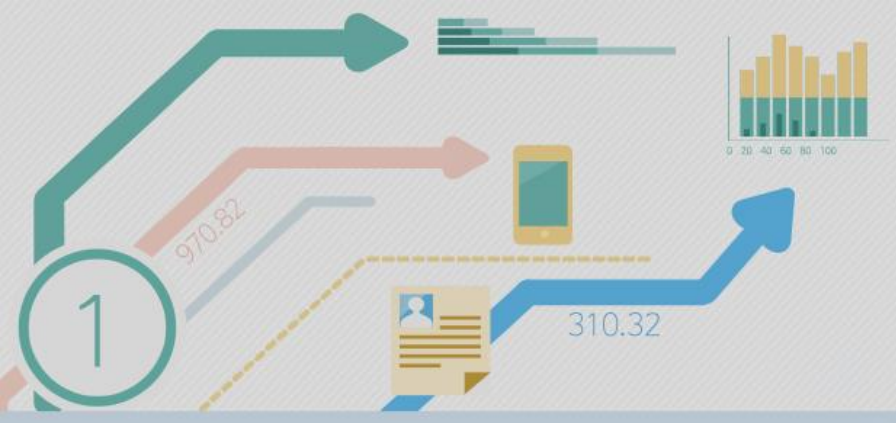


빅데이터 분석 결과 시각화

실시간 민원 신고 시각화

한국기술교육대학교
온라인평생교육원

학습내용

- 실시간 데이터 수집
- 자동차 민원 데이터 시각화
- 세탁기 민원 데이터 시각화

학습목표

- 데이터 로딩 함수를 활용한 실시간 데이터 수집방법에 대해 설명 할 수 있다.
- PHP를 활용하여 자동차 민원을 시각화 할 수 있다.
- R 프로그램을 활용하여 세탁기 민원을 시각화 할 수 있다.

실시간 데이터 수집

1. 데이터 로딩

◆ 데이터 로딩 함수

- 데이터는 다양한 형식으로 존재함
  - CSV

```
d3.csv("data.csv", function(err, data)
{ console.dir(data); });
```

  - JSON

```
d3.json("data.json", function(err, data)
{ console.dir(data); });
```

  - XML

```
d3.xml("data.xml", function(err, data)
{ console.dir(data); });
```

  - TEXT

```
d3.text("data.txt ", function(err, text)
{ console.dir(text); });
```

- D3는 이러한 파일들을 비동기(Async Request)로 가져오는 함수를 제공함

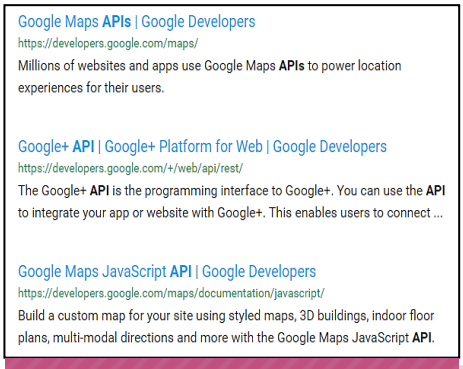
실시간 데이터 수집

1. 데이터 로딩

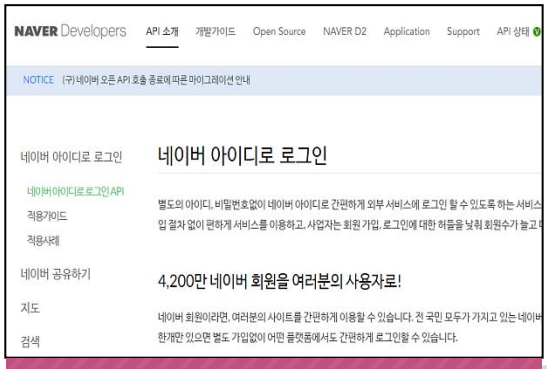
◆ API를 제공하는 방식

- 외부에서 데이터를 사용하도록 공개하는 사이트들은 각자의 API를 제공함

  예) 구글, 네이버, 애플, 페이스북 등



구글API 개발자 사이트



네이버API 개발자 사이트

〈출처 : https://developers.google.com/s/results/?q=API&hl=ko〉
〈출처 : https://developers.naver.com/products/login/api〉

