

Python을 활용한 웹 크롤러 만들기

2018. 04. 12

CONTENTS

1

크롤링의 정의

2

개발 환경 구축

3

웹드라이버 이해

4

Selenium 이해

CONTENTS

5

웹드라이버를 이용한 selenium 주요 api 습득

6

크롤링 타겟 사이트 분석 및 데이터 접근
실습

7

Beautiful Soup 이해 및 api 습득

8

수집 데이터 전처리 및 DB 처리

학습 목표

- 크롤링에 작업 이해
- Selenium을 활용한 웹드라이버의 자동
공정화 처리 이해
- BeautifulSoup를 이용한 데이터 추출 및
디비 처리 이해



1. 크롤링의 이해

◉ keyword

- 크롤링이란?
웹 페이지를 그대로 가져와서 거기서 데이터를 추출해 내는 행위
- 크롤러
크롤링 소프트웨어
- 법적유무

A person's hands are shown holding a smartphone with a white screen. The background is dark with out-of-focus, circular bokeh lights in shades of yellow, orange, and blue. A semi-transparent dark blue horizontal bar is at the bottom, containing a yellow decorative element and the text '2. 개발 환경 구축'.

2. 개발 환경 구축

◉ 주제

- 언어
 - python.exe. 설치 (3.x)
- 모듈
 - selenium 설치
 - bs4 설치
- 웹 드라이버
 - Chrome, Phantom
- 에디터
 - vs code 설치
 - plugin 설치

A person's hands are shown holding a smartphone, with the screen glowing. The background is a dark, out-of-focus city night scene with warm, yellow and orange bokeh lights. A semi-transparent dark blue banner is at the bottom, containing a yellow chevron icon and the section title.

3. 웹 드라이버의 이해

keyword

- 웹드라이버 정의
 - 자동화 설계
 - 시나리오에 따른 움직임


Mozilla GeckoDriver	0.20.0	change log	issue tracker	Implementation Status	Released 2018-03-08
Google Chrome Driver	2.36	change log	issue tracker	selenium wiki page	Released 2018-03-02
Opera	2.29		issue tracker	selenium wiki page	Released 2017-06-27
Microsoft Edge Driver			issue tracker	Implementation Status	
GhostDriver	(PhantomJS)		issue tracker	SeConf talk	
HtmlUnitDriver	2.28.1		issue tracker		Released 2017-11-19
SafariDriver			issue tracker		
Windows Phone			issue tracker		
Windows Phone	4.14.028.10		issue tracker		Released 2013-11-23
Selendroid - Selenium for Android			issue tracker		
ios-driver			issue tracker		
BlackBerry 10			issue tracker		Released 2014-01-28
Appium			issue tracker		
CrossWalk			issue tracker		Released 2014-05-05
QtWebDriver	1.3.1	change log	issue tracker	wiki page	Released 2015-06-17
jBrowserDriver			issue tracker		
Winium.Desktop	latest	change log	issue tracker	wiki, talks & demos	
Winium.StoreApps	latest	change log	issue tracker	wiki, talks & demos	
Winium.StoreApps.CodedUi (Early stage WIP)	latest		issue tracker	talks & demos	

A person's hands are shown holding a smartphone, with the screen glowing. The background is dark with out-of-focus, colorful bokeh lights in shades of yellow, orange, and blue. A semi-transparent dark banner is at the bottom, containing a yellow decorative bar and the section title.

4. Selenium의 이해

◉ keyword


- 정의
- 웹드라이버 띄우기
- 에이전트 조작
- 프록시 조작

A person's hands are shown holding a smartphone, with the screen glowing. The background is dark with out-of-focus, colorful bokeh lights in shades of yellow, orange, and blue. A semi-transparent dark banner is at the bottom, containing a yellow chevron icon and text.

5. 웹 드라이버를 이용한 Selenium의 주요 API 습득

◉ keyword

- 페이지 접속
- 우회 접속
- 로그인 및 검색 등 폼처리
- 찾기
- 추출하기

A person's hands are shown holding a smartphone, with the screen glowing. The background is dark with out-of-focus, warm-toned circular lights (bokeh).

6. 크롤링 타겟사이트 분석 및 데이터 접근 실습

◉ keyword

- search
- result
- rotation
- selenium의 최대치

A person's hands are shown holding a smartphone, with the screen glowing. The background is dark with out-of-focus, warm-toned bokeh lights in shades of yellow, orange, and blue. A semi-transparent dark banner is at the bottom, containing a yellow decorative element and the title text.

7. Beautiful Soup 이해 및 API 습득

◉ keyword

- when?
- DOM 접근
- 콘텐츠 획득

A person's hands are shown holding a smartphone, with the screen glowing. The background is dark with out-of-focus, warm-toned lights (bokeh) in shades of yellow, orange, and blue. A semi-transparent dark banner is at the bottom, containing a yellow decorative element and the section title.

8. 수집 데이터의 전처리 및 DB 처리

◉ keyword

- 디비 접속 처리
- sql 처리
- 크롤링 데이터 삽입