2020 혁신성장 청년인재 집중양성 사업

빅데이터 수업 2주차

Web Crawling and Scraping

데이터 수집 1

#Python #Node.js #WebBot

© RedInfo inc.

시작하기에 앞서

"석유의 저장량이 매년 늘어나는 이유?"

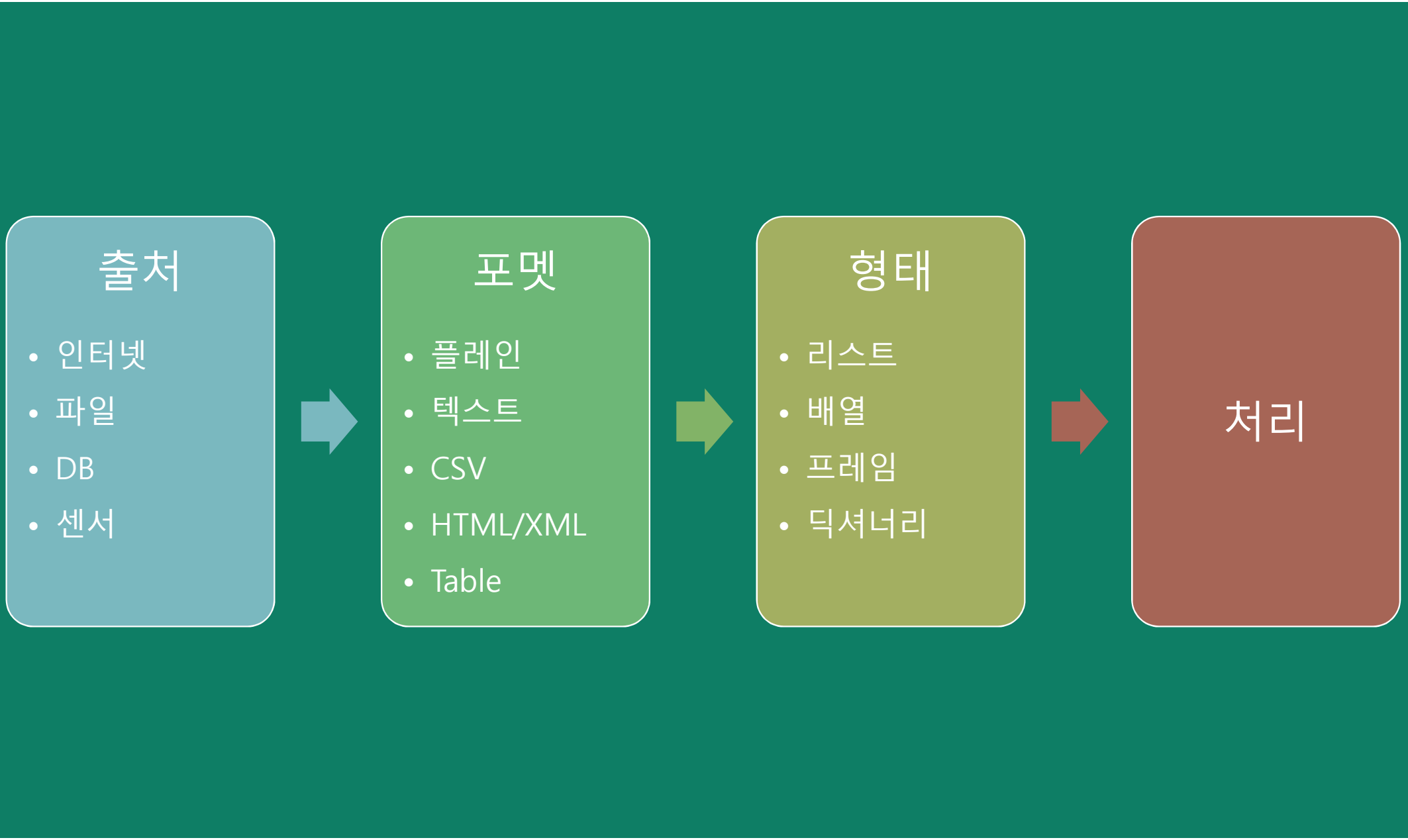
탐사
기술의
발달

시추
기술의
발달

셰일 오일

데이터 파이프라인 이란?

