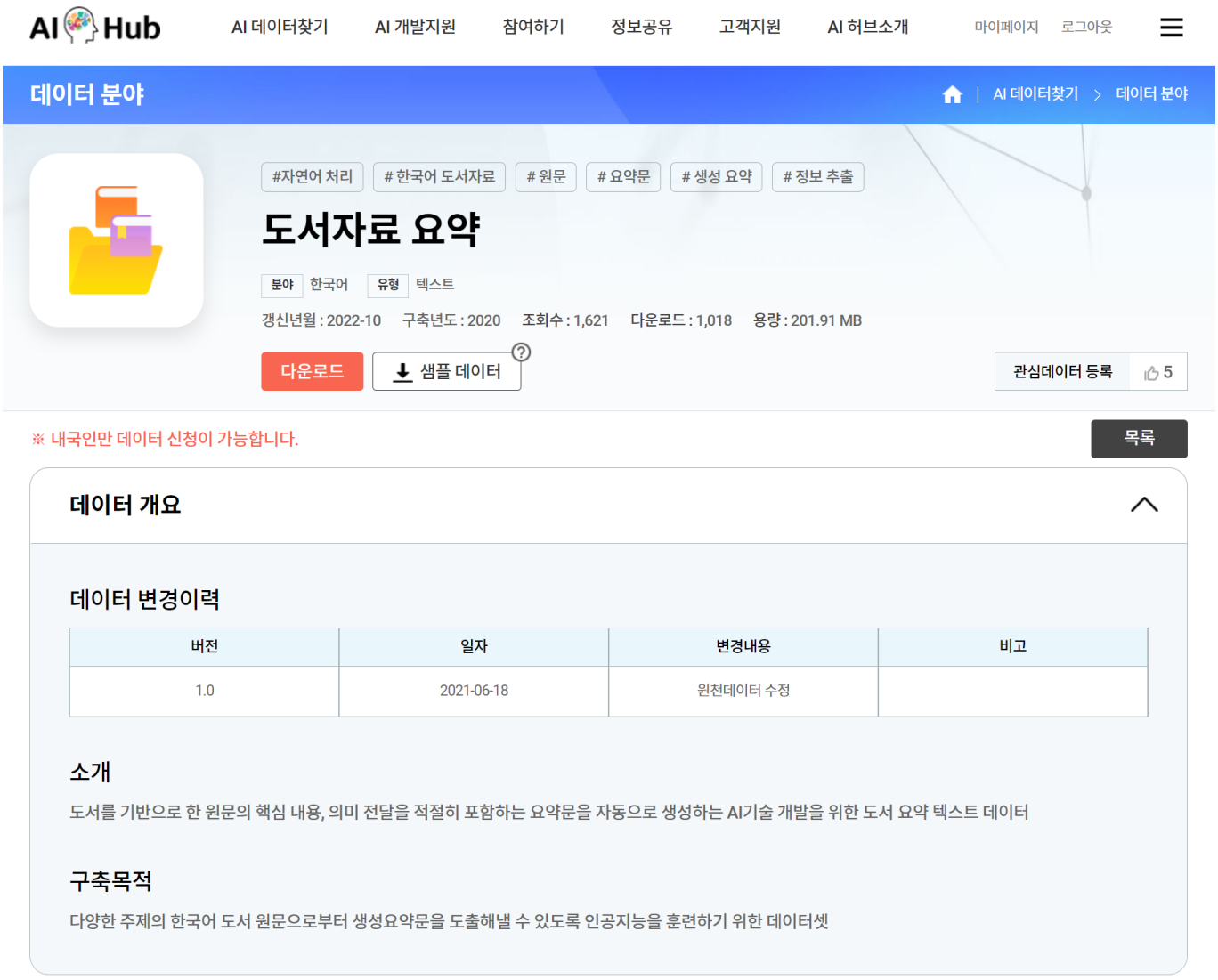
과제04: 데이터 처리 실습

언어 데이터 처리 2022-12-14 장태준 [컴퓨터 공학부 2017-17018]
https://github.com/jtjun/NLP/blob/main/hw04/hw04.md

