Journal of KIIT. Vol. 22, No. 4, pp. 23-28, Apr. 30, 2024. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2024.22.4.23

# 자연어 프롬프트 기반 데이터셋 생성 및 시각화 시스템

황유경*[1], 신진영*[2], 유석종**

# Dataset Creation and Visualization System based on Natural Language Prompt

Yu-Gyeong Hwang*[1], Jin-Young Shin*[2], and Seok-Jong Yu**


## 요 약

디지털 기술의 발전으로 급격히 증가하고 있는 데이터로부터 유용한 인사이트를 도출하는 데이터 분석이 중요해지고 있다. 특히, 웹상의 비정형 데이터를 추출하기 위해서는 웹 스크래핑 기술이 필요하지만, 웹페이지마다 HTML 구조가 상이하여 활용하기 어렵다. 따라서 본 연구에서는 웹 데이터 활용의 어려움을 개선하고자 자연어 프롬프트 기반의 데이터셋 생성, 편집 및 시각화 시스템을 제안한다. 제안 시스템은 자연어 프롬프트의 유형을 딥러닝 모델을 통해 분류하고 추출된 키워드에 따라 자동 웹 스크래핑을 수행하여 데이터셋을 생성, 편집, 시각화할 수 있다. 구현된 시스템의 성능 분석을 위해 도서, 뉴스 및 영상 도메인의 대표 웹사이트에 적용하고 수행 결과를 제시하였다. 본 시스템 사용자는 교육 연구 분석용 데이터셋을 얻기 위해 복잡한 웹 스크래핑을 위한 스크립팅 대신 자연어 프롬프트를 사용하여 데이터셋 생성, 편집 및 분석 작업을 수행할 수 있다.

## Abstract

It is important to analyze data for deriving useful insights with the rapid growth of data. Web scraping techniques are needed to extract unstructured data from the web, but it is difficult to utilize due to the different HTML structure of each web page. In this study, we propose dataset creation and visualization system based on natural language prompt to improve the difficulty of web data utilization. The proposed system uses a deep learning model to classify the types of natural language prompt, it can perform automated web scraping based on the extracted keywords and create, edit and visualize datasets. We applied the system to websites from various domains to evaluate its performance. The system enables users to create, edit, and analyze datasets using natural language prompts instead of writing complex web scraping scripts to obtain datasets for educational research analysis.


Keywords
dataset, NLP, web scraping, prompt, CNN model

* 숙명여자대학교 소프트웨어학부 학사과정
- ORCID[1]: https://orcid.org/0009-0004-5723-5879
- ORCID[2]: https://orcid.org/0009-0006-2737-5509
** 숙명여자대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: Feb. 19, 2024, Revised: Mar. 05, 2024, Accepted: Mar. 08, 2024
· Corresponding Author: Seok-Jong Yu
Dept. of Computer Science, Sookmyung Women's University, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

# Ⅰ. 서 론

디지털 기술의 발전으로 데이터 양이 급격히 증가함에 따라 대규모 데이터로부터 가치 있는 정보를 추출하는 능력이 중요해지고 있다. 대량의 데이터에서 유용한 인사이트를 도출하기 위해서는 데이터 분석에 필요한 양질의 데이터셋 확보가 핵심과제이다. 일반적으로 웹상의 비정형 데이터를 데이터셋으로 구축하기 위해서 Selenium[1] 또는 Beautiful Soup[2] 프로그래밍 도구나 웹사이트에서 제공하는 API를 이용한다. 그러나 이 방법은 웹사이트마다 다른 페이지 구조를 분석하여 적절한 웹 스크래핑 코딩 작업을 수행해야 하는 문제점이 있다. 이를 개선하고자, GUI 방식의 ScrapeStorm[3]과 플러그인 방식의 Listly[4] 서비스가 개발되었다. 이 방법은 코드 작성 없이 웹 스크래핑이 가능하지만, 웹페이지 구조에 따라 생성된 데이터셋의 품질에 편차가 발생한다. 또한, 데이터셋 생성 이외에 편집이나 시각화 분석 기능은 제공하지 않고 있다. 이에 대한 개선 방향으로 본 연구에서는, 웹페이지 구조를 이해하여 직접 코드를 작성하거나 스크래핑 도구 사용법을 학습하지 않고 데이터셋을 생성할 수 있는 방법을 제안하고자 한다. 이에 대한 구체적인 방법으로, 사용자가 입력한 자연어 프롬프트를 CNN 모델[5]로 분석하여 필요한 메타 정보를 추출하고, 해당 사이트에 대한 데이터셋을 자동 생성할 수 있다. 제안 시스템(NTD, Natural Language to Dataset)은 하부 스크래핑 라이브러리인 Selenium과 Beautiful Soup으로 구현되었으며, 도서, 뉴스, 영상 도메인의 대표 사이트에 적용하여 생성한 데이터셋 품질을 비교하는 성능 평가를 수행하였다.