1. 데이터 수집에 있어 raw level의 artifact를 획득 하는 것
2. Artifact에서 데이터 추출
3. 처리에 적합한 데이터로 변환


출처
인터넷
파일
DB
센서
포멧
플레인
텍스트
CSV
HTML/XML
Table
형태
리스트
배열
프레임
딕셔너리
처리

왜 리눅스죠?

리눅스를 사용하는 이유

네이버 뉴스 헤드라인 목록을 1시간에 한번씩 yyyy_mm_dd_hh_naver_news_headline_title.txt 로 저장한다.

무엇이 필요할까요?

리눅스를 사용하는 이유

1. 1시간 마다 특정 코드를 실행해 줄 데몬
   - Crontab

2. News.naver.com에 접속하여 데이터를 가져오는 것을 수행할 코드
   - Python
     - urllib or requests(HTTP 통신-수집)
     - beautifulSoup(HTTP 추출)
     - write() 함수

3. 저장된 파일을 특정 파일로 이동
   - mv 명령어

시작

A table of Contents

Part 1

Windows 개발환경 구축

Windows 개발환경 구축

Anaconda 설치(individual Edition)

https://www.anaconda.com/products/individual

ANACONDA
Powered by Continuum Analytics

Anaconda Installers

Windows

Python 3.7
64-Bit Graphical Installer (466 MB)
32-Bit Graphical Installer (423 MB)

Python 2.7
64-Bit Graphical Installer (413 MB)
32-Bit Graphical Installer (356 MB)

MacOS

Python 3.7
64-Bit Graphical Installer (442 MB)
64-Bit Command Line Installer (430 MB)

Python 2.7
64-Bit Graphical Installer (637 MB)
64-Bit Command Line Installer (409 MB)

Linux

Python 3.7
64-Bit (x86) Installer (522 MB)
64-Bit (Power8 and Power9) Installer (276 MB)

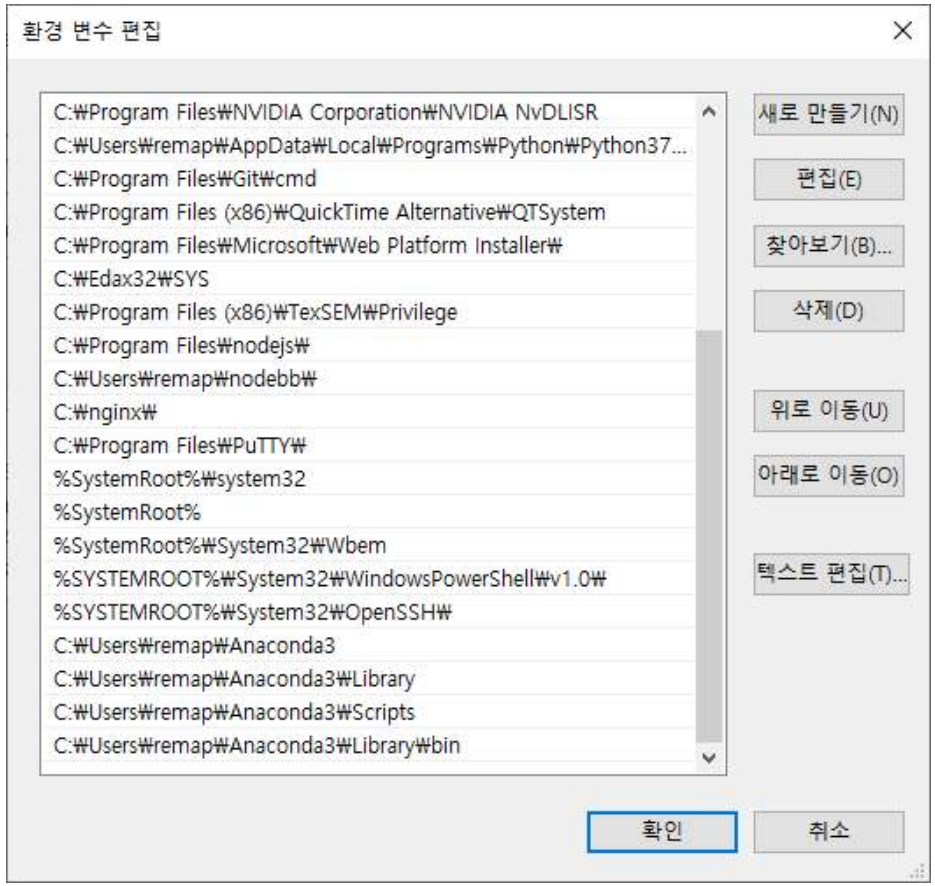
Python 2.7
64-Bit (x86) Installer (477 MB)
64-Bit (Power8 and Power9) Installer (295 MB)

