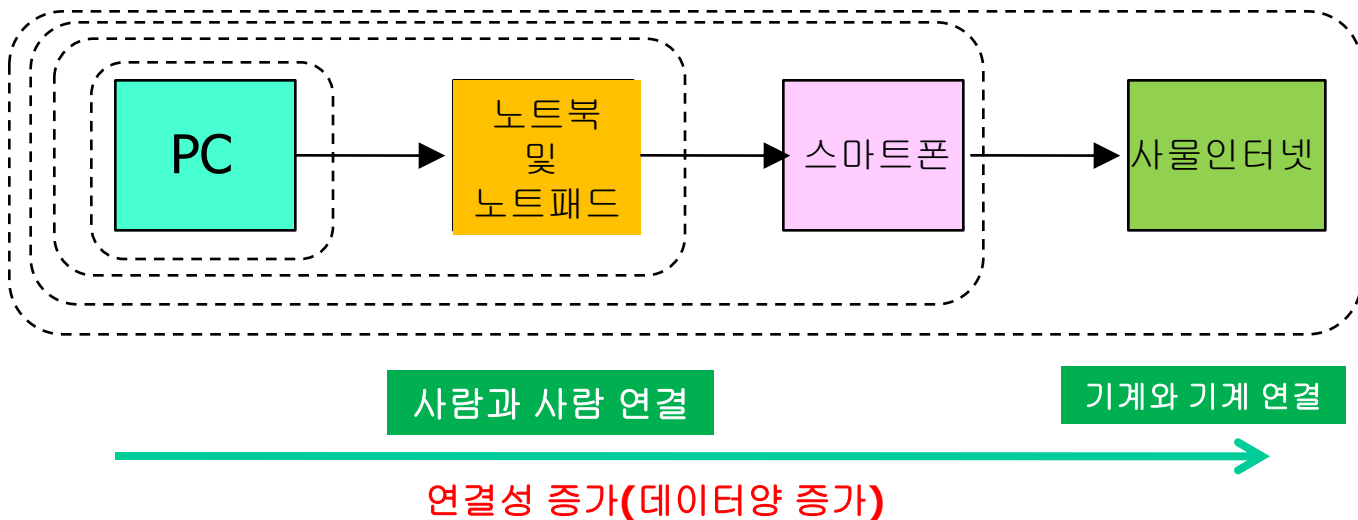
◆ 데이터 생산 및 저장비용의 하락

- 데이터 생산에 드는 수고와 비용이 줄어들음
- 저장매체 가격의 하락
- 필름카메라 **vs.** 디지털카메라

빅데이터 출현 배경(2)

◆ 전자적 연결성의 증가

- 사람과 사람 => 기계와 기계



빅데이터 출현 배경(3)

◆ 컴퓨팅 기술의 발전

- **CPU** 속도 증가
- 데이터 관리기술 발전



<90년대 그래픽처리 연구용 이미지>

빅데이터의 특징(1/2)

◆ 엄청난 데이터

- 대용량 + 처리가 어렵다
- 예:
 - 1억명의 고객 데이터
 - 한글 시의 문장 이해

빅데이터의 특징(2/2)

◆ 규모(Volume)

- 수 **TB** ~ 수**PB**
- 월마트의 거래데이터 : **2,500TB**

◆ 다양성(Variety)

- 구조적 데이터: 판매데이터 등
- 비구조적 데이터 : 뉴스기사, **SNS**게시물, 유튜브동영상 등

◆ 속도(Velocity)

- 수 분내에 **SNS**게시물 올라옴

데이터 유형(1/3)

엑셀화일-구조적(정형적)

번호	이름	국어	영어	수학
1	홍길동	90	97	88
2	심청이	95	98	90
3	준향이	85	88	77

데이터의 위치에 따라 의미 결정됨

텍스트문서-비구조적(비정형적)



데이터스트림즈, 中 빅데이터 시장 도전장
 ZDNet Korea | 13시간전 | 네이버뉴스 | [🔗](#)
 (지디넷코리아=김우용 기자) 데이터관리소프트웨어 전문업체 데이터스... 빅데이터 등 SW 기술을 소개해 좋은 반응을 얻었고, CEO가 관심을 갖고... ▲ 미영상 데이터스트림즈 대표 아울러 데이터스트림즈는 해외 시장...

