빅데이터의 개요 및 분석 과정

Soongsil University

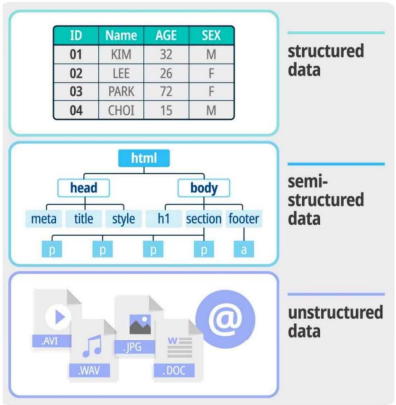
숭실대학교
베어드교양대학
강의선
iami86@ssu.ac.kr

빅데이터 정의

- 일반적인 데이터베이스 소프트웨어가 수집, 저장, 관리, 분석할 수 있는 범위를 초과하는 대규모의 데이터**(맥킨지 (Mckinsey))**
- 향상된 시사점과 더 나은 의사결정을 위해 사용되는 것으로 비용 효율이 높고 혁신적이며 대용량 고속 및 다양성을 가지는 정보 자산**(가트너(Gartner))**

빅데이터 분류

- 정형데이터(structured data)
  - 미리 정해 놓은 형식과 구조에 따라 저장된 데이터
    예) 관계형 데이터베이스의 테이블, 스프레드시트, CSV 등
- 반정형데이터(semi-structured data)
  - 일정한 규칙의 고정된 필드에 저장되어 있지 않지만 데이터의 구조 정보를 데이터와 함께 제공하는 파일형식 데이터
    예)XML, HTML, JSON, 웹문서, 웹로그 등
- 비정형데이터(unstructured data)
  - 정의된 구조가 없이 데이터 자체만으로 내용에 대한 질의 처리를 할 수 없는 데이터
    예) 소셜 데이터, 텍스트 문서, 동영상/이미지/음성 데이터, 문서(PDF) 등



[이미지 출처]: 정보통신용어사전
(http://terms.tta.or.kr/dictionary/dictionaryView.do?word_seq=175128-2)

(정형) 데이터 구성

- 데이터(표)는 행(row)과 열(column)로 구성
  - 레코드 = 행(Row)
    - ✓ 하나의 단위로 다루어지는 데이터의 집합
  - 칼럼 = 열(Column)
    - ✓ 표의 세로 축, 열에 해당
    - ✓ 변수(일반적인 컴퓨터 분야)
    - ✓ 속성(인공지능 분야)
    - ✓ 특징(패턴인식 분야)
- 칼럼의 종류
  - 수치형(Numeric) : 정수형, 실수형, Bool형
  - 범주형(Categorical) : 순서형, 텍스트

학번	이름	학과	학년
20211021	박형식	영어영문학과	1
20205412	공유	화학과	2
20210578	아이유	수학과	1
19983125	송중기	경영학과	4

1. 레코드는 개수는?
2. 칼럼의 개수는?
3. 수치형 칼럼 개수는?
4. 범주형 칼럼 개수는?

(정형) 데이터 구성

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

```
1 titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB
```

- 데이터 정보 확인하기

1. 레코드는 개수는?
2. 칼럼의 개수는?
3. 수치형 칼럼 개수는?
4. 범주형 칼럼 개수는?

Soongsil University

빅데이터 분석 과정

목표설정 → 데이터 수집 → 데이터 정제 → 데이터 탐색 → 데이터 모델링 → 결과 발표 및 자동화

- 무엇을?
- 어디 사용?
- 어떤 데이터?

- 데이터 유형과 종류에 따라 맞는 수집 기술 사용

- 데이터 여과 정제, 통합,축소, 변환

- 데이터 변수간 관계 및 상호작용, 분포 등을 탐색

- 변수를 선택하여 모델을 구성 및 평가
- 통계분석모델
- 머신러닝/딥러닝 모델

Soongsil University

공공 데이터를 이용한 데이터 분석

Soongsil University

숭실대학교
베어드교양대학
강의선
iami86@ssu.ac.kr

공공데이터 제공 사이트 예

사이트	설명
https://www.data.go.kr	한국 정부에서 제공하는 공공데이터(공공데이터포털)
https://kostat.go.kr	통계청에서 공개하는 데이터
https://opendata.hira.or.kr	한국 보건 의료 빅데이터 개방 시스템
https://www.localdata.kr	한국 지방 행정 인허가 데이터
https://www.mcst.go.kr	한국문화체육관광부 문화 데이터
https://data.seoul.go.kr/	서울 열린데이터 광장
https://data.gg.go.kr	경기도 공공데이터 개발 포털
https://www.data.gov/	미국 정부의 공공데이터
https://data.worldbank.org/	세계 은행에서 제공하는 개방 데이터
https://open.fda.gov/	미국 식약청의 개방 데이터

Soongsil University

APT 실거래가 데이터 수집

1. **목표설정** :동작구 상도동의 아파트 실거래가를 조사해보자.
   - 평당 거래가격 / 층별 거래가격...

2. **데이터 수집** : 공공데이터 <APT 실거래> 데이터 확보

3. **데이터 정제** : 분석 가능한 데이터로 전처리하기 (Excel 또는 python활용)
   1) 16번째 행을 열이름으로 설정하고 불필요한 행 삭제하기
   2) 칼럼 이름 단순화하기
   3) 불필요한 열 삭제하기
   4) 거래금액에서 쉼표 지우기
   5) 평수 칼럼 추가

4. **데이터 탐색 & 데이터 모델링**
   - pandas로 불러오기, 데이터 살펴보기
   - 간단한 분석 맛보기 (Seaborn)

APT 실거래가 데이터 수집(python 활용)

- 2주차_상도동아파트 실거래분석(python). ipynb 참조

수고하셨습니다.