파이썬을 이용한 다양한 형식의 웹 데이터 크롤링 기법

Crawling Methods for Web Data of Various Formats Using Python

저자
(Authors)

승리, 윤수진, 우영운
Li Seung, Sujin Yun, Young Woon Woo

출처
(Source)

한국정보통신학회 종합학술대회 논문집 23(2), 2019.10, 343-346(4 pages)

발행처
(Publisher)

한국정보통신학회
The Korea Institute of Information and Communication Engineering

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09262454

APA Style

승리, 윤수진, 우영운 (2019). 파이썬을 이용한 다양한 형식의 웹 데이터 크롤링 기법. 한국정보통신학회 종합학술대회 논문집, 23(2), 343-346

파이썬을 이용한 다양한 형식의 웹 데이터 크롤링 기법

승리* · 윤수진 · 우영운

동의대학교

Crawling Methods for Web Data of Various Formats Using Python

Li Seung* · Sujin Yun · Young Woon Woo

Dong-eui University

E-mail : victoria1016@naver.com / ywwoo@deu.ac.kr

요 약

이 논문에서는 카페나 블로그 형식의 다양한 웹 데이터를 자동으로 수집하기 위한 각종 기법들을 제안하였다. 제안한 맞춤식 수집 기법들과 HTML 실렉터를 활용할 수 있는 Python 언어와 Beautiful Soup 라이브러리를 이용하였으며, 특수한 형태로 구성되어 있는 카페, 블로그 등에 게시된 텍스트 데이터를 자동으로 모두 수집할 수 있었다. 제안한 기법들을 활용하여 다양한 형태의 구조로 이루어져 있는 각종 특수한 웹 페이지들에 대해서도 Python 웹 크롤링 프로그램에 의해 자동으로 대량의 데이터를 수집할 수 있었다. 이를 통해 다양한 대화 지식이 필요한 챗봇 구현이나, 빅데이터 분석 연구에 활용될 수 있을 것으로 예상한다.

ABSTRACT

In this paper, we proposed various techniques for automatically collecting various web data in cafe or blog format. We used the Python language and Beautiful Soup library, which can use the proposed custom collection techniques and HTML selector, and could automatically collect all the text data posted in cafes and blogs composed of special forms. By using the proposed technique, a large amount of data could be automatically collected by Python web-crawling program for various web pages with various structures. Through this, it is expected to be used for chatbot implementation that requires diverse conversation knowledge, or big data analysis research.

키워드

Web-crawling, Python, BeautifulSoup, HTML selector, Big data

Ⅰ. 서 론

최근 데이터 분석을 위해 웹 문서로부터 대량의 데이터를 수집하는 사례가 늘고 있다[1][2].

사용자가 직접 웹 페이지 내의 텍스트 데이터를 수동으로 수집할 수 있을 만큼 데이터 양이 많지 않은 경우에는 문제가 되지 않는다. 하지만 수집해야할 데이터의 양이 많을 때는 수동으로 하는 방법보다는 좀 더 적절한 방법이 필요하다. 이 논문에서는 챗봇에 활용하기 위한 데이터들을 수집하기 위해 Python 프로그램 등을 활용하여 자동으로 모든 페이지를 방문하면서 대량의 텍스트 데이터를 추출하는 방법과 이에 활용되는 HTML 실렉터(selector) 기능 활용 방안 기법을 제시한다.[3]

Ⅱ. 제안 기법

크롤링을 하려면 기본적으로 url주소가 필요하다. 하지만 많은 양의 데이터를 다루는 사이트의 경우 여러 페이지로 나뉘고 각 데이터마다 고유번호를 붙여 놓은 경우가 많다.

데이터를 크롤링하는 입장에서 각 데이터를 나타내는 페이지의 주소 값이 일정하고 유추가능하

* speaker

다면 큰 문제가 없을 테지만 만약, 유추할 수 없는 경우라면 어려움을 겪을 수 있을 것이다.

이 논문에서는 육아 관련 챗봇 개발을 위해 육아 관련 데이터를 크롤링 하는 과정에서 각 사이트의 구성형태에 따른 크롤링 방식을 제시한다. 이는 url을 구하는 방법을 중점적으로 크게 2가지로 나뉜다.

