Chapter. 02

데이터 크롤링

I 개요

FASTCAMPUS
ONLINE
금융공학/퀀트I

강사. 서찬웅

이번 시간에 공부할 내용을 알아보겠습니다.

1. 크롤링이란 무엇인가요?
2. 자바 스크립트 사용 여부

I 크롤링이란 무엇인가요?

- 크롤러(crawler)는 조직적, 자동화 된 방법으로 www을 탐색하는 컴퓨터 프로그램입니다.
  - 크롤러가 하는 작업을 크롤링(crawling) 혹은 스파이더링(spidering)이라 부른다. – Wikipedia
  - 쉽게 표현하면 웹 페이지에서 정보를 추출하기 위한 프로그램입니다.
  - 크롤러는 스파이더(Spider) 혹은 (Bot)이라고도 부릅니다.
  - 크롤링은 웹 페이지에서 데이터를 수집하여 다운로드 받는 작업을 의미합니다.

- 파이썬으로 크롤링을 하는 이유는?
  - 웹 데이터 수집을 위한 라이브러리가 많고, 파이썬 언어의 간결성이 더해져서 다른 언어에 비해 코드 양 및 시간에 대한 이점이 있습니다.
  - requests, selenium, beautifulSoup등의 라이브러리를 사용해서 간단하게 구현
    - 다른 언어도 동일한 라이브러리 및 비슷한 라이브러리도 존재합니다. 다만 파이썬이 다른 언어에 비해서 편하고 시간이 절약됩니다.

I 크롤링의 방법에는 어떤 방법이 있나요?

- 웹 페이지에서 데이터를 가져올 때 2가지를 생각하셔야 합니다.
  1. 우리가 원하는 데이터가 html안에 존재하는 경우
  2. 데이터가 자바스크립트 및 ajax 기술을 사용하여 페이지가 구성된 경우

- 첫번째의 경우 requests 라이브러리를 사용하여 단순하게 처리가 가능합니다. 하지만 requests의 단점으로는 자바스크립트로 코딩된 데이터는 가져올 수 없다는 문제점이 있습니다. 이럴 때 두번째 방법으로 해결하면 됩니다.

- requests가 자바스크립트를 해석할 수 없기 때문에 우리는 selenium이라는 라이브러리의 힘을 빌려야 합니다. 그리고 더불어 Chrome, firefox, PhantomJS 등의 프로그램의 도움이 함께 필요합니다.

- requests를 이용하여 데이터를 수집하는 것보다 selenium을 사용하여 웹 데이터를 수집하는 것이 속도 측면에선 빠릅니다.

I 크롤링이란 무엇인가요?

- 크롤러(crawler)는 조직적, 자동화 된 방법으로 www을 탐색하는 컴퓨터 프로그램입니다.
  - 크롤러가 하는 작업을 크롤링(crawling) 혹은 스파이더링(spidering)이라 부른다. – Wikipedia
  - 쉽게 표현하면 타켓 대상 웹 페이지에서 정보를 추출하기 위한 프로그램입니다.
  - 크롤러는 스파이더(Spider) 혹은 (Bot)이라고도 부릅니다.
  - 크롤링은 웹 페이지에서 데이터를 수집하여 다운로드 받는 작업을 의미합니다.

- 왜 파이썬으로 크롤링을 해야 할까요?
  - Part 1에서 공부했듯이 다른 언어에 비해서 생산성이 높고, 크롤링을 할 수 있는 많은 도구들이 있습니다.
    - requests, BeautifulSoup, Selenium, scrapy등 패키지를 사용하면 크롤링을 편하게 진행할 수 있음

I 정리

- 크롤링의 정의
- 자바스크립트를 사용하는 페이지안에 있는 데이터 가져오는 방법
- 자바스크립트를 사용하지 않은 페이지에서 데이터를 쉽게 가져오는 방법

감사합니다

FAST CAMPUS
ONLINE
서찬웅 강사.