본 논문의 2장에서는 기존 웹 스크래핑 기술을 소개하고, 자연어 질의 기반 시스템과의 특징을 비교하였다. 3장에서는 본 시스템의 구조와 처리 과정을 서술하였고, 4장에서는 수행 결과와 성능 평가 결과를 제시하고, 5장에서 결론을 맺는다.

# Ⅱ. 관련 연구

## 2.1 웹 스크래핑 기술

웹 스크래핑(Web scraping)이란 웹페이지에서 필요한 데이터만을 추출하여 활용 가능한 데이터셋 형태로 생성하는 작업을 의미한다. 대표적인 웹 스크래핑 프레임워크인 Selenium은 웹드라이버를 통해 특정 웹페이지에서 검색 키워드를 입력하거나 특정 버튼 클릭 행위를 자동화하여 원하는 데이터가 포함된 웹페이지를 생성한다. 그 다음, 생성된 페이지의 HTML 소스를 추출하는 동적 웹 스크래핑 작업을 수행한다. Beautiful Soup은 HTML과 XML 문서들의 구문을 분석하고 필요한 데이터를 추출하는 Python 패키지이다[2]. 제안 시스템에서는 사용자의 직접적인 조작 없이 동적 웹페이지의 데이터를 추출하고, 효율적인 파싱을 위해 Selenium과 Beautiful Soup을 함께 사용한다[6].

장동훈의 연구(DH)[7]에서는 사용자 정의에 따라 동작하는 Selenium 기반 웹 크롤링 시스템을 제시하였다. 이는 사용자가 입력한 키워드를 기준으로 동작하기 때문에 제안 시스템보다 사용자 요구사항의 처리 범위가 제한적이다. 장종욱의 연구(JW)[8]에서는 Selenium을 통해 신문 기사를 웹 스크래핑하고 시각화하였다.

ScrapeStorm(SS)[3]은 GUI에서 입력된 URL에 대해 인공지능 알고리즘을 기반으로 웹페이지 구조를 자동 인식하여 데이터를 추출하고 CSV, HTML 등 다양한 형태로 웹 스크래핑 결과를 제공한다. Listly(LT)[4]는 웹 브라우저 플러그인 방식의 프로그램으로, 사용자가 선택한 웹페이지 영역의 데이터를 추출한다.

표 1은 기존 웹스크래핑 서비스 및 선행 연구와 제안 시스템을 비교한 결과이다. 비교 항목은 인터페이스 방식, 데이터셋 생성, 편집, 시각화 기능, 결측치(Null value) 여부와 지원 사이트 범위이다.

표 1. 웹 스크래핑 서비스 및 연구 비교
Table 1. Comparing web scraping services and researches

| | SS | LT | DH | JW | NTD |
|---|---|---|---|---|---|
| Interface | GUI | Plugin | Options | unknown | NL prompt |
| Create | O | O | O | X | O |
| Edit | X | X | X | X | O |
| Visualize | X | X | X | O | O |
| Null value | O | O | X | X | X |
| Support site | various | various | limited | unknown | limited |

### 2.2 자연어 질의 기반 시스템 연구

김나영의 연구[9]와 박예나의 연구[10]에서는 자연어 질의를 기반으로 데이터 분석 시스템을 구현하였지만, 이는 각각 주식과 부동산이라는 특정 도메인에 한정되어 있다. 반면, 제안 시스템은 자연어 프롬프트를 분석하여 도메인을 결정하므로 다양한 도메인으로 확장이 용이한 장점이 있다.

## III. 자연어 프롬프트 기반 데이터셋 생성 및 시각화 시스템

### 3.1 시스템 구조 설계

그림 1은 본 시스템에서 자연어 프롬프트를 이용하여 데이터셋을 생성, 편집하고 시각화하는데 필요한 처리 모듈과 데이터의 흐름을 나타낸 그림이다.