Windows 개발환경 구축

1. Window키+R
2. sysdm.cpl ,3

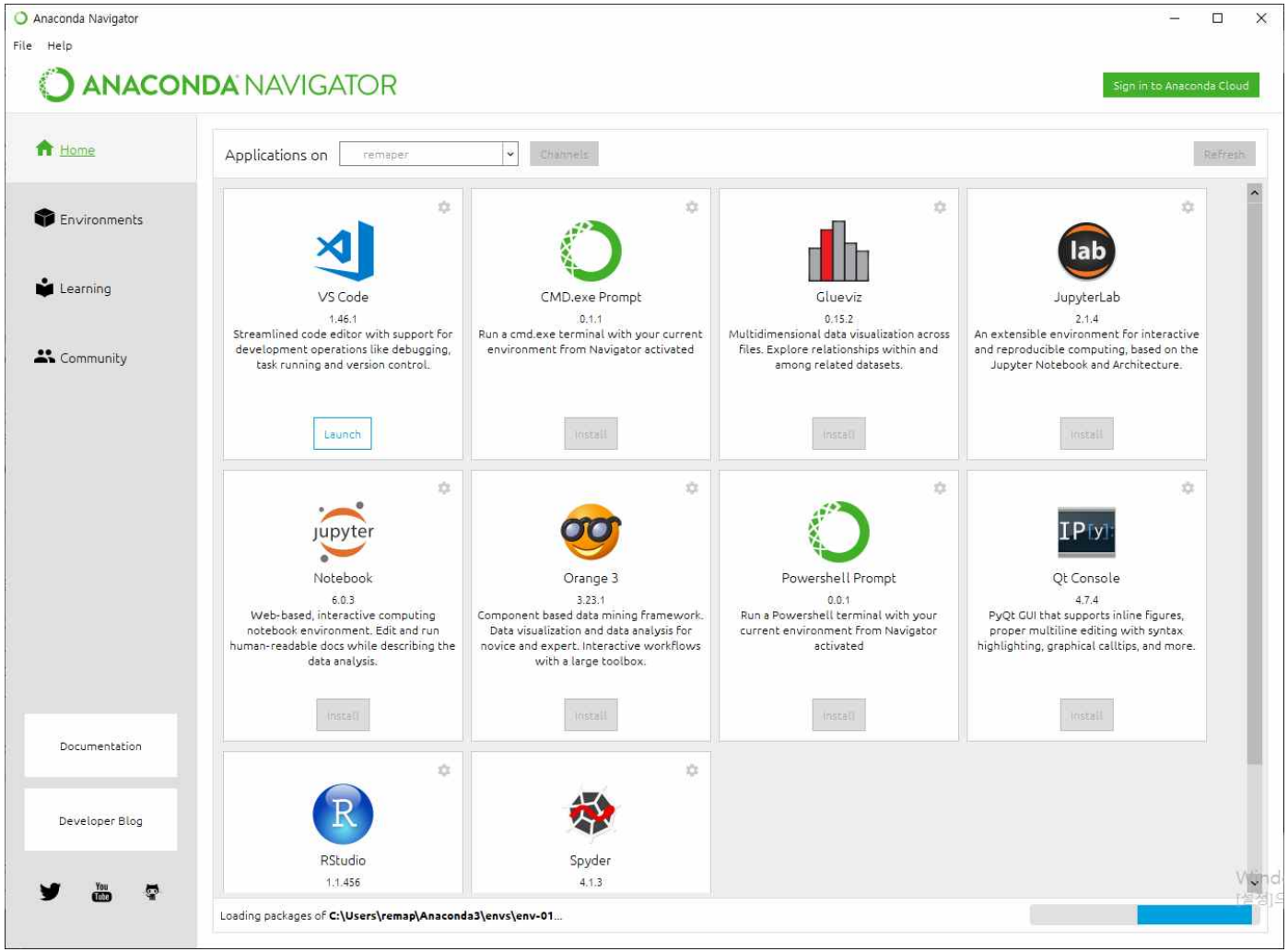
(쉼표 앞에 한 칸 띄어 써야 한다.

