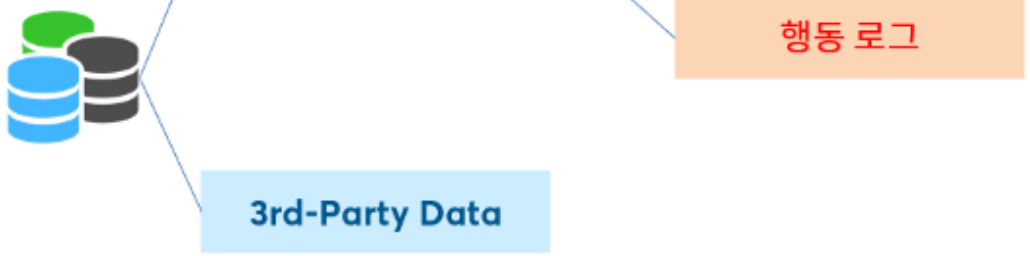
Python 데이터 분석 01

2021-2 KCA

김지환

데이터 분석이란?

- 데이터 분석을 하는 이유 : 인사이트 추출 + 통계적 분석



가입완료, 예약완료,
결제완료, 결제취소,
전화번호 변경…

가입하기 버튼 클릭,
배너 스와이프,
검색 실행 …

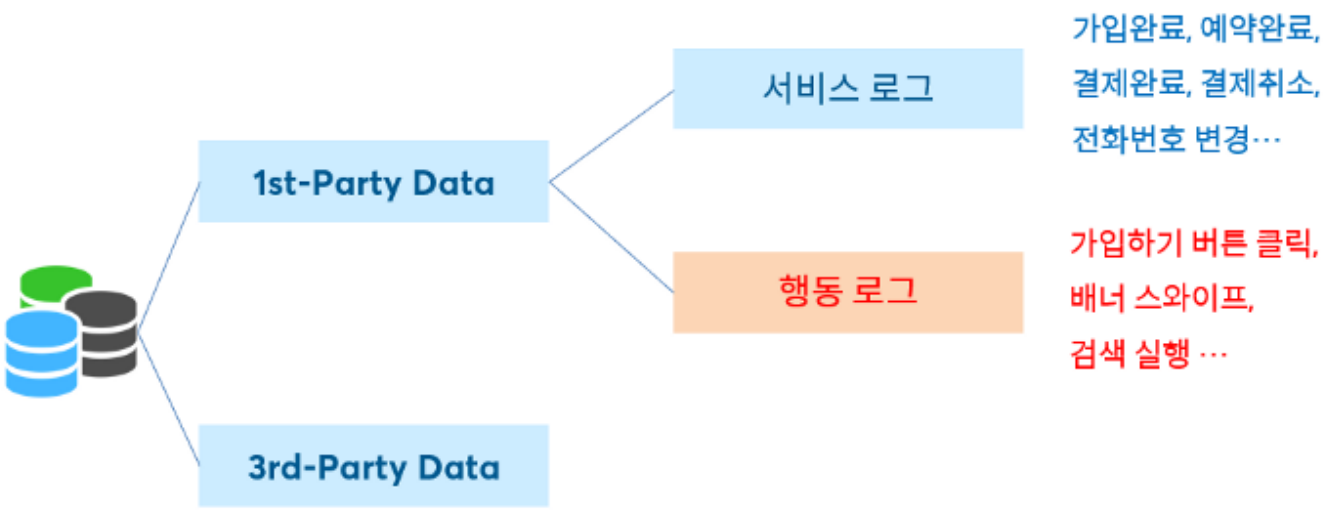
어떤 웹/앱/서비스가 출시!!

하루에 100명만 사용한다 쳐도.
수많은 로그 발생!

뭔가 페이지를 이동하고, 클릭하고, 무슨 영상을 보고, 어디에 좋아요를 누르고 등등...

⇒ 로그를 수집해서 분석
⇒ 더 좋은 사용자 경험. UX
⇒ 사용자 증가 + 회사 수익 증가

여담) 퍼스트 파티, 써드 파티란?



퍼스트 파티 : 자사 서비스와 관련된 것들.

써드 파티 : 외부 업체 서비스와 관련된 것들.

<광고에 한해서>
퍼스트 파티 쿠키는 자사의 서비스에서만 사용 가능하다.
(A사의 쿠키라면 A사가 만든 사이트 내에서만)

써드 파티 쿠키는 여러 사이트를 돌아다닐 때도 사용 가능하다.
(대표적으로 구글이 있지만, 최근 크롬에서 써드 파티 쿠키를 제공하지 않을 것이라 밝힘)

여담 2) 쿠키란?

- 웹 서비스에서 사용자마다 구별 가능하게 만든 것.
- 사용자 컴퓨터에 저장이 되며, “항상 로그인 유지“ 등의 옵션을 가능하게 함.
- 사용자 맞춤 광고를 띄우는 데에도 쓰임.

데이터 분석의 단계

- 수집
- 가공(전처리)
- 저장
- 분석
- 시각화

데이터 수집

- 무엇을 분석할까?

- 수집 방식 : 크롤링 (직접 수집), 캐글, 공공데이터, IMDB, 데이터 api 등..

예1) 실시간 검색어를 만들어 보고 싶다
> 웹 상의 뉴스기사 등을 긁어오는 크롤러 제작

예2) 버스 배차 정보를 알고 싶다
> 공공데이터에서 버스 도착시간 api 이용

여담) 크롤링이란?

- 복잡한 웹에서 원하는 정보(보통은 text)만을 추출하기 위해서 사용.
- python beautifulsoup / selenium 유명
- HTML 태그로 찾아 들어가는 식으로 만들어진다.
- 웹의 통신 및 구조에 대해서도 알면 좋음

여담) API 란?