- API를 제공할 때에는 키(Key)가 필요함
  - API를 통해 데이터를 입력 · 이용하려는 대상을 인증하기 위함
- 대체로 널리 쓰이는 방법
  - 공개키(Public Key)와 개인키(Private Key) 한 쌍을 발급하고 이를 통해 대상 인증

● 실시간 데이터 수집

1. 데이터 로딩

◆ 데이터베이스 구성 - SQL(관계형 데이터베이스)

장점	단점
• 범용성 : 다양한 용도로 사용 • 고성능 : 일반적으로 높은 성능 • 데이터의 일관성 보증 • 정규화에 따른 갱신 비용 최소화	• 대량의 데이터 입력 처리 어려움 • 갱신이 발생한 테이블의 인덱스 생성 및 스키마 변경, 컬럼의 확장의 어려움 • 단순히 빠름

● 주요 제품 종류(제품명 / 제조사)

- Oracle / Oracle
- MS-SQL Server / Microsoft
- MySQL / Oracle(SunMicroSystems)
- DB2 / IBM
- Infomix / IBM
- Sybase / Sybase
- Derby / APache
- SQLite / Opensource

데이터 사이에 관계가 존재하는 경우	데이터 사이에 관계가 필요 없는 경우
• SQL 데이터베이스 사용	• NoSQL 데이터베이스 사용

● 실시간 데이터 수집

1. 데이터 로딩

◆ 데이터베이스 구성 - NoSQL

- SQL을 사용하지 않는다는 의미
  - Not Only SQL : SQL이 필요 없다는 의미가 아니고, 개선 / 보안의 의미
  - Non-Relational Operational Database SQL : 관계형 데이터베이스가 아님
- NoSQL의 장점
  - 대용량 데이터
  - 데이터 분산 처리
  - Cloud Computing
  - 빠른 읽기 / 쓰기 속도
  - 유연한 데이터 모델링
- NoSQL의 종류

① key / value
  - 휘발성 / 영속성
  - Memchached, Tokyo Tyrant, Flare, Roma, Redis

② Document
  - 스키마 정의 없음
  - MongoDB, CouchDB

③ Big Table(Column 형) DB
  - 뛰어난 확장성, 검색에 유리
  - Hbase, Casandara, Hypertable

2. 데이터 수집

◆ 반정형 데이터의 특징

● 반정형 데이터(Semi-Structure Data) 정의

- 정형 데이터 : 데이터의 스키마 정보를 관리하는 DBMS와 데이터 내용이 저장되는 데이터 저장소로 구분됨
- 반정형 데이터

  - 데이터 내부에 정형데이터의 스키마에 해당되는 메타데이터를 가짐

  - 일반적으로 파일 형태로 저장됨

● 반정형 데이터의 형태

```
[{“Letter”:“소음”, “Freq”:“85”},{“Letter”:“서비스”, “Freq”:“95”},
 {“Letter”:“불만”, “Freq”:“91”},{“Letter”:“수리”, “Freq”:“88”},
 {“Letter”:“고객”, “Freq”:“84”},{“Letter”:“직원”, “Freq”:“78”},
 {“Letter”:“센터”, “Freq”:“72”},{“Letter”:“시간”, “Freq”:“65”},
 {“Letter”:“전화”, “Freq”:“61”},{“Letter”:“접수”, “Freq”:“57”},
 {“Letter”:“견인”, “Freq”:“51”},{“Letter”:“마모”, “Freq”:“41”},
 {“Letter”:“타이어”, “Freq”:“32”},{“Letter”:“사고”, “Freq” “31”}]
```

● 반정형 데이터의 예

URL 형태로 존재	HTML
오픈 API 형태로 제공	XML, JSON
로그형태	웹로그, IoT에서 제공하는 센서 데이터

2. 데이터 수집

◆수집방법의 분류

● 수집 데이터의 형태와 종류에 따른 분류

- 크롤링
- ETL(데이터수집프로그램)
- 로그수집
- ftp
- http
- RDB 수집방법

● 수집 데이터의 형태에 따라 물리적으로 저장형태가 분류됨

● 수집 데이터의 연동방법에 따라 수집방법이 결정됨

① 원본 데이터 요청 후 확인

② 소켓(Socket) 통신으로 연동하는 DBMS 수집

- 주로 DBMS에서 벤더가 제공하는 드라이버를 통해 데이터를 연동함

③ 스트리밍 방식으로 연동하는 로그 데이터, 센서 데이터 수집

- 주로 시스템에서 발생하는 데이터로 스트리밍 방식의 연동방법
- tcp, 블루투스, RFID 등 여러 가지 통신 프로토콜이 존재함

④ ftp 프로토콜을 사용하는 이진 파일 수집

- 비정형 데이터를 수집할 때 필요한 방법
- 데이터 수집 후 수집한 데이터를 그대로 사용할 경우도 있지만, 서비스 활용을 위해 데이터의 파싱이 필요함