1. 데이터

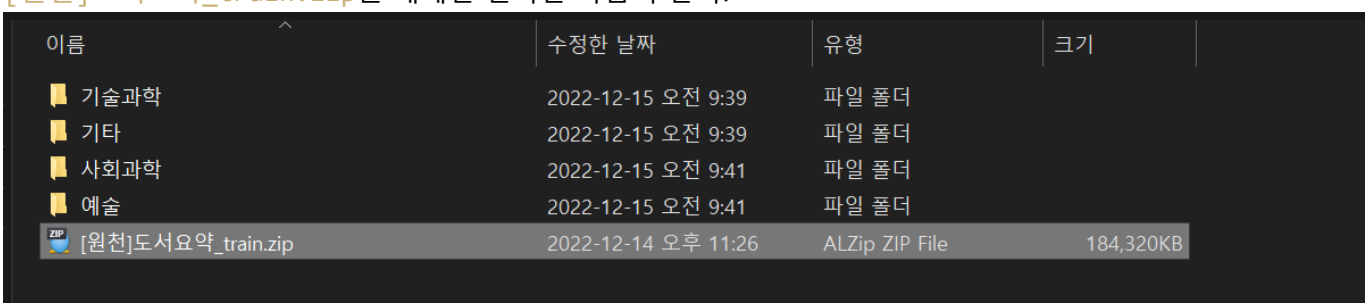
A. 소제목

Ai Hub에 가입/로그인한 뒤, 데이터를 다운 받았다.



이 링크를 통해 위의 페이지에 접속할 수 있다.

`[원천]도서요약_train.zip`을 해제한 결과는 다음과 같다.



2. 파일 개수

```
find .
```

명령어를 입력하는 경우, 모든 파일 목록이 출력 된다.

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

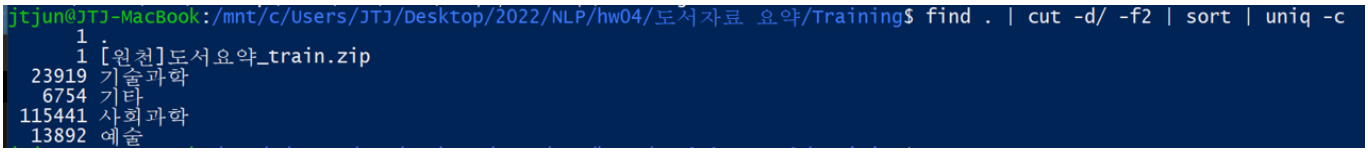
명령어의 경우, 각 파일의 directory명 즉, '기술과학', '기타', '사회과학', '예술' 4가지가 파일 개수 별로 출력 된다.

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

명령어의 경우, 위의 결과가 정렬되어서 출력된다.

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

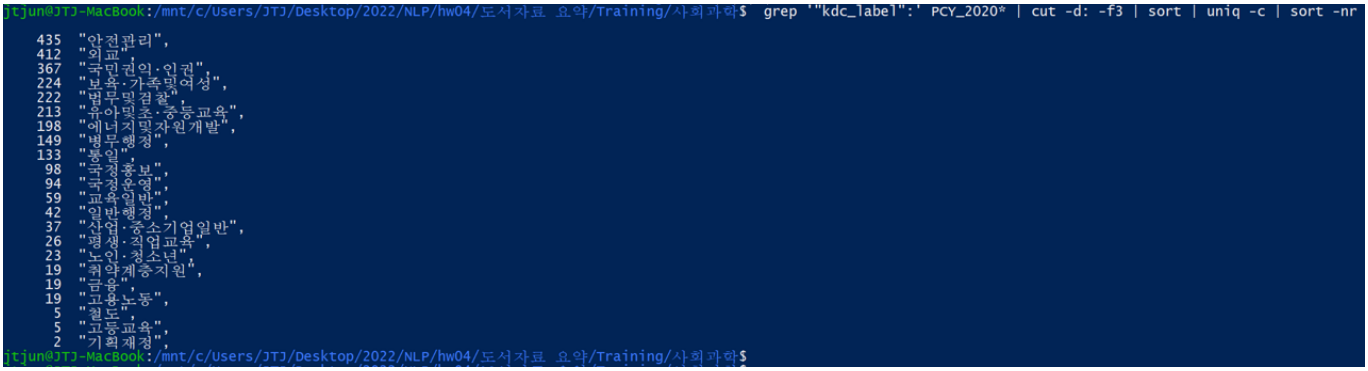
최종 명령어의 경우, 폴더 별 파일 개수의 통계가 출력된다.