▶ 데이터스트림즈, 메모리 기반의 빅데이터... ITWorld | 17시간전
 ▶ '빅데이터' 데이터스트림즈 中전출 매일경제 | 13시간전 | 네이버뉴스
 관련뉴스 전체보기 >

일선 시·군 '정보공개' 인식... 경기도 '빅데이터 시대' 엿박자
 경기일보 | 7시간전 | [🔗](#)
 빅데이터를 활용해 도민에게 각종 정보를 제공, 도정 참여 기회를 확대하겠다는 경기도와 달리 일선 시·군들이 정반대 행보를 보이고 있다. 정보 공개율이 저조하다는 지적을 사고 있기 때문이다. 23일 경기도에...

[뉴스 더보기 >](#)

데이터의 의미는 위치와 관계없음

데이터의 의미를 어떻게 파악할 것인가?

데이터 유형(2/3)

엑셀화일-구조적(정형적)

번호	이름	국어	영어	수학
1	홍길동	90	97	88
2	심청이	95	98	90
3	준향이	85	88	77

XML화일-반구조적

```
<?xml version="1.0" encoding="UTF-8" standalone="true"?>
- <dataGrid>
  - <fields>
    <field>기관</field>
    <field>경비지도사선임인원수</field>
    <field>일반</field>
    <field>기계</field>
  </fields>
  - <records>
    - <record>
      <기관>총계</기관>
      <기계>319</기계>
      <경비지도사선임인원수>5,558</경비지도사선임인원수>
      <일반>5,220</일반>
    </record>
    - <record>
      <기관>서울청</기관>
      <기계>190</기계>
      <경비지도사선임인원수>2,743</경비지도사선임인원수>
      <일반>2,543</일반>
    </record>
    - <record>
      <기관>부산청</기관>
      <기계>14</기계>
      <경비지도사선임인원수>370</경비지도사선임인원수>
      <일반>356</일반>
    </record>
    - <record>
      <기관>대구청</기관>
      <기계>9</기계>
      <경비지도사선임인원수>221</경비지도사선임인원수>
      <일반>212</일반>
    </record>
  </records>
</dataGrid>
```

Tag에 의하여 데이터의 의미 표현

데이터 유형(3/3)

◆ 빅거래 데이터

- 기업들이 기존에 보유하고 있는 시스템에 존재하는 데이터
- 예: 거래 데이터(OLTP), 분석용 데이터(DW, OLAP)
- 기존 DBMS(오라클, MySQL등)에서 주로 처리

◆ 빅상호작용 데이터

- 사람과 사람, 사람과 기계, 기계와 기계 간의 상호작용으로 발생하는 데이터
- **SNS나 IoT(Internet of Things, 사물인터넷)** 등에서 생성
- 예: 텍스트, 이미지, 센서 데이터

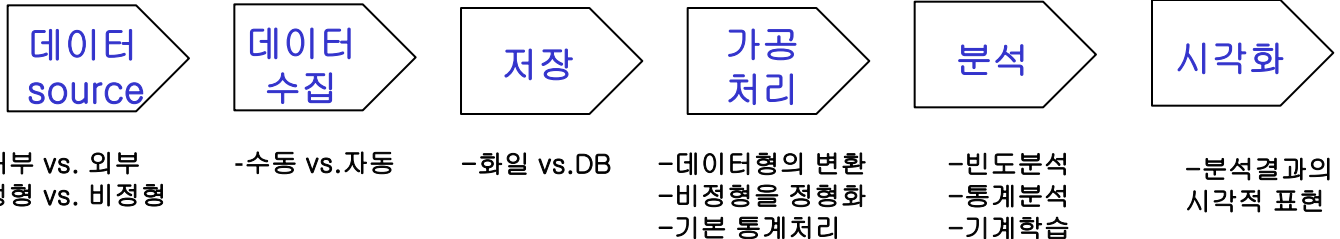
빅데이터 분석과 IT융합(1/2)