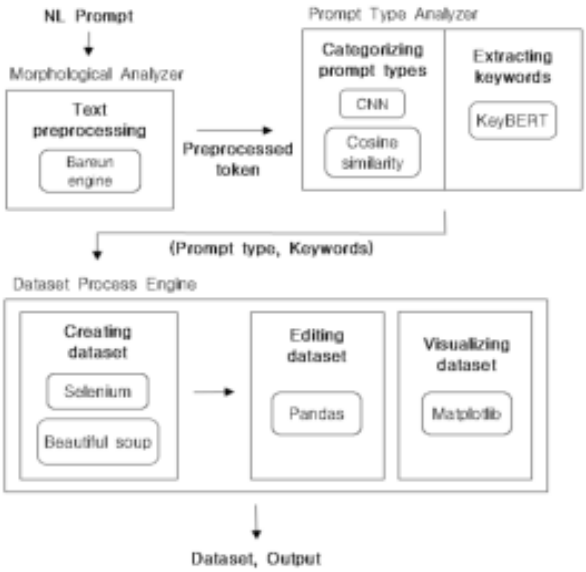


그림 1. 자연어 프롬프트 기반 데이터셋 생성 시스템
Fig. 1. Dataset creation system based on NL prompt

1) 데이터 시각화 및 사용자와의 상호작용을 지원하는 Streamlit 파이썬 라이브러리를 통해 개발한 사용자 인터페이스로 자연어 프롬프트(NL Prompt)를 입력받는다.

2) 입력받은 자연어 문장을 전처리하고 필요한 키워드를 추출한다.

3) 프롬프트 유형 분석기(Prompt type analyzer)를 활용하여 입력 문장을 분류하여 데이터셋 생성, 편집, 시각화 중 적합한 모듈에 전달한다.

4) 각 모듈별로 처리한 결과를 화면에 출력한다.

### 3.2 자연어 프롬프트 유형 분석

자연어 프롬프트를 전처리하기 위해 한국어 모호성 처리 성능이 좋은 바른(Bareun) 형태소 분석기 엔진[11]을 사용하여 문장을 토큰화하고 불용어 토큰을 제거한다. 다음으로 구글의 BERT 모델에 기반한 양방향 문장 학습으로 문맥을 파악하는 KeyBERT[12] 모델을 활용하여 키워드를 추출한다. 사전을 정의하여 추출된 키워드와 도메인, 장르, Pandas 메서드 이름과 같은 특징(Feature)을 매핑한다. 프롬프트 유형 분류기를 구현하기 위해 이미지 분류와 텍스트의 지역적인 특징 학습에 효과적인 심층 신경망 CNN 모델[13]과 문장을 벡터로 표현하여 유사성을 측정하는 코사인 유사도를 활용하였다. 프롬프트 유형 분류를 위하여, AI Hub 사이트[14]에서 자연어 기반 질의(NL2SQL) 검색 생성 데이터를 활용하여 78,244개의 한국어 질의문을 생성/편집/시각화 유형으로 라벨링한 데이터셋을 구축한 후 CNN 모델로 학습하였다. 표 2는 라벨링한 데이터셋의 예시이고, 표 3은 분류된 프롬프트 유형의 예시이다.

표 2. 훈련용 데이터셋 라벨링 결과
Table 2. Labeling result of the training dataset

| Korean query examples | label type |
|---|---|
| 위도가 36인 중개소명을 찾아줘 | create |
| 국공립 어린이집의 이름을 만 0세 반수가 많은 순으로 정렬해줘 | edit |
| 업태명이 업으로 끝나는 업소의 수를 보여줘 | visualize |

표 3. 자연어 프롬프트 유형
Table 3. Types of natural language prompt

| prompt type | domain | prompt sub-type |
|---|---|---|
| create | book | genre |
| | news | category |
| | video | search keyword |
| edit | book | add column, sort |
| | news | delete column |
| | video | delete row |
| visualize | book | ratio graph |
| | news | text frequency graph |
| | video | bar graph |

### 3.3 프롬프트 유형별 처리 과정

(1) 데이터셋 생성 : 프롬프트에서 추출된 도메인이 '도서'일 경우, 키워드 리스트에서 장르, 기간, 데이터 수 키워드를 웹 스크래핑 코드의 파라미터로 전달한다. 이때, 존재하지 않는 키워드는 기본값으로 대체되며, 이 정보를 포함한 동적 URL을 생성한다. 그 다음, Selenium을 통해 해당 URL에 접근하여 HTML 소스를 추출하고, Beautiful Soup으로 파싱하여 도서 데이터셋을 생성한다.

(2) 데이터셋 편집 : 과정 (1)에서 생성된 데이터셋의 칼럼 이름에 해당하는 키워드와 하위 프롬프트 유형에 적합한 Pandas 라이브러리를 결정하여 정렬, 필터링과 같은 데이터셋 편집 작업을 수행하고 그 결과를 출력한다.

(3) 데이터셋 시각화 : 데이터셋 시각화는 편집과정과 유사한 방식으로, 선택된 칼럼을 하위 프롬프트 유형에 적합한 차트로 시각화하여 제공함으로써 데이터의 분석 결과를 직관적으로 파악할 수 있다.

