파이썬 라이브러리를 이용한 효율적인 크롤링 기법에 관한 연구

A Study on the Efficient Crawling Techniques using Python Libraries

저자 (Authors)
성주원, 김승현, 김상철
Seong JuWon, Sung Hyun Kim, Sang-Chul Kim

출처 (Source)
한국통신학회 학술대회논문집 , 2019.1, 988-989(2 pages)
Proceedings of Symposium of the Korean Institute of communications and Information Sciences , 2019.1, 988-989(2 pages)

발행처 (Publisher)
한국통신학회
Korea Institute Of Communication Sciences

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08003643

APA Style
성주원, 김승현, 김상철 (2019). 파이썬 라이브러리를 이용한 효율적인 크롤링 기법에 관한 연구. 한국통신학회 학술대회논문집, 988-989

파이썬 라이브러리를 이용한 효율적인 크롤링 기법에 관한 연구

성주원, *김승현, **김상철
국민대학교, *U.S. Air Force, **교신저자

lumyjuwon@gmail.com, *thehappyones3@gmail.com, sckim7@kookmin.ac.kr

A Study on the Efficient Crawling Techniques using Python Libraries

Seong JuWon, *Sung Hyun Kim, Sang-Chul Kim

Kookmin University, *U.S. Air Force

요 약

본 논문에서는 파이썬 크롤링 주요 라이브러리인 Beautifulsoup 과 동적 크롤링을 진행하기 위해 Selenium 웹 브라우저를 이용하여 다양한 크롤링 기법과 사용 방법에 대해 연구한다. 기계 학습(Machine Learning)에 필요한 데이터를 공급하기 위해 멀티 프로세스를 이용한 대용량 데이터 수집 기법과 동적 웹에서 얻어낼 수 있는 데이터 수집 방법은 필수적인 요소이다. 본 논문에서는 연구한 멀티 프로세스를 활용한 크롤링은 네이버 인터넷 뉴스를 대상으로 진행되었으며 동적 웹에서 얻을 수 밖에 없는 데이터들은 selenium 드라이버를 이용하여 수집하였다.

Ⅰ. 서 론

그래픽카드와 하드웨어의 발전으로 인해 예전엔 구현하기 힘들었던 인공지능 기술들이 급격히 개발되고 있다. 하드웨어 발전 부분의 큰 기여도 있었지만 수 많은 데이터들로 인하여 머신러닝에 빠른 발전을 도모해낼 수 있었다.

머신러닝에 필요한 데이터 공급과 BM(business model)에서 필요로 하는 데이터 보유가 중요해지고 있다. 그러나 데이터를 생산하거나 보유하고 있는 상황이 아닌 경우에는 머신러닝을 진행하기 위한 데이터를 공급하기 매우 어려운 상황이다. 효율적인 머신러닝을 이루어 내기 위해서는 방대한 데이터들을 필요로 한다. 이러한 문제를 해결하기 위해 방대한 데이터들이 모여 있는 웹에서 데이터를 수집하는 크롤링과 효율적인 크롤링 기법을 통해 데이터 수집 과정에 걸리는 시간을 단축할 수 있다.

문법에 적합하며 대용량의 데이터가 축적된 인터넷 뉴스를 빠른 시간 안에 수집할 수 있는 방법에 대해서 구현 및 알아보려고 한다.

Ⅱ. 본론

본 논문에서는 파이썬 크롤링 주요 라이브러리인 Beautifulsoup 과 동적 크롤링을 진행하기 위해 Selenium 웹 브라우저를 이용하여 다양한 크롤링 기법과 사용 방법에 대해 연구한다. 데이터 수집을 위한 웹 페이지 대상은 네이버 인터넷 뉴스이며 수집 할 카테고리는 정치, 경제, 사회, 생활문화, 세계, IT 과학이다. 데이터 저장 형식은 csv 로 이루어진다.

Ⅱ-I. 라이브러리 소개

Beautifulsoup 은 웹 페이지에서 가져온 HTML 또는 XML 에서 쉽게 데이터를 가져올 수 있도록 도와주는 파이썬 기반의 라이브러리이다. HTML 또는 XML 에 포함 돼 있는 데이터들을 특정 형식에 맞게 추출해 낼 수 있도록 도와주는 라이브러리다 [1].

Selenium 은 웹 테스트 프레임워크다. Selenium 프레임워크에서 이용할 수 있는 브라우저 Chrome, Internet Explorer 등을 이용해 동적 데이터들을 포함한 HTML, XML 을 가져올 수 있다 [2].

Ⅱ-II. 크롤링

네이버 인터넷 뉴스는 동적 웹 페이지이지만 사용자에게 제공될 때는 정적 HTML, XML 제공되기 때문에 selenium 을 이용할 필요가 없다.

HTML 을 서버에서 받아올 때는 받아오는 시간 간격을 주는 것이 중요하다. 시간 간격 없이 HTML 을 불러올 경우에는 서버에서 접속 자체를 거절하는 일이 발생하기 때문이다. 네이버 인터넷 뉴스 같은 경우에는 0.01 초 시간 간격을 주면 접속 거절이 발생하지 않았다.