⑤ 크롤링 : http 프로토콜 사용하는 스크립트 파일 수집

- 웹상에 존재하는 데이터는 반정형 데이터로 존재하므로 수집기 내에서 활용 가능한 데이터의 형태로 파싱처리 후 시스템에 저장함

2. 데이터 수집

◆크롤링 기초 실습

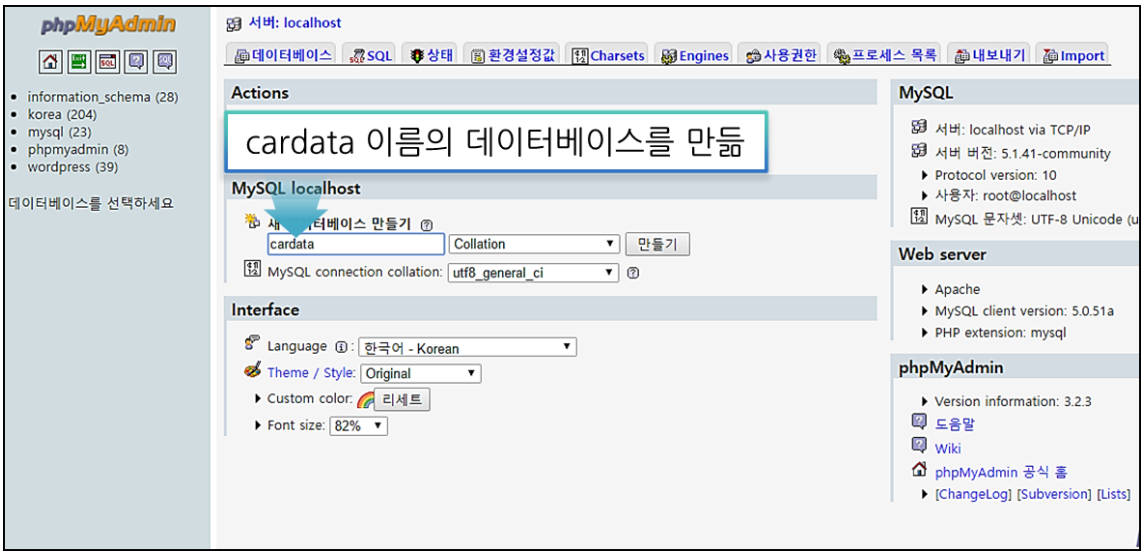
● R 스크립트로 크롤링 하는 방법

```
setwd(₩c:"₩₩R")          ←작업 디렉토리 설정
library(rvest)            ←웹사이트에서 원하는 부분만 스크랩 해주는 패키지
library(httr)             ←HTTP 통신
url = '주소값'
response = GET(url)       ←url 주소 가져옴
htxt = html(response)     ←HTML 코드 불러옴
comments = html_nodes(htxt, 'span.comment')   ←HTML 소스 분석
links = html_nodes(comments, 'a')             ←a로 링크된 소스 분석
html_text(links)                              ←글자만 불러옴
a = repair_encoding(html_text(links))         ←윈도일 경우
                                                한글 깨짐 방지
a
write(a,"크롤링.txt")                         ←크롤링 된 데이터 저장
```

자동차 민원 데이터 시각화

1. PHP를 이용하여 DB를 JSON으로 출력

◆ MySQL DB 설정



◆ 테이블

- PHP를 이용해 JSON으로 출력하기 위하여 DB 내용을 테이블로 만듦
- 데이터 값은 인터넷을 통해 얻을 수 있음

			Letter	Freq
☐			소음	85
☐			서비스	95
☐			불만	91
☐			수리	88
☐			고객	84
☐			직원	78
☐			센터	72
☐			시간	65
☐			전화	61
☐			접수	57
☐			견인	51
☐			마모	41
☐			타이어	32
☐			사고	31