C:₩Users₩remap₩Anaconda3
C:₩Users₩remap₩Anaconda3₩Library
C:₩Users₩remap₩Anaconda3₩Library₩bin
C:₩Users₩remap₩Anaconda3₩Scripts



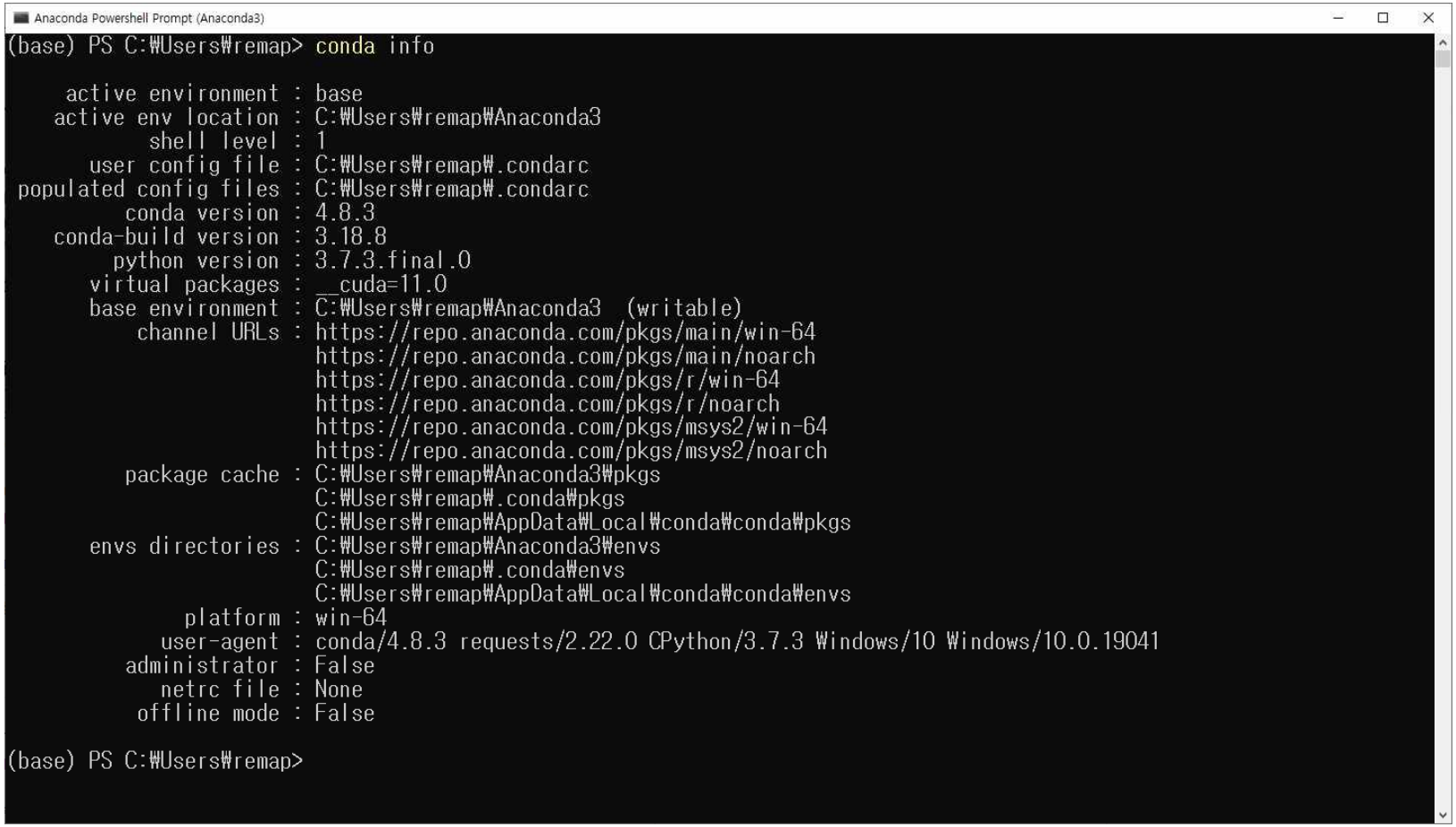
Windows 개발환경 구축

Anaconda Navigator



Windows 개발환경 구축

Anaconda powershell prompt



Windows 개발환경 구축

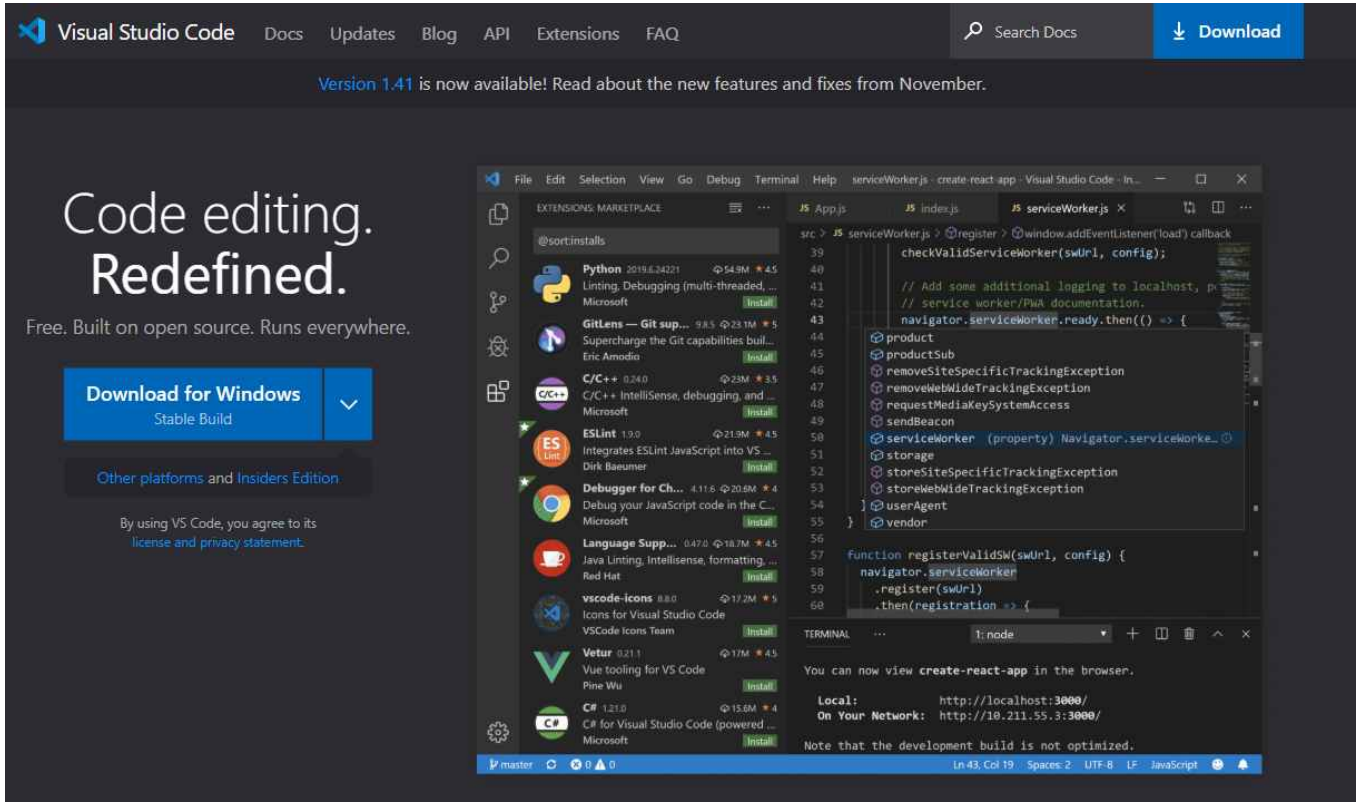
Anaconda powershell prompt

- python --version
- conda create -n env-01 python=3.7
- conda env list
- conda activate env-01
- conda deactivate

Windows 개발환경 구축

Visual Studio Code 설치

https://code.visualstudio.com/



Windows 개발환경 구축

확장프로그램 설치

- python
- python for VSCode
- Python ExtensionPack
- code Runner

python 가상환경(anaconda) setting - settings.json

```
{
    "python.pythonPath": "C:\\Users\\remap\\Anaconda3\\envs\\env-01\\python.exe"
}
```