HTML 분석 시에는 Chrome 과 같은 브라우저에서 인터넷 개발 도구를 통하여 원하는 데이터들의 태그와 형식들을 확인할 수 있다.

HTML 을 열어 필요한 데이터를 찾는 것도 좋은 방법이긴 하지만 구독성이 떨어져 분석하기 힘든 경우가 많기 때문에 브라우저 인터넷 개발 도구를 이용하는 것이 효율적이다.

반면 정적 HTML 이 주어지지 않고 동적 HTML 이 주어지는 웹 페이지에서는 selenium 웹 드라이버를 이용하여 HTML 을 추출해낼 수 있었다. selenium 웹 드라이버는 HTML 을 받아올 때 클라이언트 입장에서 가져오기 때문에 동적 요소들이 정적 HTML 로 변환 된다. 그러나 웹 드라이버에서 동적 데이터가 포함 된

HTML 을 가져올 때는 로딩 시간이 걸리므로 동적 내용을 가져올 때까지의 시간 간격을 설정해야 한다 [3].

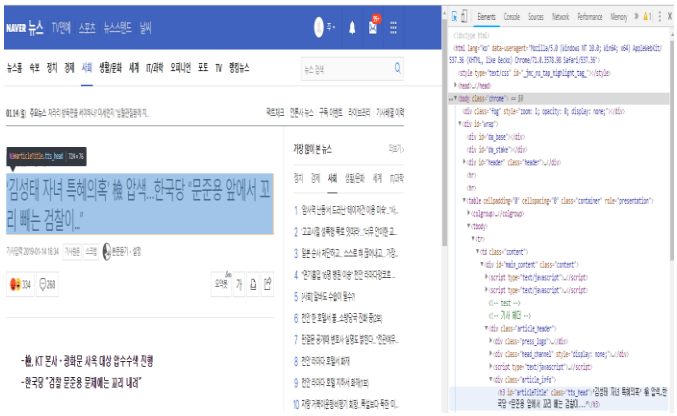

그림 1. Chrome 인터넷 개발 도구를 이용한 HTML 분석

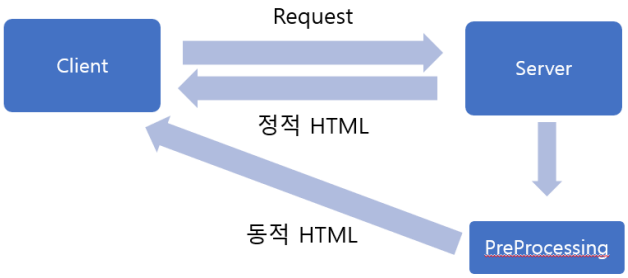

그림 2. 동적 웹 페이지 HTML

Ⅱ-Ⅲ. 시간 단축 방법

서버에 Request 를 할 때 멀티 프로세스를 이용한 크롤링은 데이터 수집 시에 한 개의 프로세스로 여러 범주의 데이터들을 순차적으로 처리하는 것이 아닌 병렬적으로 여러 범주를 동시에 처리하는 것이다. 데이터 수집 시에 각 프로세스 간의 통신이 필요 없기 때문에 IPC(Inter Process Communication)은 구축하지 않을 예정이다 [4].

그림 3. 순차적 크롤링

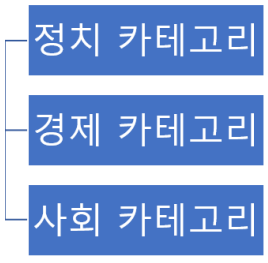

그림 4. 병렬적 크롤링

Ⅲ. 결론

본 논문에서는 크롤링을 통한 데이터 수집 시에 주의해야 하는 요인들과 정적 HTML 및 동적 HTML 크롤링하는 방법과 시간을 절약할 수 있는 멀티 프로세스 크롤링에 대해서 설명했다. 머신러닝과 같은 기술들을 개발할 때 반드시 필요한 데이터들을 웹 크롤링을 통하여 대용량으로 수집할 수 있었고 수집 시에 걸리는 시간을 줄이기 위해 멀티 프로세스를 이용했다. 하지만 예상치 못했던 에러들이 발생하기 때문에 예외처리를 반드시 해야 하며 예외처리 이후에도 에러가 발생할 수 있으니 크롤링이 멈춘 시점을 따로 기록하여 그 시점부터 다시 이어갈 수 있는 기술들이 필요한 것 같다. 이 연구를 통해 인터넷 기사 뿐만 아니라 더 많은 분야에서 크롤링을 통해 품질이 좋은 데이터들을 효율적으로 수집할 수 있을 것이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 연구결과로 수행되었음 (2016-0-00021)

참 고 문 헌

[1] Beautiful Soup Documentation, https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[2] Selenium Documentation, https://www.seleniumhq.org/.

[3] Dynamic HTML, https://en.wikipedia.org/wiki/Dynamic_HTML.

[4] Abraham Silberschatz, Operating System Concepts 10th, pp, 106-126.