

과제04: 데이터 처리 실습

언어 데이터 처리 2022-12-14 장태준 [컴퓨터 공학부 2017-17018]

<https://github.com/jtjun/NLP/blob/main/hw04/hw04.md>

1. 데이터

A. 소제목

Ai Hub에 가입/로그인한 뒤, 데이터를 다운 받았다.

AI Hub | AI 데이터찾기 | AI 개발지원 | 참여하기 | 정보공유 | 고객센터 | AI 허브소개 | 마이페이지 | 로그아웃

데이터 분야

#자연어 처리 #한국어 도서자료 #원문 #요약문 #생성 요약 #정보 추출

도서자료 요약

분야: 한국어 | 유형: 텍스트

갱신년월: 2022-10 | 구축년도: 2020 | 조회수: 1,621 | 다운로드: 1,018 | 용량: 201.91 MB

다운로드 | 샘플 데이터

관심데이터 등록 | 5

※ 내국인만 데이터 신청이 가능합니다.

목록

데이터 개요

데이터 변경이력

버전	일자	변경내용	비고
1.0	2021-06-18	원천데이터 수정	

소개

도서를 기반으로 한 원문의 핵심 내용, 의미 전달을 적절히 포함하는 요약문을 자동으로 생성하는 AI기술 개발을 위한 도서 요약 텍스트 데이터

구축목적

다양한 주제의 한국어 도서 원문으로부터 생성요약문을 도출해낼 수 있도록 인공지능을 훈련하기 위한 데이터셋

이 링크를 통해 위의 페이지에 접속할 수 있다.

[원천]도서요약_train.zip을 해제한 결과는 다음과 같다.

이름	수정한 날짜	유형	크기
기술과학	2022-12-15 오전 9:39	파일 폴더	
기타	2022-12-15 오전 9:39	파일 폴더	
사회과학	2022-12-15 오전 9:41	파일 폴더	
예술	2022-12-15 오전 9:41	파일 폴더	
[원천]도서요약_train.zip	2022-12-14 오후 11:26	ALZip ZIP File	184,320KB

2. 파일 개수

```
find .
```

명령어를 입력하는 경우, 모든 파일 목록이 출력 된다.

```
find . | cut -d/ -f2
```

명령어의 경우, 각 파일의 directory명 즉, '기술과학', '기타', '사회과학', '예술' 4가지가 파일 개수 별로 출력 된다.

```
find . | cut -d / -f2 | sort
```

명령어의 경우, 위의 결과가 정렬되어서 출력된다.

```
find . | cut -d / -f2 | sort | uniq -c
```

최종 명령어의 경우, 폴더 별 파일 개수의 통계가 출력된다.

```
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training$ find . | cut -d/ -f2 | sort | uniq -c
1 .
1 [원천]도서요약_train.zip
23919 기술과학
6754 기타
115441 사회과학
13892 예술
```

3. 주제별 빈도

```
grep '"kdc_label":' PCY_2020* | cut -d: -f3 | sort | uniq -c | sort -nr
```

위 명령어를 통해 '사회과학' 폴더를 확인한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/사회과학$ grep '"kdc_label":' PCY_2020* | cut -d: -f3 | sort | uniq -c | sort -nr
435 "안전관리",
412 "의료",
367 "국민권의 인권",
224 "보육가족및여성",
222 "법무및경찰",
213 "에너지및환경",
198 "에너지및환경개발",
149 "법무행정",
133 "문화",
98 "국립중앙도서관",
94 "국립중앙도서관",
59 "교육인사관리",
42 "일반행정",
37 "상업및공업",
26 "농림수산",
23 "농림수산",
19 "과학기술",
19 "과학기술",
19 "과학기술",
5 "철학",
5 "고등교육",
2 "기획재정",
```

추가적으로, '기술과학' 폴더에 대해 동일한 명령어를 실행한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/사회과학$ cd ../기술과학/
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/기술과학$ grep '"kdc_label":' PCY_2020* | cut -d: -f3 | sort | uniq -c | sort -nr
252 "보건의료",
126 "자연",
70 "농업, 농촌",
53 "방송통신",
49 "과학기술연구",
43 "공학, 공업일반",
16 "식품의약품안전",
2 "해양수산, 어촌",
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/기술과학$
```