첫 번째는 단순히 URL을 통해 원하는 페이지의 HTML값을 크롤링하는 경우이다. 이를 '직접적인 페이지 주소방식' 이라 하겠다.

이 방식은 URL을 통해 직접적으로 페이지를 크롤링 할 수 있거나 URL의 특정 부분만 바꾸어 주면 되는 경우이다. 즉, 고유번호의 값의 범위가 명확하게 정해져 있기 때문에 따로 고유번호를 추출하는 등의 방식을 생략하고 이 고유번호 값을 유추할 수 있는 경우라 할 수 있다.[4][5]

그림 1. 페이지 주소와 추출할 Q&A 내용

그림 2. 일정 범위의 고유값을 포함한 유추 가능한 주소와 Q&A내용

두 번째 경우는 사이트의 페이지 주소만으로는 바로 크롤링을 할 수 없는 경우이다. 이는 '간접적인 페이지 주소방식' 이라 하겠다.

이 경우는 크롤링 할 대상인 Q&A의 전체 내용이 페이지에 나타나 있지 않는 구성을 띄거나 url을 구성하는 각 Q&A의 고유 번호의 값의 순서가 일정하지 않고 범위 또한 알 수 없을 경우이다.(즉, 실제 url의 값을 유추할 수 없을 경우이다.)[6]

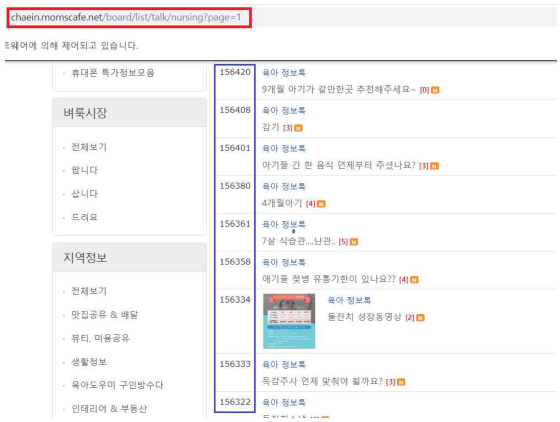

그림 3. 페이지주소와 추출해야 할 데이터의 고유번호

따라서 고유 번호를 수집하여 직접적인 주소를 구해야 크롤링이 가능하기 때문에 페이지 주소에서 각 Q&A의 고유 번호를 먼저 크롤링 해야하므로 이와 같은 과정을 거쳐야 한다.

① 각 페이지별 HTML코드를 txt파일로 저장함
② HTML코드 내에서 게시글로 연결되는 url을 추출하여 이에 속한 고유번호들을 추출
③ 고유번호들을 txt파일에 목록 형식으로 저장
④ txt파일에 모아놓은 고유번호들을 이용하여 'url+고유번호'형식으로 게시글에 접속하여 데이터 크롤링

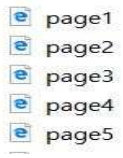

그림 4. 각 페이지의 HTML 코드 저장

```
for i in lst:

    id_ = i.replace('\n', '')

    r = requests.get(f'http://www.mamslibe.com/bbs/board.php?bo_table=cunsolt&'+id_)
```

그림 5. url + 고유번호 형식으로 크롤링

그림 6. 게시글 별 고유번호

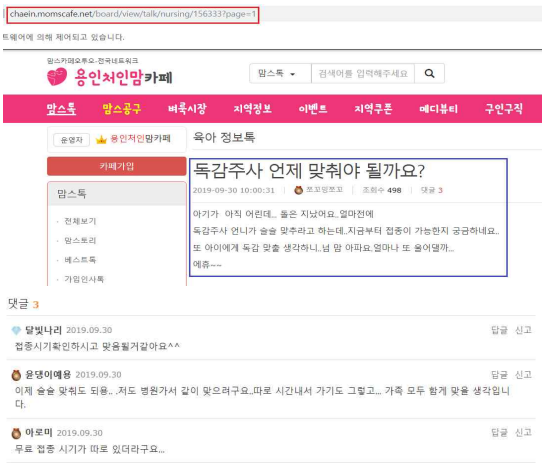

그림 8. 추출한 고유번호로 완성한 실제 url과 Q&A의 내용

다음은 별도로 사이트에서 질문의 답변내용을 페이지 소스에 제공하지 않는 경우를 다루었다. 이러한 경우 이전의 beautifulsoup를 이용한 크롤링 방식으로는 답변의 내용을 수집할 수 없다.

