

M3.

크롤링을 위한 파이썬 라이브러리

강사 프로필



김 호 성

| 학력사항

- 상명대학교 경영학 학사
- 상명대학교 빅데이터과학 학사

| 경력사항 (프로젝트 등)

- 데이터마케팅코리아 데이터엔지니어, 프론트엔드 개발자
- 텍스트 분석솔루션 FoxTA 2.0 개발 담당
- AI 마케팅 분석 솔루션(Ma:deri) 프론트엔드, 백엔드 개발
- 한화투자증권 텍스트분석 분류체계 개발 지원
- 솔루션 Gamification 및 DPI(Digital Power Index) 기획 참여
- 공공사업, 컨설팅 데이터 시각화 담당(QlikSense, R, Python)

| 강의 경력사항

- 예술경영지원센터 크롤링 강의 지원

커리큘럼

과정 목표

- 01 온라인(웹) 상의 다양한 데이터를 수집하기 위해 데이터의 형태와 웹의 기본구조를 이해한다.
- 02 크롤링을 위해 파이썬의 다양한 라이브러리의 기능과 역할을 학습한다.
- 03 실제 온라인에서 제공되는 다양한 형태의 데이터들을 원하는 방식으로 수집할 수 있도록 코드를 작성한다.

모듈 명	주요 내용
M1. 데이터의 형태	<ul style="list-style-type: none">• 데이터를 정의하고, 강의에서 활용할 데이터에 대한 유형을 소개• JSON, CSV, TSV, XLSX, 등 데이터의 형태를 소개하고 간단한 활용법 학습• Database에 대한 소개 및 SQLite를 활용한 간단한 활용법 숙지
M2. 웹의 기본구조	<ul style="list-style-type: none">• 웹의 개념과 간단한 역사 및 작동원리에 대한 내용을 파악• 실제 다양한 웹페이지를 통해 웹을 구성하고 있는 요소(HTML, Javascript, JSON 등)들을 학습
M3. 크롤링을 위한 파이썬 라이브러리	<ul style="list-style-type: none">• 크롤링을 위한 다양한 파이썬 라이브러리에 대한 소개• 주피터노트북을 활용하여 각 라이브러리에 대한 실습 진행
M4. 웹페이지를 활용한 데이터 수집	<ul style="list-style-type: none">• 실제 동작하는 웹페이지를 분석하고 수집하여, 원하는 형태의 데이터로 변환하는 실습 진행• 대상 : 네이버금융, 네이버뉴스, DART, 증권사이트 등
M5. API를 활용한 데이터 수집	<ul style="list-style-type: none">• 공공기관이나 민간기업에서 제공하고있는 OpenAPI를 활용하여 데이터를 수집하는 실습 진행• 대상 : 공공데이터(부동산, 금융), SNS API, 쇼핑몰 API 등
M6. 데이터 수집 시스템 만들기	<ul style="list-style-type: none">• 위에서 작성된 파이썬 코드를 활용하여, 정기적인 수집 시스템을 만드는 실습 진행

*상기 커리큘럼의 내용 및 교육시간은 기관 및 교육과정의 사정에 따라 변동이 가능합니다.

M3 세부 목차

M3. 크롤링을 위한 파이썬 라이브러리

M4를 위한 API KEY발급

Jupyter 설치 및 크롤링 유의사항

urllib 기본 사용 및 예제

requests 및 BeautifulSoup 사용

Selenium을 통한 동적 크롤링

M3 목차

Jupyter 설치 및 크롤링 유의사항

cmd 명령 프롬프트 - pip3 install jupyter

```
Microsoft Windows [Version 10.0.17134.590]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\User>pip3 install jupyter
```

