JSON

- 서버와 웹 앱 사이의 데이터 전송을 위해 사용하는 데이터 형식

-

- Object : 중괄호를 이용하여 감싸고 있음

- name은 string으로 되어 있고, name : value

- Value는 큰따옴표

- {object} / name : value / [array]

a. Jackson

* 코드가 심플하고 고용량(100MB 이상)의 JSON 데이터 처리 성능이 탁월
* 패키지가 무거운편

**b. Google-gson**

* 소스코드, 레퍼런스 찾기 용이
* 가벼운 JSON데이터 처리 성능이 탁월하고 전반적으로 고성능
* 상대적으로 가벼운 jar속에 하나의 패키지로 구성

c. JSON-lib

* 더글라스 크록포트, 디펜던시 다수 존재

d. Flexjson

* 자바 객체를 JSON으로 직렬화하거나 비직렬화 할 수 있는 경량 라이브러리
* 다른 외부 라이브러리와 의존성이 없음
* 성능은 무난하나 아주 적은 패키지 사이즈가 장점

e. Json-io

* JsonReader, JsonWriter 두 클래스로 구성 -> 직렬화 담당하는 stream객체가 필요 없음
* JDK의 ObjectInputStream, ObjectOutputStream보다 빠른 직렬화 성능을 제공

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Python \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

데이터가 1만건 정도라면 Python, 10만개 이상이라면 R

무언가를 만든다면 Python, 통계적 분석 -> R

Data Cleaning(가공하지 않은 데이터를 정리해서 가공)면에서는 Python

Data Exploration(데이터 탐색)의 경우 파이썬이 더 유리(pandas package)

<http://krksap.tistory.com/523>

R  
 - 연구목적 최적화  
 - 방대한 양의 패키지와 즉시 사용 가능한 테스트 SET 제공  
 - 데이터 시각화(ggplot2, ggvis, googleVis, rCharts)  
 - 생태계,   
 - 느린 속도(읽기 어려운 코드의 경우)  
 - 어려움

Python  
 - 데이터 분석 결과를 다른 웹 앱과 접목  
 - 통계적인 코드를 DB에 포함하는 경우  
 - IPython Notebook

<http://media.fastcampus.co.kr/knowledge/dataanalysis-python-r/>

웹 데이터 크롤링(Web Crawling, Web Scraping, Spider)

* “XML”, “stringr”(문자열 정리) 패키지가 필요

통계프로그램 R( RStudio IDE)

* 통계분석 패키지 프로그램은 SAS, SPSS, MINITAP 이렇게 3가지가 존재
* SAS의 경우 데이터 사이언스, 심화통계 CRM
* SPSS는 안에 Syntax를 이용하여 프로그래밍, 리서치 회사(빈도 분석)
* Minitap은 품질관리를 통한 6시그마
* R은 “무료”프로그램

서버

* 웹서버(80)
* 데이터베이스 서버
* DNS서버(1023, 53)
* FTP서버(21)
* Ssh 서버(22)
* 메일 서버(110, 25, 143)

1. 원하는 웹 페이지에 request를 보내 결과 html을 받음

2. 받은 html을 파싱 한다.

3. 필요한 정보만 추출한다.

a. requests 모듈 (pip install requests)

파이썬에 내장된 urllib 모듈을 편하게 사용하기 위해 만든 것

b. Beautiful soup  
 많이 쓰이는 파이썬용 파서로 html, xml을 파싱할 때 주로 사용

<https://medium.com/@mjhans83/%ED%8C%8C%EC%9D%B4%EC%8D%AC%EC%9C%BC%EB%A1%9C-%ED%81%AC%EB%A1%A4%EB%A7%81-%ED%95%98%EA%B8%B0-908e78ee09e0>

<http://yoonpunk.tistory.com/4>

<http://itissmart.tistory.com/52>

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Java \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Jaunt Library, Jsoup

* Web-Scraping & Web-Automation, including JSON

Jsoup

* 기본적으로 HTML 형식의 string을 넘겨주면 자바에서 사용할 수 있는 DOM 객체로 만들어 주는 Parser(웹페이지 읽어오는 기능 제외, AsyncHttpClient로 읽어온 후 변환)
* Navigating과 관련된 API는 크게 2가지
* 전통적인 방식의 Navigating api  
  getElementById(), getElementsByTag()… 와 같은 메쏘드를 이용하여 특정 Element를 찾아 낼 수 있음

Jaunt의 무료버전은 Jsoup의 MIT License에 비하여 더욱 자유롭다. MIT License의 경우 모든 코드에 MIT copyright/permission notice를 적어 넣어야 함

Two-year version의 경우 jar파일을 재분배해주지 않기 때문에 MIT License보다 더욱 제한적이다. 이것은 서버쪽 프로젝트에 적합하다 그리고 어떤 개인적 프로젝트에 분산될 수 없다. ($24)

기업형 버전의 경우 사업 라이선스다. ($950)

무료 버전의 경우 각 달의 마지막에 만료가 되고 다시 다운 받아야 함

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Jsoup \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Document: Jsoup얻어온 결과 HTML 전체 문서

Element Document의 HTML요소

Elements Element가 모인 자료형 for, while등의 반복문을 사용 할 수 있음

Connection: Jsoup의 connect혹은 설정 메소드들을 이용해 만들어지는 객체, 연결을 하기 위한 정보를 담고 있다.

Response: Jsoup가 URL에 접속해 얻어온 결과. Document와 다르게 status 코드, status 메시지나 charset같은 헤더 메시지와 쿠키등을 갖고 있다.