- Application Programming Interface
- 특정 작업을 하게끔 만든 함수들의 집합
- 특정 신호에 특정 응답을 한다고 약속된 것.

- [Web API]

특정 web 주소에 들어가면 json 형태로 응답 값을 던져주는 것.
우리는 그것을 받아서 이용하면 된다!

보통 공공API / 카카오, 네이버 인증 API 등은 인증 키 발급이 필요함.

api 예시 : https://jsonplaceholder.typicode.com/
https://jsonplaceholder.typicode.com/todos/

데이터 가공(전처리) + 저장

- 데이터를 어떤 형태로 저장할까?
- 전처리 : 그 형태로 만들어 주는 것!
- 다양한 파일 포맷
- 보통 csv, json 많이 쓰임!
- SQL로 DB에 넣어 놓기도 함!
- 각 파일 변환에는 parser(구문 분석기)가 필요

▶ 기계 판독이 가능한 형태의 포맷 단계별 구분·비교

구분	1단계	2단계	3단계	4단계	5단계
기계 판독이 가능한 형태	미충족포맷 (포털등록불가)	최소충족포맷	오픈포맷*		
특징	특정 소프트웨어에서 읽을 수만 있는 데이터로 자유로운 수정, 변환 불가	특정 소프트웨어에서 읽고 수정, 변환 가능	모든 소프트웨어에서 읽고 수정, 변환 가능	URI**를 기반으로 데이터 속성 특성 관계를 기술하고 있는 데이터 구조	웹상의 다른 데이터와 연결, 공유 가능
예시	PDF	HWP, XLS, JPG, PNG, WMV, MPEG, MP3, SWF	CSV, JSON, XML	RDF	LOD

* 오픈포맷(open format) : 모든 소프트웨어에서 자유롭게 활용(수정, 편집 등)할 수 있는 형태의 데이터(예: CSV, JSON, XML 등 3단계 이상의 포맷)

** URI(Uniform Resource Identifier, 통합자원식별자) : 웹 상의 특정 콘텐츠를 다른 콘텐츠(텍스트, 이미지, 동영상 등)들과 구별하여 인식 · 확인할 수 있는 고유값(유일식별자)

데이터 분석

- 흔히 데이터 분석? 하면 떠오르는 부분.
- 각종 통계학, 수치해석적 지식 (회귀분석, 보간법 ...) 을 코드로 구현 (라이브러리에서 가져다 쓰기...)
- 최근에는 머신러닝, 딥러닝을 도입해서 예측 모델을 생성하기도 함.

데이터 시각화

- 목적에 맞게
- python의 matplotlib, matlab의 plot 그리기 등 툴은 다양함.
- 논문 사용? => LaTex 변환

파이썬 사용 이유

- 파이썬으로 가능한 것은 엑셀로도 대부분 가능하다

- 엑셀 안의 VBA (Visual Basic Application) 기능으로 무려 코드도 짤 수 있다!

- 그런데 왜 파이썬?

⇒언어가 더 간편함
⇒복잡한 수식 직접 수정 가능(수치해석, 통계 등)
⇒자동화가 편리함 (매일 아침 00시에 돌리기 등)
⇒DB와 연동, 서버 연동 등이 편리함
⇒다양한 파일 포맷 읽어오기 가능

파이썬 문법 복습 : for, iterator

- iterator : 값을 차례대로 꺼낼 수 있는 객체 (list, tuple, range, string..)

C와는 다르게, 파이썬의 for 문에는 iterator기만 하면, 모두 들어갈 수 있음. C는 숫자만.

for i in <iterator>:

i는 iterator에서 하나씩 꺼내와진다.

- for i in range(1, 10):

range로 많이 배우는데, 꼭 range만 쓸 필요는 없다!!

range -> list : list(range(a, b, c))

파이썬 문법 복습 : 리스트 슬라이싱

[1:-2] 의미??

a = [1, 2, 3, 4, 5, 6]

- a[0:3] >> 1, 2, 3 [0번 부터, 3번 이전까지]

C 배열에서 포인터 접근을 생각해보자!

-가 붙으면 뒤에서부터 센다!

* iterator들은 슬라이싱이 가능함! (range, string ….)

유용한 파이썬 문법 : zip()

a = [1, 2, 3]

b = ['a', 'b', 'c']

a, b는 iterator

- for i, j in zip(a, b):

> a, b에 i, j로 각각 접근 가능

- for k in zip(a, b):

> k는 a, b가 묶인 튜플로 나옴

유용한 파이썬 문법 : f string (python 3.6~~)

- f string

문자열 만들 때 f를 붙여주면 다른 변수를 {}안에 바로 집어넣기 가능!

마치 JS의 백틱(``) 과 비슷

```
a = 11
문자열 = f"안녕하{a}세요"
>> 안녕하11세요
```

+ 정규 표현식 (python 기본 모듈 re)

반복되는 문자 처리가 간편해짐!

-> 그때 그때 찾아서 쓰면 된다!

라이브러리(패키지) 추가 방법

- 파이썬은 pip (패키지 관리자) 존재!!
- 간편하게 설치, 제거, 관리 가능!

- 패키지 > 모듈 > 클래스, 함수, 변수

```
import 모듈
...
모듈.함수()
```

```
from 모듈 import 변수, 함수, 클래스
...
함수()
```

```
import 모듈 as 이름
...
이름.함수()
```

```
from 모듈 import 함수 as 이름
...
이름()
```

```
import 패키지.모듈
...
패키지.모듈.함수()
```

```
from 패키지.모듈 import 변수, 함수, 클래스
...
함수()
```

참고 : https://dojang.io/mod/page/view.php?id=2441
점프 투 파이썬 & 파이썬 코딩 도장 추천!!