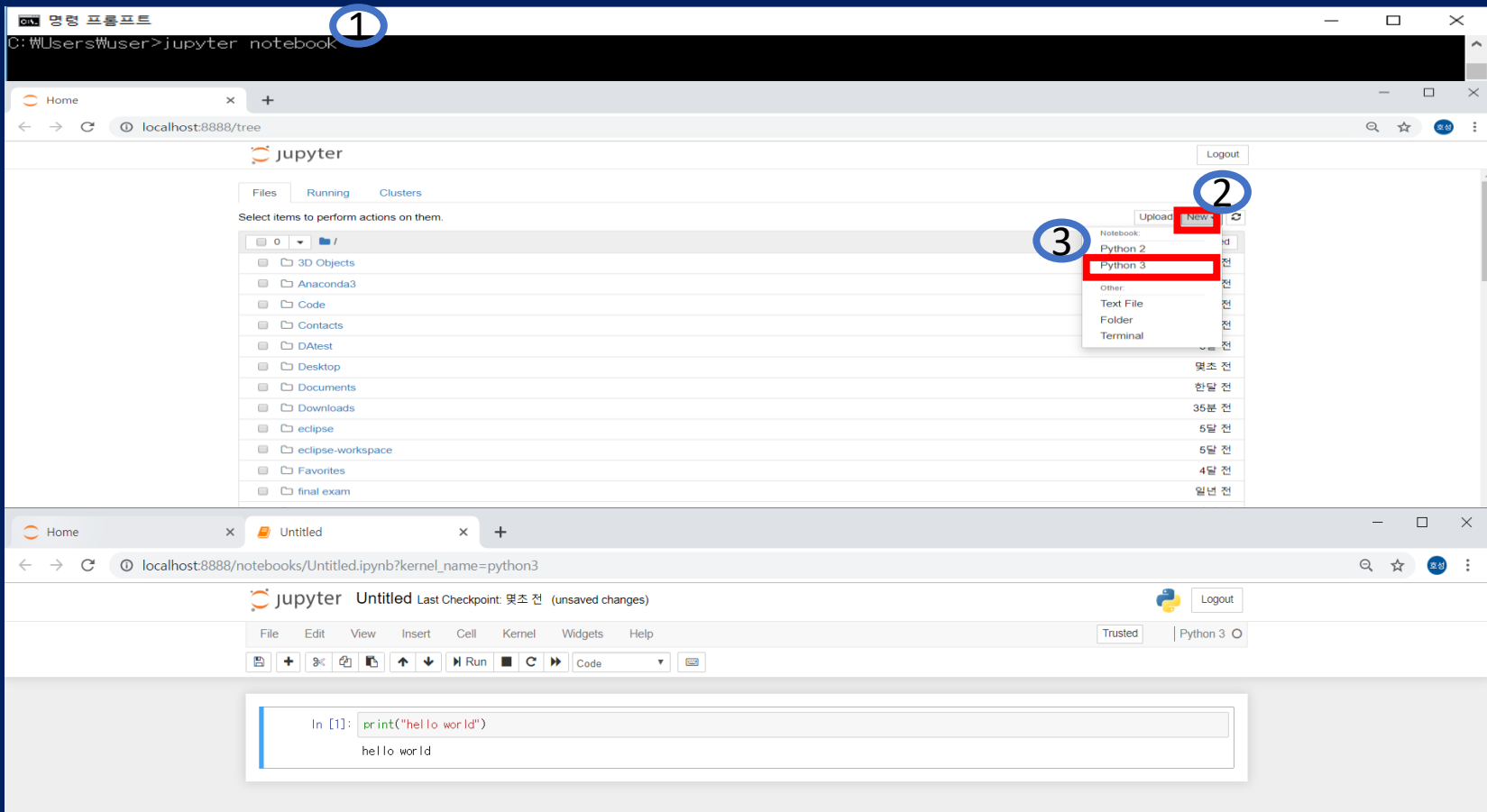
cmd 명령 프롬프트

```
Collecting MarkupSafe>=0.23 (from Jinja2->notebook->jupyter)
  Downloading https://files.pythonhosted.org/packages/5b/d4/1deb3c5dc3714fb160c7e2116fc6dff36a063d9156a9328cce54ef35cc52/MarkupSafe-1.1.1-cp37-cp37m-win32.whl
Collecting jsonschema!>=2.5.0,<=2.4 (from nbformat->notebook->jupyter)
  Downloading https://files.pythonhosted.org/packages/aa/69/df679dfbdd051568b53c38ec8152a3ab6bc533434fc7ed11ab034bf5e82f/jsonschema-3.0.1-py2.py3-none-any.whl (54kB)
100% |#####| 61kB 2.6MB/s
Collecting pywinpty>=0.5; os_name == "nt" (from terminado>=0.8.1->notebook->jupyter)
  Downloading https://files.pythonhosted.org/packages/4a/20/4ab8f904a95048b781ecb00b56142b10388ae66728e8193c2dc712d13c55/pywinpty-0.5.5-cp37-cp37m-win32.whl (1.3MB)
100% |#####| 1.3MB 4.5MB/s
Collecting wcwidth (from prompt-toolkit<2.1.0,>=2.0.0->jupyter-console->jupyter)
  Downloading https://files.pythonhosted.org/packages/a7/bd/e2f4753c5fa93932899243b4299011a757ac212e9bc8ddf062f38df4e78b/wcwidth-0.1.7-py2.py3-none-any.whl
Collecting webencodings (from bleach->nbconvert->jupyter)
  Downloading https://files.pythonhosted.org/packages/f4/24/2a3e3df732393fed8b3ebf2ec078f05546de641fe1b667ee316ec1dcf3b7/webencodings-0.5.1-py2.py3-none-any.whl
Collecting parso>=0.3.0 (from jedi>=0.10->ipython>=5.0.0->ipykernel->jupyter)
  Downloading https://files.pythonhosted.org/packages/2c/a7/8a50738eb27e204aa271abe170dec7bdeb07128ed892fb3a92f14a69bae3/pyrsistent-0.15.1.tar.gz (10kB)
100% |#####| 102kB 4.3MB/s
Collecting attrs>=17.4.0 (from jsonschema!>=2.5.0,<=2.4->nbformat->notebook->jupyter)
  Downloading https://files.pythonhosted.org/packages/23/96/d828354fa2dbdf216eaa7b7de0db692f12c234f7ef888cc14980ef40d1d2/attrs-19.1.0-py2.py3-none-any.whl
