# Fastcampus

## Data Science SCHOOL with R

## Scraping with Google Sheets

# Introduce

## 최우영

- Co-founder, Developer at disceptio
- Solution Architect, Web Developer, Instructor
- python web crawling bootcamp(gilbut, 2018 expected)
- Skills: Python, Golang, Julia, Node.js, Google tag manager ...

blog: https://blog.ulgoon.com/

github: https://github.com/ulgoon/

email: me@ulgoon.com

Google Sheets

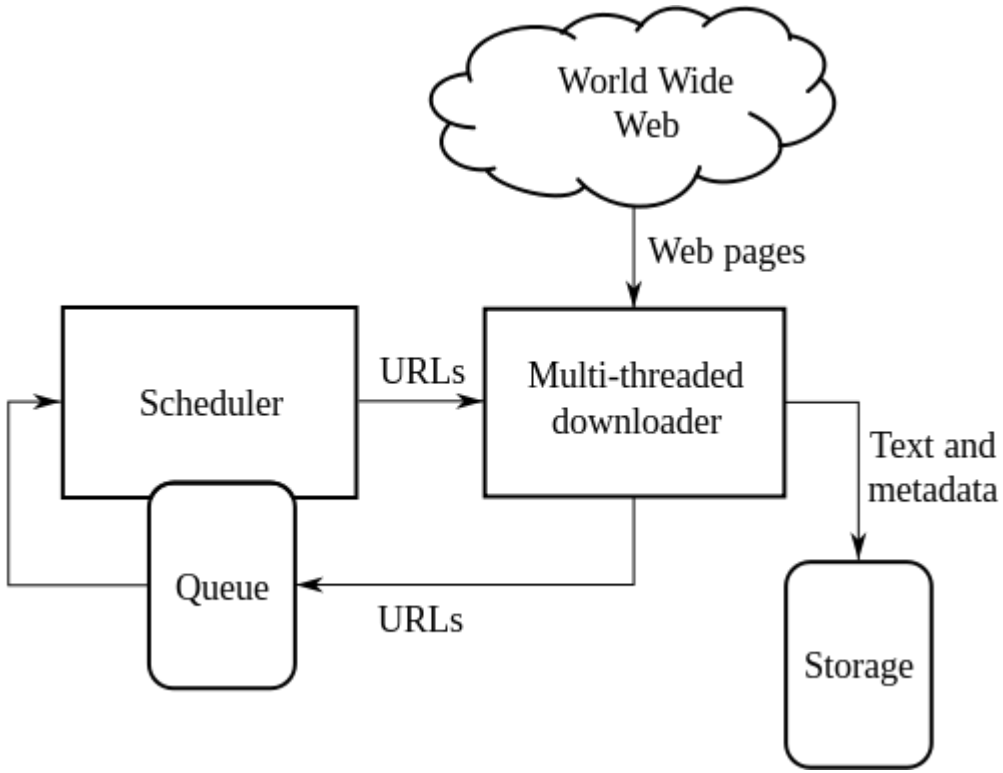
# Web Scraping with Google Sheets

https://docs.google.com/spreadsheets/d/1kVTI157__2Etw4vp1EmWIEOwVSnC0BzCKZOiHY1jzTI/edit?usp=sharing

# Crawling, Scraping, Parsing

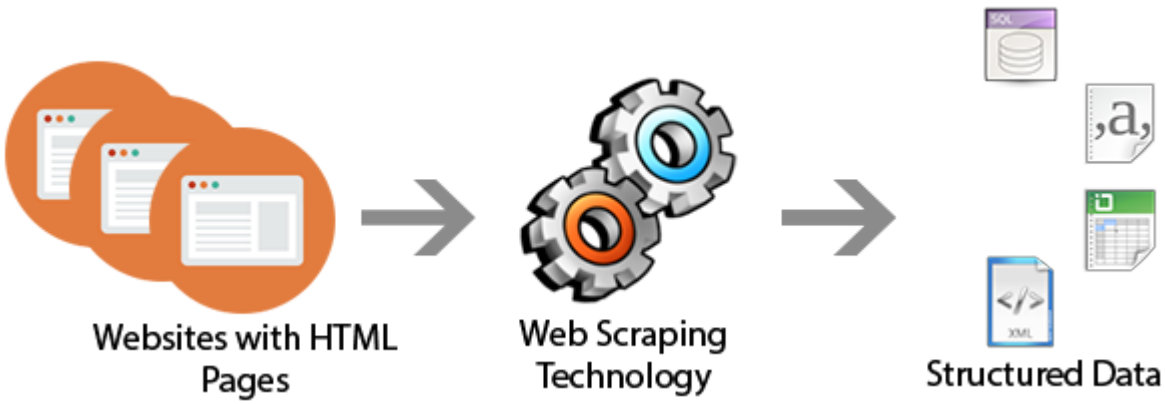
# Crawling

Crawling: 조직적 자동화된 방법으로 월드 와이드 웹을 탐색하는 것



# Scraping

Scraping: 데이터를 수집하는 행위



## Parsing

Parsing: 문장 혹은 문서를 구성 성분으로 분해하고 위계관계를 분석하여 문장의 구조를 결정하는 것

SCRAP with Google sheets

# Caution!!

저작권 침해 위반 소지

- 웹사이트 운영자의 크롤링 금지 룰을 어길경우
- 월권하여 데이터베이스에 접근
- 타인의 경제적 이익을 침해할 경우
- 개인정보를 수집할 경우(전화번호, 주소, ..)