4. 출판 연도별 파일 개수

```
ls | xargs grep '"published_year":' | cut -d: -f3 | sort | uniq -c
```

위 명령어를 통해 '사회과학' 폴더를 확인한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/사회과학$ ls | xargs grep '"published_year":' | cut -d: -f3 | sort | uniq -c
47 "1991",
82 "1993",
45 "1996",
21 "1997",
74 "1998",
190 "1999",
102 "2000",
105 "2001",
82 "2002",
26 "2003",
223 "2006",
16 "2007",
72 "2008",
40 "2009",
185 "2010",
373 "2011",
962 "2012",
2222 "2013",
7797 "2014",
18171 "2015",
21366 "2016",
12922 "2017",
10951 "2018",
6099 "2019",
806 "2020",
32461 null,
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/사회과학$
```

추가적으로, '기술과학' 폴더에 대해 동일한 명령어를 실행한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/기술과학$ ls | xargs grep '"published_year":' | cut -d: -f3 | sort | uniq -c
131 "2009",
198 "2010",
340 "2011",
848 "2012",
763 "2013",
1402 "2014",
4176 "2015",
6249 "2016",
2374 "2017",
899 "2018",
703 "2019",
348 "2020",
5487 null,
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/기술과학$
```

5. 출판 연도별 주제 빈도

```
ls | xargs grep '"published_year":' | cut -d: -f3 > ../published_year.txt
ls | xargs grep '"kdc_label":' | cut -d: -f3 > ../kdc_label.txt
cd ..
paste published_year.txt kdc_label.txt | grep '"2020"' | cut -f2 | sort | uniq -c | sort -nr
```

위 명령어를 순서대로 실행한 결과는 다음과 같다.

```

tjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/사회과학$ ls | xargs grep '"published_year":' | cut -d: -f3 > ../published_year.txt
tjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/사회과학$ ls | xargs grep '"kdc_label":' | cut -d: -f3 > ../kdc_label.txt
tjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/사회과학$ cd ..
tjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training$ paste published_year.txt kdc_label.txt | grep '"2020"' | cut -f2 | sort | uniq -c | sort -nr
136 "법무행정",
108 "행정학",
103 "법무법집합",
94 "국정홍보",
71 "사회학, 사회문제",
50 "유아교육, 중등교육",
46 "보육·가족및여성",
42 "일반행정",
34 "법학",
29 "정치학",
26 "경제학",
19 "외교",
16 "교육",
10 "사회과학",
8 "국방, 군사학",
7 "통계학",
6 "문학",
1 "종교, 민속학",
tjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training$

```

```

paste published_year.txt kdc_label.txt | grep '"2019"' | cut -f2 | sort | uniq -c
| sort -nr

```

다음으로 위 명령어를 실행한 결과는 다음과 같다.

```

tjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training$ paste published_year.txt kdc_label.txt | grep '"2019"' | cut -f2 | sort | uniq -c | sort -nr
519 "국민권의인권",
459 "안전관리",
439 "외교",
434 "일반행정",
402 "에너지및자원개발",
373 "경제학",
349 "교육학",
326 "법무법집합",
296 "보육·가족및여성",
280 "사회학, 사회문제",
274 "법무행정",
251 "교육노동",
243 "행정학",
234 "국정홍보",
171 "정치학",
169 "유아·초·중등교육",
133 "통계",
133 "법학",
97 "교육일반",
89 "국정홍보",
88 "평생·직업교육",
57 "법제",
49 "인간정수년",
48 "국가발전",
37 "산업·중소기업일반",
36 "지방행정·재정지원",
33 "사회과학",
20 "국방, 군사학",
16 "지역및도시",
11 "항공·공항",
6 "무역및투자유치",
5 "철도",
5 "고등교육",
4 "기획재정",
3 "금융",
tjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training$