## IV. 구현 및 성능 평가

### 4.1 생성 데이터셋의 품질 비교

ScrapeStorm(SS), Listly(LT), 제안 시스템(NTD)을 각각 사용하여 각 도메인별로 사용자 수가 많고 최신 데이터 수집이 용이한 교보문고 베스트셀러, 네이버 뉴스, 유튜브 'AI' 검색 웹페이지를 스크래핑하였다. 표 4는 생성된 데이터셋의 크기(행x열)와 결측치(Null value)를 중심으로 정량적 비교한 결과이다. SS는 유튜브에서 257개, LT는 네이버 뉴스를 제외한 모든 경우에 결측치가 발생한 반면, NTD는 모든 사이트에서 결측치가 발생하지 않았다.

표 4. 생성된 데이터셋 품질 비교(행x열, ()=결측치)
Table 4. Comparison of created datasets(row*col, ()=null)

| site/method | | SS | LT | NTD |
|---|---|---|---|---|
| Kyobo book | IT | - | 10x31 (35) | 20x9 (0) |
| | all | - | 40x32 (128) | 50x9 (0) |
| Naver news | | 100x5 (0) | 25x5 (0) | 70x5 (0) |
| Youtube | | 86x10 (257) | 20x28 (182) | 100x3 (0) |

다음은 제안 시스템의 정성적 성능 분석 결과이다.

(1) 데이터 유용성 : SS는 웹 스크래핑을 할 때 웹사이트 구조에 따라 파싱이 불가능하여 결측치가 발생하는 경우가 있으며, 생성되는 데이터셋 크기를 지정할 수 없다. 반면, 제안 시스템은 원하는 크기의 데이터셋을 결측치 없이 생성할 수 있다. LT를 사용한 스크래핑의 경우, 교보문고 웹페이지에 로드된 모든 데이터를 추출하여 최대 32개 칼럼의 데이터셋를 생성하게 된다. 이는 제안 시스템과 달리 불필요한 데이터를 포함한 것으로 별도의 정제 과정이 필요하다.

(2) 사용 편의성 : SS와 LT를 활용하기 위해서는 먼저 웹페이지 구조를 이해하고 프로그램의 컴포넌트를 조작해 직접 웹 스크래핑 시나리오를 구성해야 하므로 사용 방법이 복잡하다고 할 수 있다. 반면 제안 시스템은 필요한 정보를 자연어 프롬프트 형식으로 입력하기 때문에 사용 편의성이 우수하다고 할 수 있다.

(3) 시스템 확장성 : 제안 시스템은 데이터 분석을 위해 평균, 정렬과 같은 편집 기능과 시각화 기능을 제공하고 있다. 반면, SS와 LT는 추출할 칼럼 선택 기능 외에 추가 기능을 제공하지 않는다.

### 4.2 시스템 인터페이스 구현 결과

그림 2, 그림 3, 그림 4는 본 시스템의 수행 결과이다.


그림 2. IT 도서 데이터셋 생성 결과
Fig. 2. Creation of an IT book dataset

(1) 데이터셋 생성 : 프롬프트를 통하여 교보문고 IT 도서 웹페이지를 스크래핑하여 데이터셋을 생성한 결과이다(그림 2).

(2) 데이터셋 편집 : IT 도서 데이터셋에서 '선호도' 칼럼을 기준으로 내림차순 정렬 결과이다(그림 3).

(3) 데이터셋 시각화 : 종합 카테고리 도서 데이터셋의 장르별 비율을 파이 차트로 시각화한 결과이다(그림 4).