● 자동차 민원 데이터 시각화

1. PHP를 이용하여 DB를 JSON으로 출력

◆ PHP 한글 코드 깨짐 방지

● 한글이 깨지는 문제

- 한글에 대한 고려 없이 PHP 소스코드를 만들어 놓으면 한글은 JSON으로 변환 중에 깨짐
- DB, PHP소스코드, PHP header 모두 UTF-8인데도 불구하고 MySQL_fetch_array를 이용해 SQL 응답 내용을 PHP배열로 저장을 하는 과정에서 한글이 깨짐

→ 한글은 urlencode() method를 이용해 한번 감싸주고 나중에 JSON으로 변환 할 때 urldecode() method를 이용해 디코드 함

● 숫자가 String형으로 출력되는 문제

- 숫자가 String형으로 출력되면 데이터 이용할 때 애로사항이 많음

→ SQL 응답 내용을 PHP Array로 저장 할 때 앞 부분에 (int)를 기입하여 꼭 int형으로 저장함

자동차 민원 데이터 시각화

1. PHP를 이용하여 DB를 JSON으로 출력

◆ PHP DB 호출 코드

- 자신이 호스팅한 사이트에 아이디와 패스워드 DB이름을 넣음

```
<?php
$db_host = “호스트이름”;
$db_user = “아이디”;
$db_passwd = “패스워드”;
$db_name = “DB이름”;
```

```
<?php
$db_host = "localhost";
$db_user = "root";
$db_passwd = "apmsetup";
$db_name = "cardata";
```

실습 예

자동차 민원 데이터 시각화

1. PHP를 이용하여 DB를 JSON으로 출력

◆ 데이터베이스 연결

```
MySQL_connect($db_host,$db_user,$db_passwd);
MySQL_select_db($db_name);
MySQL_query('set session character_set_connection=utf8;');
MySQL_query('set session character_set_results=utf8;');
MySQL_query('set session character_set_client=utf8;');
```

① 데이터베이스 연결

- MySQL에 설정된 데이터베이스를 호출하는 것

```
MySQL_connect($db_host,$db_user,$db_passwd);
MySQL_select_db($db_name);
```

② 쿼리 실행

- MySql 입출력 인코딩 형식을 설정하는 명령어

```
MySQL_query('set session character_set_connection=utf8;');
MySQL_query('set session character_set_results=utf8;');
MySQL_query('set session character_set_client=utf8;');
```

자동차 민원 데이터 시각화

1. PHP를 이용하여 DB를 JSON으로 출력

◆ JSON 호출

- 배열을 이용해 cars에 있는 데이터를 json으로 변환하여 호출함

```
$result = MySQL_query("SELECT * FROM cars;");
$rows = array();
while($row = MySQL_fetch_array($result))
{
$JSONdata =array("Letter" => urlencode($row['Letter']),
"Freq" => urlencode($row['Freq']));
  array_push($rows,$JSONdata);
}

echo (strip_tags(urldecode(json_encode($rows))));
```