◆ 빅데이터 분석은 진정한 IT융합 기술

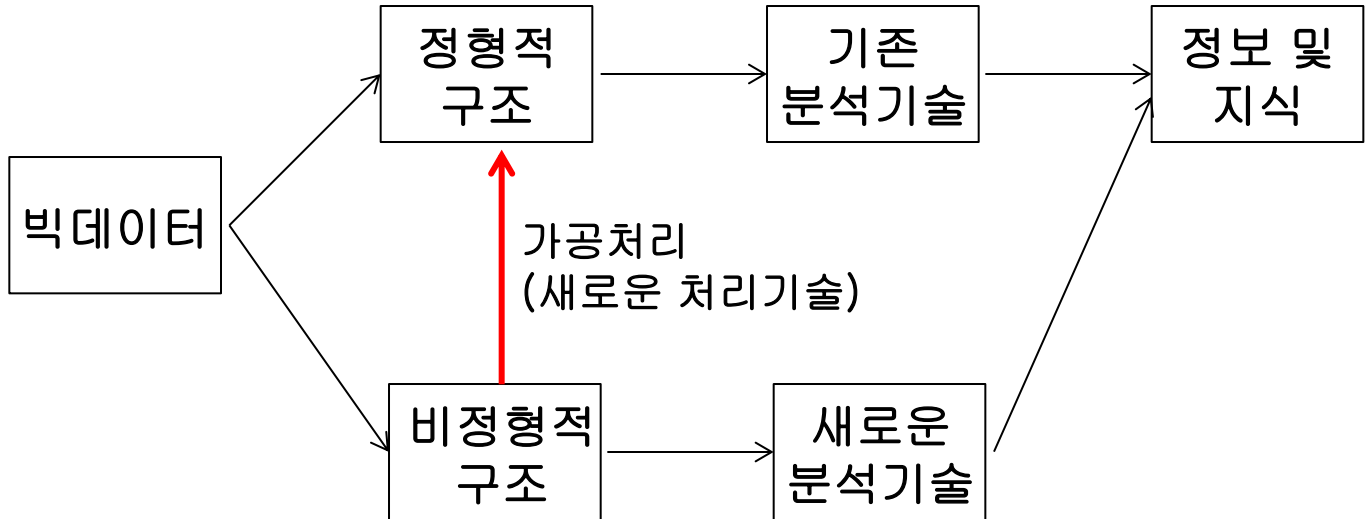
- 모든 산업에서 데이터가 발생하며, 분석을 통한 가치창출 가능

◆ 비정형(비구조적) 데이터도 결국은 정형화 되어야 분석 가능

- 가공처리 단계에서 많은 노력 소요



빅데이터 분석과 IT융합(2/2)



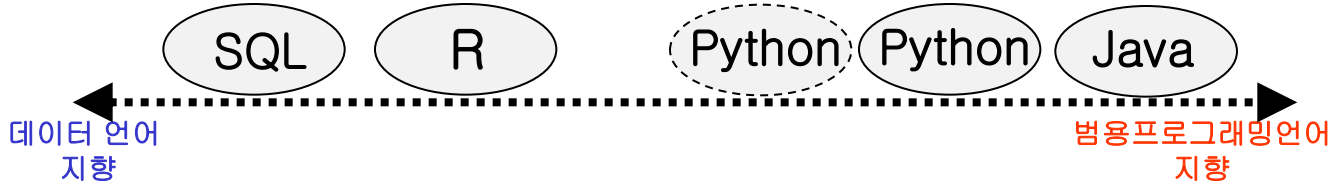
데이터 분석도구 유형

- GUI(Graphic User Interface) 기반
 - 메뉴를 선택하면서 데이터 처리
 - ✓ 엑셀, SPSS 등
- 명령어 표현(Command Driven) 기반
 - 데이터언어
 - ✓ 데이터 처리에 특화된 프로그래밍언어
 - ✓ 단어와 문법규칙에 따라 명령어 표현
 - ✓ SQL, R , Python 등

데이터 분석절차에 따른 데이터언어 비교



데이터언어의 비교



- SQL

- Structured Query Language
- 데이터베이스를 관리하기 위하여 설계된 표준 데이터언어
- 대부분의 DBMS 소프트웨어에서 지원

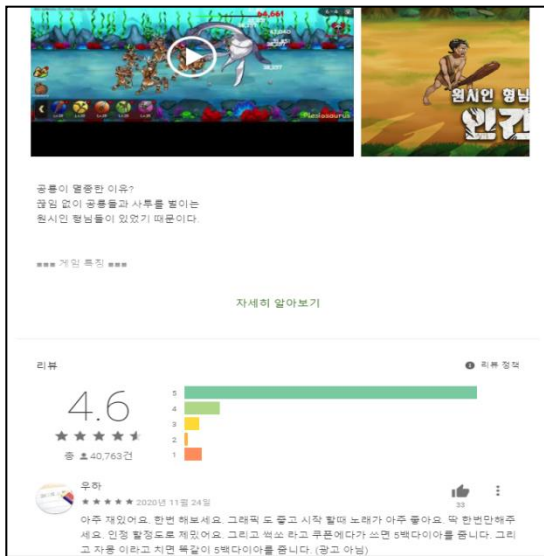
- R

- 데이터분석을 위한 통계 및 그래픽 시각화를 지원하는 오픈 소프트웨어
- S언어 근간의 명령어 표현 기반으로 데이터 처리

- Python

- 범용 프로그래밍언어와 가까움
- 데이터 분석을 위한 명령어 표현을 쉽게 할 수 있음

웹 review 데이터 수집 및 분석(1/2)