Windows 개발환경 구축

debug setting - launch.json

```
{
  "version": "0.2.0",
  "configurations": [
    {
      "name": "Python: Current File",
      "type": "python",
      "request": "launch",
      "program": "${file}",
      "console": "integratedTerminal"
    }
  ]
}
```

Windows 개발환경 구축

task setting - tasks.json

```
{
  "version": "2.0.0",
  "tasks": [
    {
      "label": "Run project label",
      "type": "shell",
      "command": "python",
      "args": ["${file}"],
      "group": {
        "kind": "build",
        "isDefault": true
      },
      "presentation": {
        "echo": true,
        "reveal": "always",
        "panel": "new",
        "focus": true
      },
      "options": {
        "env": {
          "PYTHONIOENCODING": "UTF-8"
        }
      }
    }
  ]
}
```

Windows 개발환경 구축

pyTest.py

```
import sys
import os
import platform
world_name = 'remaper'
print('hello ' + world_name + ' world')

a = 1
b = 2
c = a+b
print(c)


print(platform.architecture())

print(os.getcwd())

print(sys.executable)
```

Part 2

Python으로 HTTP 가져오기

Python으로 HTTP 가져오기

꼭 필요한 python 기초

```
print("\t Hello, world! \t\t\n")
print("\t Hello, world! \t\t\n".lstrip())
print("\t Hello, world! \t\t\n".rstrip())
print("Hello, world! \t\t\n".strip())
print("Hello, world! \t\t\n".strip().lower().capitalize())
print("Hello, world! \t\t\n".strip().upper())
print("Hello, world! \t\t\n".strip().upper().islower())
print("Hello, world! \t\t\n".strip().upper().isupper())
print("Hello, world! \t\t\n".strip().upper().isdigit())
print("Hello, world! \t\t\n".strip().upper().isalpha())
```

Python으로 HTTP 가져오기

꼭 필요한 python 기초

```
"Hello, world!".split()
"Hello, world!".split(" ")
"www.networksciencelab.com".split(".")
", ".join(["alpha", "bravo", "charlie", "delta"])
"-".join("1.617.305.1985".split("."))
" ".join("This string\n\r has many\t\tspaces".split())
"This string\n\r has many\t\tspaces".split()
```

Python으로 HTTP 가져오기

꼭 필요한 python 기초

```
# python에서 제공해주는 문자관련 함수
"www.networksciencelab.com".find(".com")
"www.networksciencelab.com".count(".")

# 단어를 카운트 해보자1
phrase = "a man a plan a canal panama"
cntr = Counter(phrase.split())
cntr.most_common()

# 단어를 카운트 해보자2
cntrDict = dict(cntr.most_common())
cntrDict
```

Python으로 HTTP 가져오기

```
pip install urlopen
pip install datatime
pip install bs4
```

Python으로 HTTP 가져오기

```
from urllib.request import urlopen
import sys
# 1. URL을 이용하여 데이터 가져오기
f = urlopen('http://news.naver.com/')
print(type(f)) # type 확인
print(f.read()) # 읽어온 페이지 확인
print(f.status) # URL 상태 400-서버다운 500-서버오류
print(f.getheader('Content-Type')) #인코딩 확인
```

Python으로 HTTP 가져오기

```
from urllib.request import urlopen

# 2. URL에서 가져온 데이터를 텍스트화 하기
encoding = f.info().get_content_charset(failobj="utf-8")
print('encoding:', encoding, file=sys.stderr)
text = f.read().decode(encoding)
```

Python으로 HTTP 가져오기

```
from urllib.request import urlopen

# 3. 저장하기
html_file = open('html_file.html', 'w')
html_file.write(text)
html_file.close()
```

Python으로 HTTP 가져오기

```
import sys

# 4. 불러오기
html_file = open('html_file.html', 'r')
read_html = html_file.read()
print(read_html)


html_file.close()
```

Python으로 HTTP 가져오기

네이버 뉴스 헤드라인 목록을
naver_news_headline_title.txt 로 저장한다.

Python으로 HTTP 가져오기

네이버 뉴스 헤드라인 목록의 뉴스 내용을
naver_news_n_yyyy_mm_dd.txt 로 저장한다.

Part 3

코드 리뷰

에러 메시지에 대처하는 방법

코드 리뷰

날씨 데이터를 가져와 보자

주가 데이터를 가져와 보자