## Google Sheets functions

- CONCATENATE(): 여러 문자열을 합쳐 하나의 문자열로 표현할 때 사용
- TODAY(): 오늘 날짜를 (MM/DD/YYYY) 형태로 표현
- NOW(): 오늘 날짜와 시간을 MM/DD/YYYY HH:MM:SS로 표현
- SPLIT(text, delimiter): delimiter를 기준으로 text를 나눌 때 사용

# IMPORTFEED

> RSS 또는 Atom 피드를 가져옵니다.

`IMPORTFEED(URL, [쿼리], [헤더], [항목_개수])`

- URL - RSS 또는 Atom 피드의 URL로 프로토콜(예: http://)을 포함
- "feed"는 제목, 설명 및 URL 등의 피드 정보를 포함하는 하나의 행을 반환
    - "feed "은 피드의 특정 속성을 반환 은 제목, 설명, 작성자 또는 URL 중 하나
- "items"는 피드의 항목을 포함하는 표 전체를 반환
    - 항목_개수를 지정하지 않을 경우 현재 피드에 게재된 모든 항목이 반환
    - "items "은 요청한 항목의 특정 속성을 반환

# IMPORTFEED

> RSS 또는 Atom 피드를 가져옵니다.

`IMPORTFEED(URL, [쿼리], [헤더], [항목_개수])`

- 헤더: [ 선택사항 - 기본값은 FALSE] - 반환된 값의 상단에 행을 추가해 열 헤더를 포함할지 여부
- 항목_개수: [ 선택사항 ] - 항목의 검색어에 대해 가장 최근 순서로 반환할 항목의 개수
- 항목_개수를 지정하지 않을 경우 현재 피드에 게재된 모든 항목이 반환

## Get Articles

https://www.theguardian.com/technology/mobilephones/

## Get Podcast Feed

http://getrssfeed.com/

https://soundcloud.com/xsfm

# IMPORTDATA

> 웹에 게재된 csv, tsv 등의 파일 데이터를 가져옵니다.

`IMPORTDATA(url)`

- url: 웹에 게재된 데이터 파일의 고유주소

# Let's get theatre data

https://ulgoon.github.io/file/theatre_20180403.csv

# IMPORTHTML

> HTML 페이지에서 표 또는 목록에 있는 데이터를 가져옵니다.

`IMPORTHTML(url, query, index)`

- url: 검토할 페이지의 URL이며 프로토콜(예: http://)을 포함
- query: 원하는 데이터가 어떤 구조에 포함되었는지에 따라 '목록' 또는 '표'
    - ul, ol, dl, table, ..
- index: 여러개의 요소가 존재할 때 선택할 요소의 순서

Let's get 2018 Russia WorldCup Table Data

https://en.wikipedia.org/wiki/FIFA_World_Cup

Rotten Tomatoes TOP BOX OFFICE

https://www.rottentomatoes.com/

# IMPORTXML

> XML, HTML, CSV, TSV, RSS 및 Atom XML 피드를 포함한 다양한 구조화된 데이터로부터 데이터를 가져옵니다.

`IMPORTXML(url, xpath_검색어)`

- url: 검토할 페이지의 URL이며 프로토콜(예: http://)을 포함
    - url 값은 따옴표로 묶거나, 적절한 텍스트를 포함하는 셀에 대한 참조
- xpath_검색어: 구조화된 데이터에서 실행되는 XPath 검색어
    - https://www.w3schools.com/xml/xpath_intro.asp

Popular news in media Daum

http://media.daum.net/

# IMPORTRANGE

> 지정된 스프레드시트에서 셀 범위를 가져옵니다.

`IMPORTRANGE(spreadsheet_key, range_string)`

spreadsheet_key: 가져올 데이터가 있는 스프레드시트의 URL
range_string - " `sheet_name` ! `range` " 형식의 문자열이며 가져올 범위를 지정

# Let's make Currency Report

Let's make Currency Report

http://finance.daum.net/exchange/exchangeMain.daum?nil_profile=stockgnb&nil_menu=exchange_top

https://finance.yahoo.com/quote/KRW=X?p=KRW=X