2. 자동차 민원 데이터 시각화 실습

◆ 자동차 민원 데이터 시각화 실습 순서

① 데이터 준비하기

② PHP로 JSON 출력 코드 작성하기

③ 자동차 민원 시각화 코드 작성하기

④ 자동차 민원 시각화 결과물 확인하기

세탁기 민원 데이터 시각화

1. 세탁기 민원 데이터 수집

① 포털사이트에서 '세탁기 민원' 데이터 검색

② 민원 데이터 수집

[1] "또한 '자동차대여(렌트)'(56.0%) 관련 민원과 '세탁기'(32.7%), '화장품세트'(27.7%) 등이 뒤를 이었다. 또한 지난해 같은 기간 대비 증가한 민원으로는..." [2] "A/S와 민원 때문에 대부분 3개월을 못버티고 그만두시는 분도 많아서... 지속적으로 구인광고를 낸다고 하네요. 드럼세탁기는 고무가스킷과 스파이더가 가장 더럽습니다....." [3] "[민원] 그가 그를 사랑하는 방식-1 W. 능소니 "전원우…." 민규가 급하게 문을 열고... 원우와 자신의 교복과 옷을 세탁기에 넣고 빨고 아까부터 신경 쓰였던 설거지를 하고..." [4] "기다려주세요." 세탁기 설치가 어려워 사용할 수 없다는 할머니의 고충을 들은 1472팀. 서둘러 장비를 챙기고, 민원인을 도우러 출동합니다. 이날 1472팀이 도착한 곳은..." [5] "이 정도 소음이면 늦은 밤 세탁기를 돌리면 민원이 들어올 거라빨래는 되도록 낮 시간에 돌리고 있어요. 장점만 있었으면 정말 좋았을 텐데 완벽한 건 없는 거겠죠??^^;;..." [6] "자칫, 이웃집으로부터 민원이 들어올 수 있으므로 가장 신경 써 줘야 합니 다. 그 외에도 세탁기 판넬간 닿는 부분 에서도 미세한 소음이 발생할 수 있 으므로 원천적으로..." [7] "9%, 층간소음 민원 4위청소기·세탁기 등 가전제품 소리가 198건으로 2.6% 층간소음 민원 주거 유형별아파트가 78.7% 연립주택은 11%아래층 82.5%위층13.7%옆집1.6% 층간소음..." [8] "이곳에 이사오고 나서 세탁기 돌렸는 데, 진동소음때문에 아랫집에서 민원 넣었었나봐요. 속상해 하시더라구요. ..... 그래서 이번에는 세탁기분해청소외에 이러한..." [9] "민원이 얼마나 많았던건지 모르겠지만.. 세탁기 통을 첨부터 새거로 가져와서 교체해주더라구요. 과거에 계속 통돌이를 고치고고치고해도 민원이 없어지지않았고...." [10] "세탁기 소음으로 인해 아랫집에서 민원이 들어왔다고 하면서 드럼세탁기 베어링 수리를 문의하신 고객님 댁은 광명시의 한 아파트입니다. 베어링이 손상되면 소음이 상당히..."

● 세탁기 민원 데이터 시각화

1. 세탁기 민원 데이터 수집

③ R 프로그램 민원 데이터 분석

- R 프로그램에서 Sort 명령어를 통해 50개의 데이터를 내림차순으로 정렬함

```
> head(sort(wordcount, decreasing=T),50)
data4
    세탁기      민원    에어컨    냉장고      처리      관리      발생
       930       793       250       202       128       118       114
      옷장      해결      시간        TV      옵션      하게      빨래
       102       100        95        91        90        90        89
   싱크대      드럼      하기      생활      사용      소음      침대
        88        85        78        76        75        75        71
      해서      소리      원룸      접수      각종 서비스 가스레인지
        70        65        64        62        61        61        60
      불편      베란다      인덕      교체      신속      청소      설치
        60        59        59        58        58        55        52
  화장실      씽크대      사항 인터넷      제공      문제      가구
        52        51        50        50        50        49        45
 가전제품      때문      제품 아파트      별도      운영 민원(고장
        43        43        43        42        41        41        40
  붙박이장
        40
```

④ JSON데이터 만들기

- DB에 넣는 작업을 하지 않고 직접 빅데이터로 분석한 연관어의 데이터와 수치를 minwon.json으로 만듦

```
[{
        "Letter": "드럼",
        "Freq": 85      },
{
        "Letter" : "소음",
        "Freq": 75},
{
        "Letter" : "원룸",
        "Freq": 64},
{
        "Letter" : "베란다",
        "Freq": 59},
{
        "Letter" : "청소",
        "Freq" : 58},
{
        "Letter" : "설치",
        "Freq" : 55},
{
        "Letter" : "화장실",
        "Freq" : 52}]
```