selenium을 사용하면 크롤링을 하길 원하는 페이지를 새롭게 직접 만들어 그 전체 내용을 새로운 페이지 소스로 만들어 제공한다.

따라서 기존 사이트의 페이지 소스에서 찾을 수 없었던 답변 내용을 selenium을 통해 찾을 수 있고 크롤링 가능하다. 새로 얻은 페이지 소스의 실렉터 정보를 이용해 각 Q&A의 질문과 답변을 수집할 수 있다.

그림 8. 기존의 방식으로는 answer의 내용을 크롤링할 수 없음

그림 9. selenium을 사용한 파이썬 코드 내용

그림 10. 위의 코드 실행 결과 answer 값

이러한 방식들로 웹페이지 게시판내의 데이터들을 크롤링할 수 있다.

III. 수행 결과

그림 11은 2장에서 제안한 웹페이지 게시판에서의 크롤링 방법을 활용하여 Python 코드이다.[7]

이 논문에서 작성한 웹 스크래핑 프로그램은 BeautifulSoup 라이브러리를 이용하여 작성하였으며 각 게시판의 질문과 대답 영역을 추출하기 위해 사용하였다.[8]

```
def getQ(text):
    soup = bs(text, 'html.parser')
    qcontent = soup.select('#bo_v_con')

    for item in qcontent:
        question = item.text
        question = extracts.sub('', question)

    return question
```

그림 11. 파이썬 웹 스크래핑 코드 일부

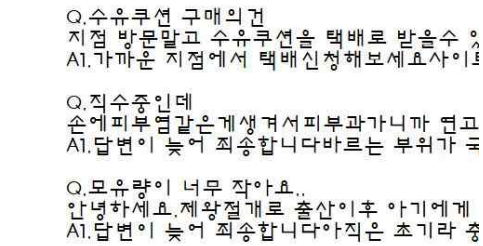

그림 12. 결과물 텍스트 파일

Ⅳ. 결 론

이 논문에서는 게시판 형식으로 된 웹페이지의 데이터들을 Python 웹 크롤링 프로그램 기능과 HTML 실렉터를 활용하는 기법을 이용하여 크롤링 하는 방법에 대해 여러 가지 방법을 제안하였다. 제안한 기법을 활용하여 “아이사랑”, “맘‘s 백과”, “용인처인맘카페”, “맘스리베” 등의 웹 사이트에서 여러 페이지로 분리되어 있는 게시판 데이터를 모두 수집할 수 있음을 확인하였다.

References

[1] K. W. Cho, S, K Bae and Y. W. Woo, “Analysis on Topic Trends and Topic Modeling of KSHSM Journal Papers using Text Mining,” *The Korean Journal of Health Service Management*, vol. 11, no. 4, pp. 213-224, Dec. 2017.

[2] Y. S. Kim and C. W. Seo, “A Study on the Analysis of Text Data Using Web Scraping of Cloud Computing,” *Proceedings of The Institute of Electronics and Information Engineering Summer Conference*, pp. 1736-1775, Jun. 2018.

[3] R. Mitchell, *Web Scraping with Python: Collecting Data from the Modern Web*, 1st edition, Sebastopol, CA:O'Reilly Media, Inc., 2015.

[4] I-sarang, Comprehensive Pregnancy and Infant Care Website. [Internet]. Available: http://www.childcare.go.kr/.

[5] 일동맘, [Internet]. Available: https://www.ildongmom.com/?home

[6] 용인처인맘카페, [Internet]. Available: http://chaein.momscafe.net/main/

[7] 맘스리베, [Internet]. Available: http://www.mamslibe.com/

[8] Beautiful Soup Documentation [Internet]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.