BigData

- 실습 코드소스 깃 참고

■ <DB 구축 - DBMS / SQLite>

<SQLite 설치>
https://sqlite.org/download.html
sqlite-tools-win32-x86-버전.zip 다운로드

<DBeaver 프로그램>
https://dbeaver.io/download/
Windows(Installer)

<Web Crawling>
Web Crawling은 웹 사이트에서 정보를 수집하는 자동화된 프로세스입니다. 다른 말로 "웹 스크래핑" 또는 "웹 스크롤링"으로도 알려져 있습니다.

Web Crawling의 주요 특징은 다음과 같습니다:

자동화된 데이터 수집: Web Crawling은 자동으로 웹 페이지를 탐색하고 데이터를 수집합니다. 이것은 대규모 정보 수집 작업에 매우 유용합니다.

정기적인 업데이트: 웹 크롤러는 정기적으로 웹 사이트를 방문하여 새로운 정보를 수집하거나 변경된 정보를 업데이트합니다.

다양한 목적: Web Crawling은 다양한 목적으로 사용됩니다. 예를 들어, 검색 엔진은 웹 크롤링을 사용하여 검색 결과를 생성하고, 가격 비교 웹 사이트는 제품 가격 정보를 수집하고, 뉴스 사이트는 최신 뉴스 기사를 수집합니다.

링크 추적: 웹 크롤러는 웹 페이지 간의 하이퍼링크를 따라 이동하며 새로운 웹 페이지를 발견합니다. 이를 통해 전체 웹을 탐색할 수 있습니다.

데이터 추출: 크롤러는 웹 페이지의 HTML 또는 다른 마크업 언어를 분석하고 원하는 데이터를 추출합니다. 이 데이터는 일반적으로 텍스트, 이미지, 링크 등 다양한 형식일 수 있습니다.

로봇 프로토콜 준수: 웹 크롤러는 로봇 프로토콜(robots.txt)을 준수하여 특정 웹 사이트에서 수집할 수 있는 페이지를 제한하거나 제한할 수 있습니다.

Web Crawling은 정보 수집, 웹 사이트 모니터링, 경쟁 정보 수집, 검색 엔진 최적화, 가격 비교, 뉴스 집계 등 다양한 분야에서 활용됩니다.

<urllib 사용>

```
import urllib.request

nateUrl = "https://www.nate.com"
htmlObject = urllib.request.urlopen(nateUrl)
html = htmlObject.read()
```

```
print(html)
```

```
b'\r\n\r\n<!DOCTYPE html>\r\n<html lang="ko">\r\n<head>\r\n\t\r\n<meta
http-equiv="X-UA-Compatible" content="IE=Edge" />\r\n<meta name="msapplication-starturl"
content="//www.nate.com/" />\r\n<meta http-equiv="Content-Type" content="text/html;
charset=utf-8" />\r\n<meta name="nate:title" content="" />\r\n<meta
name="nate:description" content="\xeb\x84\xa4\xec\x9d\xb4\xed\x8a\xb8
\xec\x9d\xb4\xec\x8a\x88UP" /> ... 형태로 긴 줄로 나타나게 됨
```

<시간 간격 크롤링 - 자주 쓰이는 크롤링 기법>

```
import csv
import time
import datetime

csvName =  'C:/Users/admin/Desktop/Python, BigData/source/CSV/datetime.csv'
with open(csvName, 'w', newline='') as csvFp:
    csvWriter = csv.writer(csvFp)
    csvWriter.writerow(['연월일', '시분초'])

count = 10
while count > 0 :
    count -= 1

    now = datetime.datetime.now()
    yymmdd = now.strftime('%Y-%m-%d')
    hhmmss = now.strftime('%H:%M:%S')
    time_list = [yymmdd, hhmmss]
    print(time_list)

    with open(csvName, 'a', newline='') as csvFp:
        csvWriter = csv.writer(csvFp)
        csvWriter.writerow(time_list)

    time.sleep(3)
```

```
['2023-09-06', '15:10:56']
['2023-09-06', '15:10:59']
['2023-09-06', '15:11:02']
['2023-09-06', '15:11:05']
['2023-09-06', '15:11:08']
['2023-09-06', '15:11:11']
['2023-09-06', '15:11:14']
['2023-09-06', '15:11:17']
['2023-09-06', '15:11:20']
['2023-09-06', '15:11:23']
```

와 같이 나타난다.

원본 소스 Code09-15

import bs4
import urllib.request ...
전연 변수부에 아래 코드 한 줄 추가

```
        # 'euc-kr'로 디코딩
        decodedPage = webPage.decode('euc-kr', 'ignore')
```

디코딩 없으면 글자 깨짐 현상 발생