```

6. 도서 설명의 단어 빈도

```

grep '"passage":' PCY_2020* | cut -d: -f3 | awk '{for(i=1;i<=NF;i++) print $i}' |
sort | uniq -c | sort -nr | head -20

```

위 명령어를 실행한 결과는 다음과 같다.

```

tjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/사회과학$ grep '"passage":' PCY_2020* | cut -d: -f3 | awk '{for(i=1;i<=NF;i++) print $i}' | sort | uniq -c | sort -nr | head -20
20 20
1751 있다.
1689 대한
1635 명
1467 있는
891 그
746 문
721 이
711 것으로
697 할
692 것이다.
690 위
628 이
621 문제
572 위
551 것이
547 하는
541 따라
491 또한
471 그리고
tjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/사회과학$

```

과제05: 데이터 처리 실습

7. CSV과 JSON을 위한 명령행 도구

먼저, **xsv**, **csvkit**, **jq**를 설치하였다.

A. **xsv** 설치

```
curl https://sh.rustup.rs -sSf | sh
```

github에 명시된 안내에 따라, 위 명령어로 **Cargo**를 먼저 설치한 뒤,

```
cargo install xsv
```

명령어를 입력하여 **xsv** 설치를 완료하였다.

B. **csvkit** 설치

[csvkit 홈페이지의 tutorial](#)에 명시된 대로

```
sudo pip install csvkit
```

명령어를 통해 csvkit을 설치하였다.

C. **jq** 설치



```
sudo apt-get install jq
```

jq 홈페이지의 download 에 명시된 명령어를 통해 **jq**를 설치하였다.

D. 실습

```
jq -r '.metadata.kdc_label' *.json | sort | uniq -c | sort -nr
```

위 명령어를 실행한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예술$ jq -r '.metadata.kdc_label' *.json | sort | uniq -c | sort -nr
3474 공연 예술 및 매체 예술
3065 예술
2196 문화 예술
1578 문화 예술
1477 문화체육관광 일반
1375 관광
210 음악
208 체육
100 오락, 스포츠
93 연극
52 오락, 운동
48 회화, 도화
15 건축
```

```
head -12 PCY_202006180835523551_0.json
```

위 명령어를 실행한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예술$ head -12 PCY_202006180835523551_0.json
{
  "passage_id": "PCY_202006180835523551_0",
  "metadata": {
    "doc_id": "PCY_202006180835523551",
    "doc_type": "도서",
    "doc_name": "2019 콘텐츠산업 통계조사 : 2018년 기준",
    "author": null,
    "publisher": "문화체육관광부",
    "published_year": "2020",
    "kdc_label": "문화예술",
    "kdc_code": "600"
  },
}
```

```
jq '.metadata | .kdc_label' *.json | head
```

위 명령어를 실행한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예술$ jq '.metadata | .kdc_label' *.json | head
"문화예술"
"문화예술"
"문화예술"
"문화예술"
"문화예술"
"문화예술"
"문화예술"
"문화예술"
"문화예술"
"문화예술"
"문화예술"
"문화예술"
```

```
jq '.metadata | .kdc_label, .published_year' *.json | head
```

위 명령어를 실행한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예수$ jq '.metadata | .kdc_label, .published_year' *.json | head
"공연예술 및 매체예술"
"2007"
"공연예술 및 매체예술"
"2007"
"공연예술 및 매체예술"
"2007"
"공연예술 및 매체예술"
"2007"
"공연예술 및 매체예술"
"2007"
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예수$
```

```
jq '.metadata | [.kdc_label, .published_year]' *.json | head
```

위 명령어를 실행한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예수$ jq '.metadata | [.kdc_label, .published_year]' *.json | head
[
  "공연예술 및 매체예술",
  "2007"
]
[
  "공연예술 및 매체예술",
  "2007"
]
[
  "공연예술 및 매체예술",
  "2007"
]
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예수$
```