공통이 많은 이유?
공인 없이 공통들과 사투를 보이는
원시인 할남들이 있었기 때문이다.

*** 개인 특징 ***

자세히 알아보기

리뷰

4.6

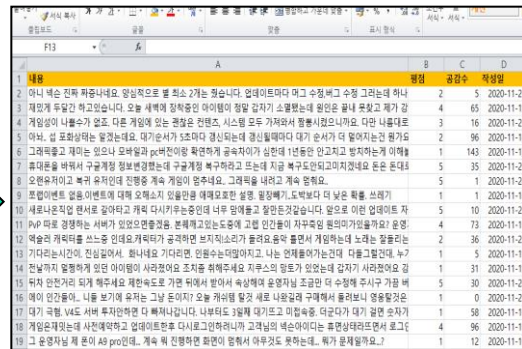
★★★★★

총 440,763건

우하

★★★★★ 2020년 11월 24일

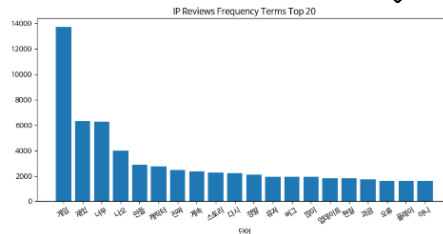
아주 재밌어요. 한번 해보세요. 그래픽도 좋고 시작 할때 노래가 아주 좋아요. 막 한번만학후
세로. 인정 할것으로 재밌어요. 그리고 책소 라고 루프에다가 쓰면 5백다이어를 뽐니다. 그리
고 라운 이하고 가면 특일이 5백다이어를 뽐니다. (광고 아님)

제목	평점	공감수	작성일
1 아니 책은 처음 나왔네요. 알았지만은 볼 땐 2차는 평순입니다. 절대이트이다. 마그 수정해그 수종 그라는데 하나	2	5	2020-11-25
2 평정계 두달간 하고있네요. 오늘 새벽에 양악종인 아이엘이 정말 듣기가 소름돋는 원인은 몰래 웃고그 제가 일	4	65	2020-11-19
4 개인적인 내용수가 없고 다른 개성있는 것은 없습니다. 시스템 으르 가져서 뽐내시킴이네요. 다만 내용대로	3	16	2020-11-25
5 아바. 성 포화상태는 괜찮네요. 대가수사가 150이다. 경신되는 결심할때대 대가 수사가 더 떨어진것에 이해	2	96	2020-11-14
6 그대들고 그대미는 원시나 모박월과 50백미의 확연하게 공작처리가 심한데 1년동안 안고지고 받쳐어느결 이해	1	143	2020-11-15
7 후다들고 바쳐서 구공계정 원본변경하는데 구공계정 복귀라고 있는데 지금 복귀도안되고미지경대로 돈은 못대	5	35	2020-11-21
8 오만유저이고 복귀 유저인데 진행중 계속 개성만 맞추네요. 그래픽을 내리고 계속 맞춰	5	1	2020-11-24
9 트럼이벤트 오픈 이벤트에 대해 오프닝식 있을만큼 대대모호한 설정. 할당해가.도박보다 더 낫은 것을 쓰게	1	1	2020-11-10
10 새로나온것인 원서로 알려하고 개작 다시키우는중인데 너무 망쳐놓고 잘만것같습니다. 앞으로 이런 업데이트 자	5	10	2020-11-25
11 PVP 따로 경쟁하는 서버가 있었으면좋겠음. 본래개고있는공예 웹 안만안게 자유적인 원미가있을까요? 운영	4	73	2020-11-11
12 엑슬라 개작터를 쓰느중 인데오개작터가 공격하면 브지식소리가 뽐내요.유악 올린서 제정하는데 노래는 잘들리는	2	36	2020-11-23
13 기다리는시간이. 전신공에서. 회사로 기다리면. 안통수는대면이치고. 나는 언제올까는전데. 다음그날인데. 누?	1	5	2020-11-15
14 전날까지 열렬하게 있던 아이엘이 사라졌어요. 오직을 위해주세요. 7루스의 일로에 갑자기 사라졌어요. 같	1	31	2020-11-19
15 뭐지 만가게 되게 할게요. 개판속으로 가면 뒤에서 받아서 육식해에 올릴것만 조금만 더 수중에 주시구 가를 바	5	30	2020-11-30
16 예이 인간들이. 나를 보기에 유라는 그날 돈이요? 오늘 개성을 할것. 새로 나와와 구해서서 올릴것만 잘들었음	1	0	2020-11-25
17 대가 국할. V4도 새로 투지전환만 다 해쳐나갔네요. 나부터도 3월째 대가토도 미합속을. 더군다나 대가 올면 순차	1	58	2020-11-16
18 제일큰게있는데 사전개작하고 업데이트한후 다시로그인하러나고 그대날의 색은아이드는 휴관상태로써서 로그	4	96	2020-11-10
19 그 운영자님 제 모이 아8 pro인데. 계속 뭐 진행하면 회원이 많아서 아무것도 못하는데. 뭐가 문제일까요?	1	12	2020-11-10