2. 세탁기 민원 데이터 시각화 실습

◆ 세탁기 민원 데이터 시각화 실습 순서

① CSS 설정하기

② 캔버스 크기 및 그래프 간격 설정하기

③ 축의 크기 설정 및 SVG 요소 추가하기

④ 데이터 로딩 및 데이터 범위 설정하기

⑤ 축의 스타일 지정하기

⑥ 바그래프 속성 설정하기

⑦ 결과화면 확인하기

핵심요약

1. 실시간 데이터 수집

- **데이터 로딩함수**
  - 데이터는 csv, json, text, xml 등의 다양한 형식으로 존재함
  - D3는 이러한 파일들을 비동기(Async Request)로 가져오는 함수를 제공함
- **API를 제공하는 방식**
  - 구글, 네이버, 애플, 페이스북 처럼 외부에서 데이터를 사용하도록 공개하는 사이트들은 각자의 API를 제공함
  - API를 통해 데이터를 입력 · 이용하려는 대상을 인증하기 위해서 API를 제공할 때에는 키(Key)가 필요함
  - 대체로 널리 쓰이는 방법이 공개키(Public Key)와 개인키(Private Key) 한 쌍을 발급하고 이를 통해 대상을 인증하는 방식임
- **데이터베이스 구성 - SQL(관계형 데이터베이스)**
  - 장점
    - 범용성 : 다양한 용도로 사용
    - 고성능 : 일반적으로 높은 성능
    - 데이터의 일관성 보증
    - 정규화에 따른 갱신 비용 최소화
  - 단점
    - 대량의 데이터 입력 처리 어려움
    - 갱신이 발생한 테이블의 인덱스 생성 및 스키마 변경, 컬럼의 확장의 어려움
    - 단순히 빠름
  - 데이터 사이에 관계가 존재하면 SQL 데이터베이스를 사용함
  - 데이터 사이에 관계가 필요 없으면 NoSQL 데이터베이스를 사용함

핵심요약

1. 실시간 데이터 수집

- **데이터베이스 구성 - NoSQL**
  - SQL을 사용하지 않는다는 의미
  - Not Only SQL : SQL이 필요 없다는 의미가 아니고, 개선 / 보안의 의미
  - Non-Relational Operational Database SQL : 관계형 데이터베이스가 아님
  - NoSQL의 장점
    - 대용량 데이터
    - 데이터 분산 처리
    - Cloud Computing
    - 빠른 읽기 / 쓰기 속도
    - 유연한 데이터 모델링
- **반정형 데이터의 특징**
  - 정형 데이터는 엑셀 데이터관리 처럼 데이터의 스키마 정보를 관리하는 DBMS와 데이터 내용이 저장되는 데이터 저장소로 구분됨
  - 반정형 데이터는 데이터 내부에 정형데이터의 스키마에 해당되는 메타데이터를 갖고 있으며, 일반적으로 파일 형태로 저장됨
  - 반정형 데이터의 예
    - URL 형태로 존재 - HTML
    - 오픈 API 형태로 제공 - XML, JSON
    - 로그형태 - 웹로그, IoT에서 제공하는 센서 데이터

핵심요약

1. 실시간 데이터 수집

- **수집방법의 분류**
  - 일반적 수집 데이터의 형태와 종류에 따라 크롤링, ETL(데이터수집프로그램), 로그수집, ftp, http, RDB 수집방법으로 분류함
  - 수집 데이터의 형태에 따라 물리적으로 저장형태가 분류되고, 연동방법에 따라 수집방법이 결정됨

    ① 원본 데이터 요청 후 확인

    ② 소켓(Socket) 통신으로 연동하는 DBMS 수집방법

    ③ 스트리밍 방식으로 연동하는 로그 데이터, 센서 데이터 수집방법

    ④ ftp 프로토콜을 사용하는 이진 파일 수집방법

    ⑤ http 프로토콜 사용하는 스크립트 파일 수집방법

핵심요약

2. 자동차 민원 데이터 시각화

- **PHP를 이용하여 DB를 JSON으로 출력**

① MySQL DB 설정

② DB 내용을 PHP를 이용해 JSON으로 출력하기 위해 테이블을 만듦

- 데이터 값은 인터넷을 통해 데이터를 얻음

③ PHP 한글 코드 깨짐 방지

- 한글은 urlencode() method를 이용해 한번 감싸주고 나중에 json으로 변환 할 때 urldecode() method를 이용해 디코드함
- 숫자는 SQL 응답 내용을 PHP Array로 저장 할 때 앞 부분에 (int)를 써주어서 꼭 int형으로 저장함

④ PHP DB 호출 코드

- 자신이 호스팅한 사이트에 아이디와 패스워드 DB이름을 넣음

⑤ 데이터베이스 연결

- 데이터 베이스 연결은 MySQL에 설정된 데이터베이스를 호출하는 것
- 쿼리 실행은 DB 값을 가지고 오기 위해 사용하는 명령어

⑥ JSON 호출

- 배열을 이용해 cars 에 있는 데이터를 json으로 변환하여 호출

핵심요약

3. 세탁기 민원 데이터 시각화

- **세탁기 민원 데이터 수집**
  - ① 포털사이트에서 '세탁기 민원' 데이터 검색
  - ② 민원데이터 수집
  - ③ R 프로그램 민원 데이터 분석
  - ④ JSON데이터 만들기
    - DB에 넣는 작업을 하지 않고 직접 빅데이터로 분석한 연관어의 데이터와 수치를 minwon.json으로 만듦