```
jq '.metadata | [.doc_id, .publisher, .published_year, .kdc_label]' *.json | jq -r '@tsv' | head
```

위 명령어를 실행한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예수$ jq '.metadata | [.doc_id, .publisher, .published_year, .kdc_label]' *.json | jq -r '@tsv' | head
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
CNTS-00047966809 연극과인간 2007 공연예술 및 매체예술
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예수$
```

```
jq '.metadata | [.published_year, .kdc_label]' *.json | jq -r '@csv' | sort | uniq -c | sed -r 's/^[ ]+([0-9]+) /\1,/' > label_year_freq.txt
head label_year_freq.txt
```

위 명령어를 실행한 결과는 다음과 같다.

```
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예수$ jq '.metadata | [.published_year, .kdc_label]' *.json | jq -r '@csv' | sort | uniq -c | sed -r 's/^[ ]+([0-9]+) /\1,/' > label_year_freq.txt
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예수$ head label_year_freq.txt
538,"2000","공연예술 및 매체예술"
1122,"2001","공연예술 및 매체예술"
442,"2002","공연예술 및 매체예술"
807,"2003","공연예술 및 매체예술"
170,"2004","공연예술 및 매체예술"
4,"2004","음악, 스포츠"
60,"2005","공연예술 및 매체예술"
48,"2005","과학, 도화"
272,"2007","공연예술 및 매체예술"
185,"2007","음악"
jtjun@JTJ-MacBook: /mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예수$
```

```
xsv sort -n -s1 -NR label_year_freq.txt | head
```

위 명령어를 실행한 결과는 다음과 같다.

```

jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예술$ xsv sort -n -s1 -NR label_year_freq.txt | head
1122,2001,공연예술 및 매체예술
990,2018,예술
807,2003,공연 예술 및 매체 예술
706,,문화 예술
538,2000,공연 예술 및 매체 예술
452,2016,문화체육관광일반
447,2010,예술
447,2019,문화체육
442,2002,공연 예술 및 매체 예술
441,2015,문화체육관광일반
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예술$

```

```
xsv sort -n -s3 label_year_freq.txt | xsv table | head
```

위 명령어를 실행한 결과는 다음과 같다.

```

jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예술$ xsv sort -n -s3 label_year_freq.txt | xsv table | head
15      2015      건축 예술
538     2000     공연 예술 및 매체 예술
1122    2001     공연 예술 및 매체 예술
442     2002     공연 예술 및 매체 예술
807     2003     공연 예술 및 매체 예술
170     2004     공연 예술 및 매체 예술
60      2005     공연 예술 및 매체 예술
272     2007     공연 예술 및 매체 예술
19      2009     공연 예술 및 매체 예술
44      2010     공연 예술 및 매체 예술
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예술$

```

```
xsv stats -n -s1 --cardinality label_year_freq.txt | xsv table
```

위 명령어를 실행한 결과는 다음과 같다.

```

jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예술$ xsv stats -n -s1 --cardinality label_year_freq.txt | xsv table
field type sum min max min_length max_length mean stddev cardinality
0 Integer 13891 1 1122 1 4 198.44285714285712 229.90753642741345 61
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training/예술$

```

8. 파일 정리

```
find . -type f | cut -d/ -f3 | sort | uniq -c | awk '$1 > 1'
```

spec에 나와있는 명령어로는 오류가 나와, 메뉴얼에 있는대로 **file** 대신 **f**를 사용하였다.

```

jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training$ find . -type f | cut -d/ -f3 | sort | uniq -c | awk '$1 > 1'
3
jtjun@JTJ-MacBook:~/mnt/c/Users/JTJ/Desktop/2022/NLP/hw04/도서자료 요약/Training$

```

3이 출력되었다. 즉, 같은 파일이 3개 있는 것으로 보인다.

```

mkdir all
find 기술과학/ -type f -exec mv {} all/ \;
find 기타/ -type f -exec mv {} all/ \;
find 사회과학/ -type f -exec mv {} all/ \;
find 예술/ -type f -exec mv {} all/ \;

```

위에서 **file** 대신 **f**를 사용한 것과 같이, 변경하여 실행하였다.