Ask Company

크롤링 차근차근 시작하기

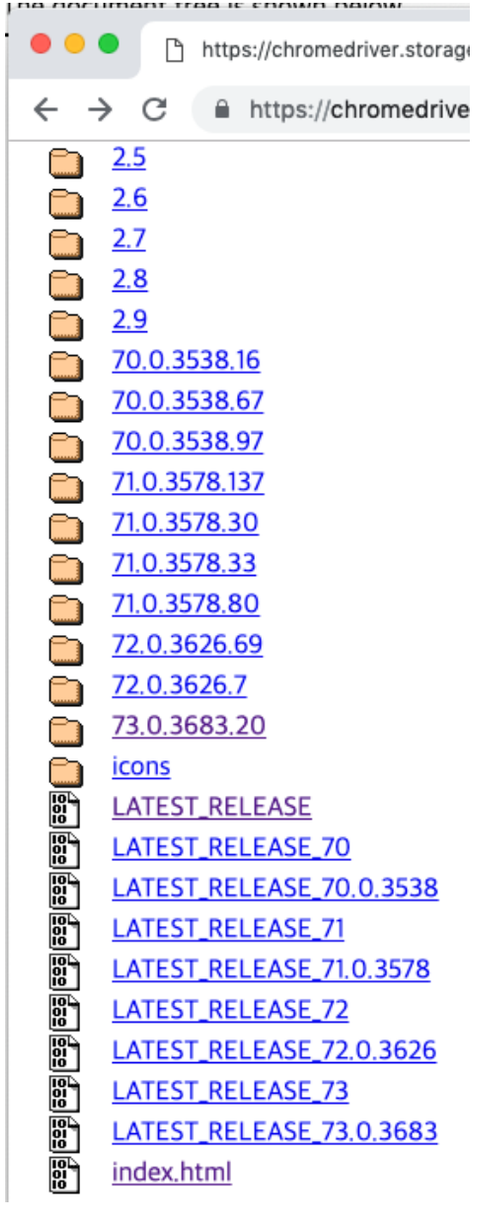
맛보기) requests와 selenium 간의 비교

여러분의 파이썬/장고 페이스메이커가 되겠습니다.

