스크래핑

Kangwuk Heo

Blockchain Professional Architect

스크래핑 (Scraping) 과 크롤링

스크래핑은 '뉴스 기사를 스크랩하다', '게시글들을 스크랩하다'처럼 사용되는 용어

스크래핑

스크랩(Scrap)이란 우리가 원하는 자료를 모으는 행위

일반적으로 웹에서 사용되는 스크래핑은 웹 스크래핑이라고도 불리며 크롤링(Crawling)로 명칭이 사용됨

스크래핑이란 웹사이트의 페이지를 그대로 가져와서 그 안에서 데이터를 추출하는 행위

스크래핑' 기술을 활용해서 원하는 데이터를 추출한 뒤 자기 회사 서비스에 분석해서 사용한다던가, 통계를 내는데 사용

크롤링 (Web Scraping)

크롤링의 사전적인 의미는 기어다닌다는 뜻으로 컴퓨터 소프트웨어 기술로 웹사이트들을 돌아다니면서 정보를 수집하는 행위

한 페이지만 방문하는 것이 아니라 그 페이지에 링크되어 있는 또 다른 페이지를 지속적으로 방문하고 이처럼 링크를 따라 웹을 돌아다니는 모습이 마치 거미와 비슷하다고하여 스파이더라는 명칭을 사용

웹은 기본적으로 HTML 형태로 되어 있고, 웹 페이지상에서의 규칙을 적용, 분석해서 원하는 정보들만 뽑아서 가져오는 것을 웹 크롤링 작업

크롤링 (Web Scraping) 매커니즘

크롤링은 총 4단계 기준으로 순차적으로 처리되는 매커니즘으로 구성되어 있음

대상 선정	• 웹 상의 데이터는 고유한 ID를 가진다. URI라고 부르며, 이는 우리가 잘 아는 웹 사이트 주소인 URL과 RUN 으로 구성 • 크롤링할 웹사이트 선정 및 URL 정보 획득
데이터 로드	• 데이터 로드는 웹 사이트를 로딩하는 것 • API 호출일 경우, XML, JSON 파일 정보 로딩 • 웹페이지 호출일 경우, 웹페이지의 HTML문서를 다운 받는 것
데이터 분석	• 로드된 데이터에서 필요한 부분을 추출하는 것 • 수집할 대상과, 수집이 필요없는 대상을 선정하는 과정
데이터 수집	• 데이터 분석 과정을 통해서 수집할 내용을 선정 • 데이터를 추출하여 파일 또는 데이터를 메모리에 저장하는 과정

감사합니다

Kangwuk Heo
calvin.heo@gmail.com