Collecting pyrsistent>=0.14.0 (from jsonschema!>=2.5.0,<=2.4->nbformat->notebook->jupyter)
  Downloading https://files.pythonhosted.org/packages/2c/a7/8a50738eb27e204aa271abe170dec7bdeb07128ed892fb3a92f14a69bae3/pyrsistent-0.15.1.tar.gz (10kB)
100% |#####| 112kB 5.1MB/s
Installing collected packages: tornado, six, decorator, ipython-genutils, traitlets, python-dateutil, jupyter-core, pyzmq, jupyter-client, pickleshare, parso, jedi, wcwidth, prompt-toolkit, colorama, backcall, pygments, ipython, ipykernel, MarkupSafe, Jinja2, pandocfilters, testpath, entrypoints, attrs, pyrsistent, jsonschema, nbformat, mistune, webencodings, bleach, defusedxml, nbconvert, Send2Trash, prometheus-client, pywinpty, terminado, notebook, widgetsnbextension, ipywidgets, jupyter-console, qtconsole, jupyter
Running setup.py install for backcall ... done
Running setup.py install for pandocfilters ... done
Running setup.py install for pyrsistent ... done
Running setup.py install for prometheus-client ... done
Successfully installed MarkupSafe-1.1.1 Send2Trash-1.5.0 attrs-19.1.0 backcall-0.1.0 bleach-3.1.0 colorama-0.4.1 decorator-4.4.0 defusedxml-0.6.0 entrypoints-0.3 ipykernel-5.1.0 ipython-7.5.0 ipython-genutils-0.2.0 ipywidgets-7.4.2 jedi-0.13.3 Jinja2-2.10.1 jsonschema-3.0.1 jupyter-1.0.0 jupyter-client-5.2.4 jupyter-console-6.0.0 jupyter-core-4.4.0 mistune-0.8.4 nbconvert-5.5.0 nbformat-4.4.0 notebook-5.7.8 pandocfilters-1.4.2 parso-0.4.0 pickleshare-0.7.5 prometheus-client-0.6.0 prompt-toolkit-2.0.9 pygments-2.3.1 pyrsistent-0.15.1 python-dateutil-2.8.0 pywinpty-0.5.5 pyzmq-18.0.1 qtconsole-4.4.4 six-1.12.0 terminado-0.8.2 testpath-0.4.2 tornado-6.0.2 traitlets-4.3.2 wcwidth-0.1.7 webencodings-0.5.1 widgetsnbextension-3.4.2
You are using pip version 19.0.3, however version 19.1.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.

C:\Users\User>
```