크롤링 대상

https://chromedriver.storage.googleapis.com/index.html

최신 버전명 찾기



requests로 시도

시도 #1

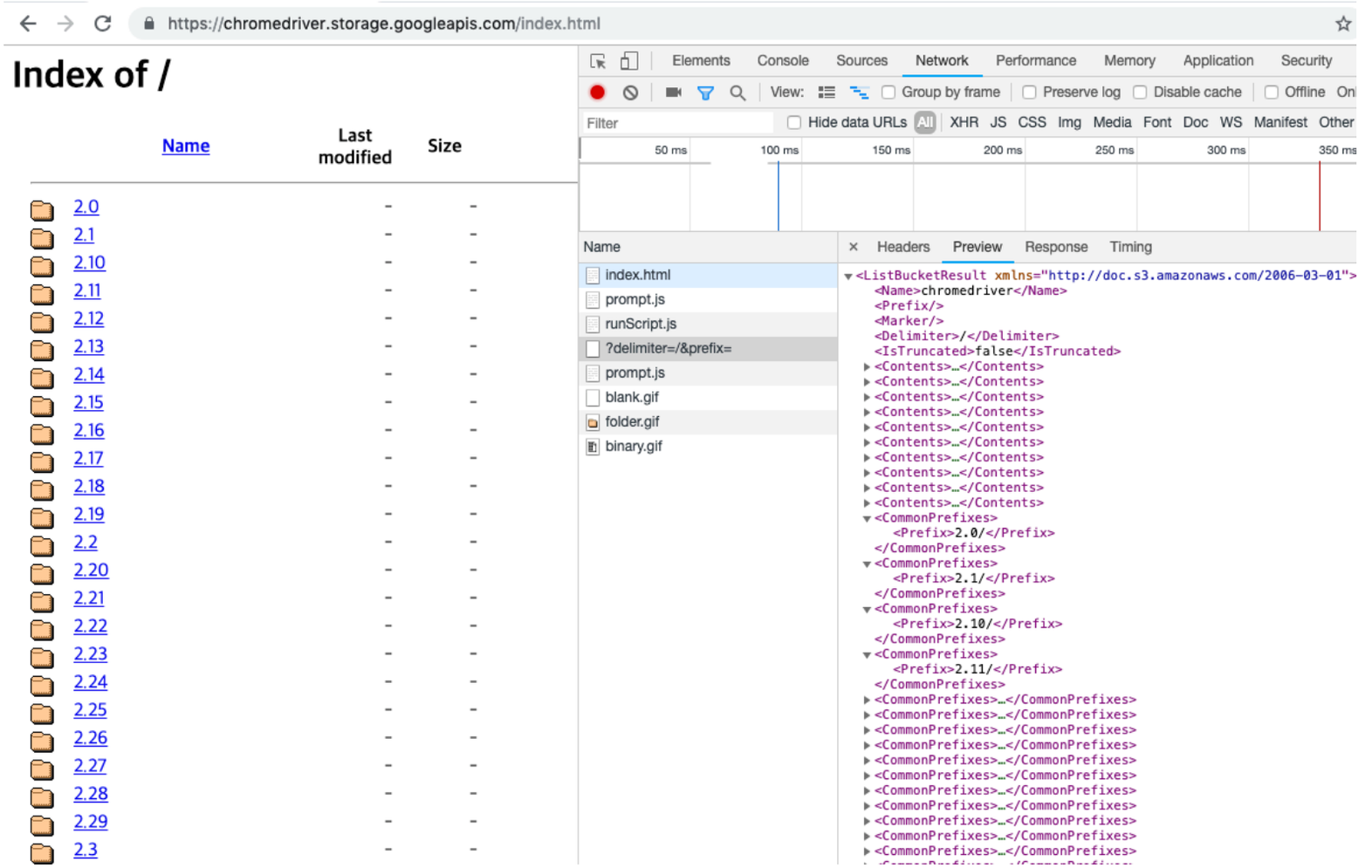
```
import requests
from bs4 import BeautifulSoup

url = "https://chromedriver.storage.googleapis.com/index.html"
res = requests.get(url)
html = res.text
soup = BeautifulSoup(html, 'html.parser')
soup.select('a')
```

살펴보니, JavaScript에 의한 컨텐츠 처리

별도 주소에서

XML 포맷 데이터 로딩하여

JavaScript를 통한 처리



https://chromedriver.storage.googleapis.com/?delimiter=/&prefix=

시도 #2

XML주소로 직접 요청을 보내어 처리

```
import re
import requests
from bs4 import BeautifulSoup

url = "https://chromedriver.storage.googleapis.com/?delimiter=/&prefix="
res = requests.get(url)
xml = res.text
soup = BeautifulSoup(xml, 'html.parser')

version_list = [
    tag.text.rstrip('/')
    for tag in soup.select('prefix')
    if re.match(r'^\d', tag.text)]

version_list.sort(key=lambda version: tuple(map(int, version.split('.'))))  # 오름차순 정렬
recent_version = version_list[-1]
print('recent_version :', recent_version)
```

Selenium으로 시도

시도 #3 Selenium으로 웹페이지 현재 DOM 내역을 읽어서 처리

```
import re
from selenium import webdriver
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait

chrome_driver_path = 'drivers/chromedriver'

with webdriver.Chrome(executable_path=chrome_driver_path) as driver:
    url = "https://chromedriver.storage.googleapis.com/index.html"
    driver.get(url)
    element_list = WebDriverWait(driver, 10).until(
        EC.presence_of_all_elements_located((By.TAG_NAME, 'a'))  # 주기적으로 체크할 함수 지정
    )
    version_list = [element.text for element in element_list if re.match(r'^\d', element.text)]
    # 아래와 같이 "현재 전체 HTML"을 받아서, BeautifulSoup을 통해 파싱을 하셔도 됩니다.
    # html = driver.page_source
    # soup = BeautifulSoup(html, 'html.parser')

# 브라우저 종료

version_list.sort(key=lambda version: tuple(map(int, version.split('.'))))  # 오름차순 정렬
recent_version = version_list[-1]
print('recent_version :', recent_version)
```

인생은 짧습니다.
파이썬/장고를 쓰세요.

여러분의 파이썬/장고 페이스메이커가 되겠습니다.

-

Ask Company