그림 3. 데이터셋 정렬 결과
Fig. 3. Sorting result of dataset

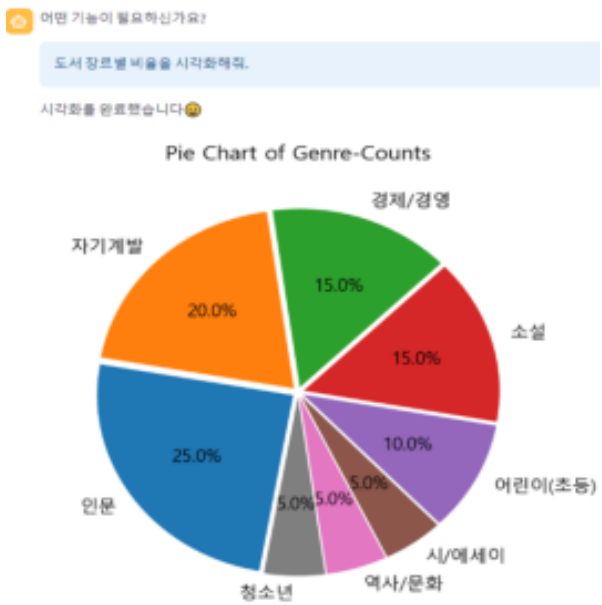


그림 4. 데이터셋 시각화 결과
Fig. 4. Visualization result of dataset

## V. 결론 및 향후 과제

본 연구에서는 사용자 편의성을 개선하고자 자연어 프롬프트를 기반으로 웹 데이터를 수집하고, 데이터셋으로 생성 및 편집하는 시스템을 구현하였다. 제안 시스템은 자연어 프롬프트 방식을 도입하여 데이터셋의 유용성, 사용자 편의성, 확장성 측면에서 기존 방법보다 우수하다고 평가할 수 있다. 반면, 웹 스크래핑 코드가 사전에 구현된 웹사이트에서만 동작한다는 한계가 있으며, 후속 과제에서 웹페이지 구조에 의존적이지 않은 스크래핑 방식에 대한 추가 연구가 필요하다. 또한, 자연어 프롬프트 유형 분석을 위해 보다 다양한 딥러닝 모델의 적용이 요구된다고 할 수 있다.

## References

[1] Selenium, https://www.selenium.dev [accessed: Jan. 22, 2024]

[2] Beautiful Soup, https://www.crummy.com/software/BeautifulSoup/bs4/doc/ [accessed: Jan. 22, 2024]

[3] ScrapeStorm, https://kr.scrapestorm.com [accessed: Jan. 22, 2024]

[4] Listly, https://www.listly.io/ko [accessed: Jan. 22, 2024]

[5] W. W. Kim and K. H. Park, "Design of Korean Text Emotion Classifier Using Convolution Neural Network", Proc. of the Korea Information Science Association, Vol. 2017, No. 6, pp. 642-644, Jun. 2017.

[6] Ruchitaa Raj N R, Nandhakumar Raj S, and Vijayalakshmi M, "Web Scrapping Tools and Techniques: A Brief Survey", 4th International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, Feb. 2023. https://doi.org/10.1109/ICITIIT57246.2023.10068666.

[7] D. H. Jang, D. H. Han, M. J. Kang, D. Y. Kim, Y. C. Ahn, Y. M. Lee, and S. J. Ko, "Design and Implementation of a User-defined Web Crawler", Proc. of Symposium of the Korean Institute of Communications and Information

Sciences, pp. 907-908, Jun. 2019.
[8] J. W. Jang, J. H. Jeong, and Y. S. Son, "A Study on the Crime Data Collection System Using Web Crawler", Proc. of KIIT Conference, pp. 93-94, Nov. 2018.
[9] N. Y. Kim and S. J. Yu, "Stock Information Retrieval and Analysis using Korean Natural Language Query", The Journal of Korean Institute of Information Technology, Vol. 20, No. 9, pp. 13-18, Nov. 2022. https://doi.org/10.14801/jkiit.2022.20.9.13.
[10] Y. N. Park and S. J. Yu, "Real Estate Environment Retrieval System based on Natural Language Query", The Journal of Korean Institute of Information Technology, Vol. 20, No. 11, pp. 23-28, Nov. 2022. https://doi.org/10.14801/jkiit.2022.20.11.23.
[11] Bareun, https://bareun.ai/ [accessed: Jan. 22, 2024]
[12] Bayan Issa, "A Comparative Study on Embedding Models for Keyword Extraction Using KeyBERT Method", 13th International Conference on System Engineering and Technology (ICSET), Shah Alam, Malaysia, pp. 40-45 Oct. 2023. https://doi.org/10.1109/ICSET59111.2023.10295108.
[13] J. J. Kim and C. B. Kim, "Implementation of Robust License Plate Recognition System using YOLO and CNN", The Journal of Korean Institute of Information Technology, Vol. 19, No. 4, pp. 1-9, Apr. 2021. https://doi.org/10.14801/jkiit.2021.19.4.1.
[14] AI Hub, https://www.aihub.or.kr/ [accessed: Jan. 22, 2024]

## 저자소개

황 유 경 (Yu-Gyeong Hwang)

2020년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정
관심분야 : 데이터 마이닝, 자연어 처리

신 진 영 (Jin-Young Shin)

2020년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정
관심분야 : 자연어 처리, 빅데이터 처리

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 전산과학과(이학사)
1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)
2001년 2월 : 연세대학교 컴퓨터과학과(공학박사)
2005년 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수
관심분야 : 데이터마이닝, 추천시스템, 정보시각화