jupyter notebook설치

1. 실행창 > cmd (명령프롬프트) 실행
2. cmd창에 `pip3 install jupyter` 입력
3. 설치가 완료되면 다음 명령을 입력 받을 준비를 함



jupyter notebook설치

- 1.cmd창에 **jupyter notebook** 입력
2. jupyter notebook이 열리면 New버튼을 클릭
- 3.New버튼을 클릭한 뒤 나오는 메뉴에서 Python3선택
4. 명령어를 입력한 뒤 Ctrl+enter키를 눌러 제대로 작동 되는지 확인



```
[1] import requests

[3] res = requests.get('http://www.naver.com')

res.text
```

'<!doctype html> <html lang="ko" data-dark="false"> <head> <meta charset="utf-8"> <title>NAVER</title> <meta http-equiv="X-UA-Compatible" content="IE=edge"> <meta name="viewport" content="width=1190"> <meta name="apple-mobile-web-app-title" content="NAVER"/> <meta name="robots" content="index,nofollow"/> <meta name="description" content="네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요"/> <meta property="og:tit le" content="네이버"> <meta property="og:url" content="https://www.naver.com/"> <meta property="og:image" content="https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png"> <meta property="og:descrip tion" content="네이버 메인에서 다양한 정보와 유용한 콘텐츠를 만나 보세요"/> <meta name="twitter:card" content="summary"> <meta name="twitter:title" content=""> <meta name="twitter:url" content="https://www.naver.com/"> <me ta name="twitter:image" content="https://s.pstatic.net/static/www/mobile/edit/2016/0705/mobile_212852414260.png"> <meta name="twitter:description" content="네이버 메인에서 ...'

구글 Colab

1. google 검색창에 colab을 검색

*웹 브라우저 크롬 권장

2. jupyter notebook과 같은 기능을 할 수 있는 것으로 구글계정만 있다면 따로 설치할 필요가 없음

3. 사용 목적에 따라서 jupyter notebook보다 좋을 수 있으나 강의에서는 jupyter를 기준으로 강의하므로 권장하지는 않음

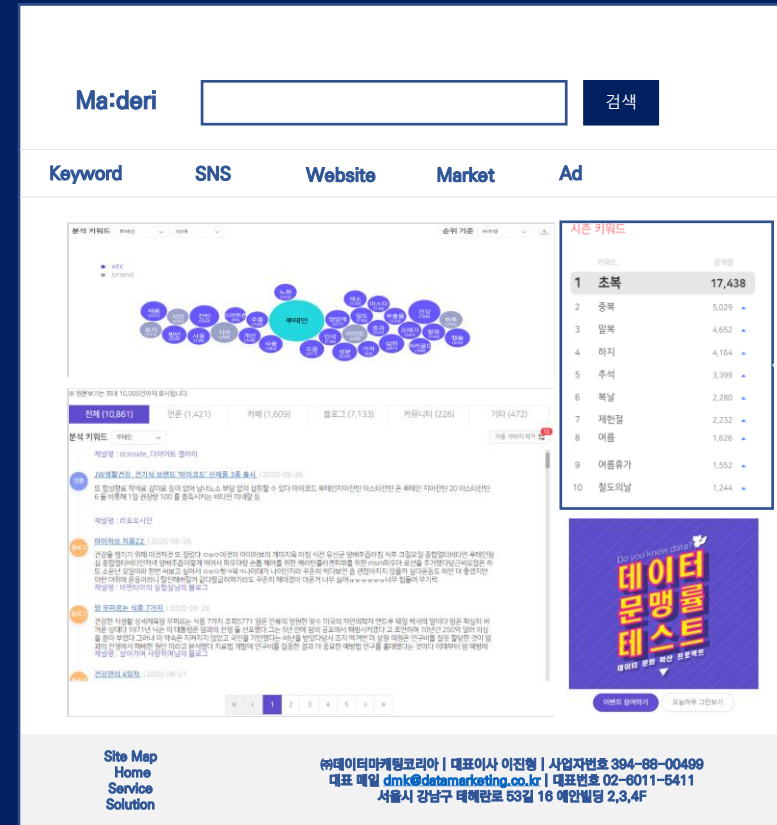
M3

실습

urllib 기본 사용 및 예제

데이터 예시

```
response = {
  "data": [
    { "rank": 1, "keyword": "초복", "val": 17438 },
    { "rank": 2, "keyword": "중복", "val": 5029 },
    ...
    { "rank": 10, "keyword": "철도의날", "val": 1244 }
  ]
}
```



〈maderi.xx.xx로 요청시〉

- html rendering
- 데이터가 비워진 영역이 있음

〈비동기 요청〉

- html, json 형태로 데이터를 받음
- 데이터를 부르고 내부 javascript함수로 컴포넌트 에 맞춰 나타남

M3

실습