단어빈도분석



로지스틱회귀분석 (주제→긍정/부정)



웹 review 데이터 수집 및 분석(2/2)

```
from selenium import webdriver
import time

driver = webdriver.Chrome(executable_path="D:\\practice\\chromedriver_win32\\chromedriver.exe")
driver.get(url)
html = driver.page_source
bsObj=BeautifulSoup(html, 'lxml')
div_reviews = bsObj.find_all("div", {"class":"d15Mdf bAhLNe"})
print(len(div_reviews))

''' 여러 페이지(10번 스크롤하면서)를 크롤링 '''
cnt = 0
while(cnt < 10):
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
    cnt += 1
    time.sleep(0.5)
html = driver.page_source
bsObj=BeautifulSoup(html, 'lxml')
div_reviews = bsObj.find_all("div", {"class":"d15Mdf bAhLNe"})
print(len(div_reviews))
div_reviews[4]
type(div_reviews)
print(len(div_reviews))

while (True): #스크롤바 첫 버튼 발견할 때까지 내림
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
    time.sleep(0.05)
    try:
        element = driver.find_element_by_xpath('//*[@class="U26fgb 00WRkf aG5Srb C00Vfc n9Lfj"]')
        if (element is not None):
            element.click()
            break;
    except Exception:
        continue

errTime = 0
successTime = 0

while (errTime < 100 and successTime < 100):
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
    time.sleep(0.05)
    try:
        element = driver.find_element_by_xpath('//*[@class="U26fgb 00WRkf aG5Srb C00Vfc n9Lfj"]')
        if (element is not None):
            element.click()
```

```
from eunjeon import Mecab

mecab = Mecab()
result2 = mecab.pos("폴사 태깅을 지원합니다")

mecab.morphs("할태소 톨리를 지원합니다")

mecab.nouns("평사에 해당하는 할태소만 추출합니다")

# NNG: 일반명사, NNP: 고유명사, VV: 동사, VA: 형용사, VCP: 긍정
def getNVM(text):
    tokenizer = Mecab()
    parsed = tokenizer.pos(text)
    pos = []
    tags = ['NNG', 'NNP', 'VV', 'VA', 'VCP', 'VCN']
    for word in parsed:
        tag = word[1]
        if (tag in tags):
            pos.append(word[0])
    return pos
text = "제주대학교 경영정보학과 출결통은 맞췄요"
print(getNVM(text))

#reviews.csv 파일을 데이터프레임으로 읽기 => 5주 3차시 강의
import pandas as pd

input_file = 'd:/practice/review.csv'
output_file = 'd:/practice/pandas_output.csv'

reviews = pd.read_csv(input_file)
print(reviews)
reviews.loc[1,1]
type(reviews.loc[1,1, '평점'])
#reviews.to_csv(output_file, index=True)
reviews.columns
reviews['내용']
type(reviews.loc[1, '평점'])
docs = []
words = []
for doc in reviews['내용']:
    docs.append(doc)
    words.append(getNVM(doc))
print(getNVM(doc))
```

감사합니다