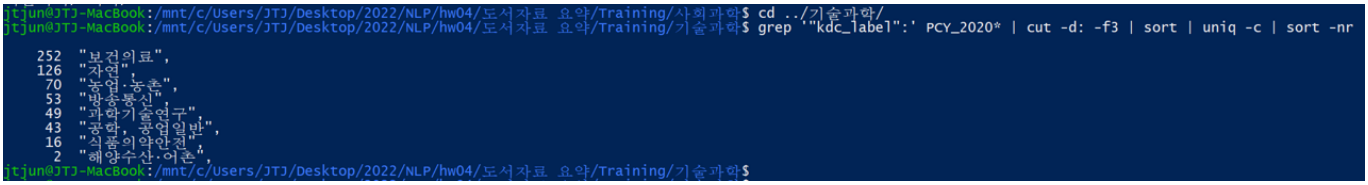
3. 주제별 빈도

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

위 명령어를 통해 '사회과학' 폴더를 확인한 결과는 다음과 같다.



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

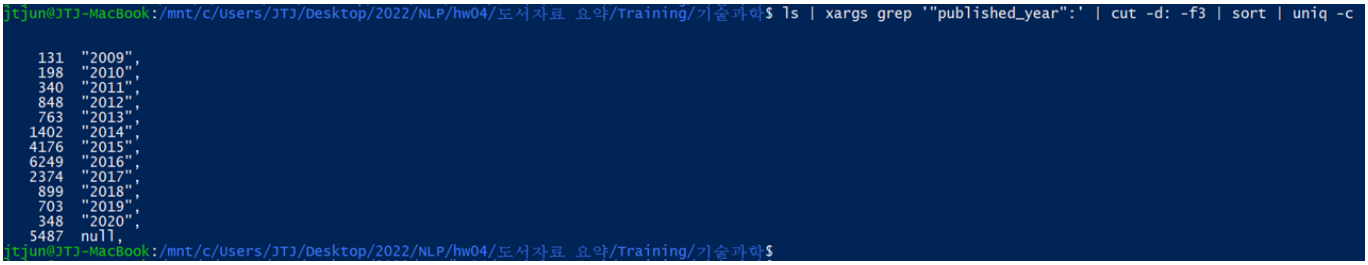


4. 출판 연도별 파일 개수

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

위 명령어를 통해 '사회과학' 폴더를 확인한 결과는 다음과 같다.

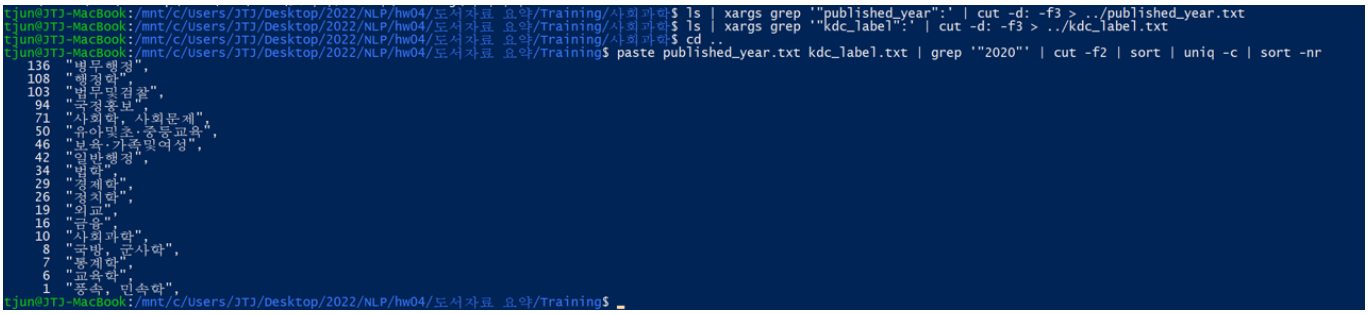
추가적으로, '기술과학' 폴더에 대해 동일한 명령어를 실행한 결과는 다음과 같다.



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.




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

다음으로 위 명령어를 실행한 결과는 다음과 같다.

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.

과제05: 데이터 처리 실습

7. CSV과 JSON을 위한 명령행 도구

먼저, `xsv`, `cvkit`, `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. `cvkit` 설치

csvkit 홈페이지의 tutorial에 명시된 대로

```
sudo pip install csvkit
```

명령어를 통해 csvkit을 설치하였다.

C. `jq` 설치

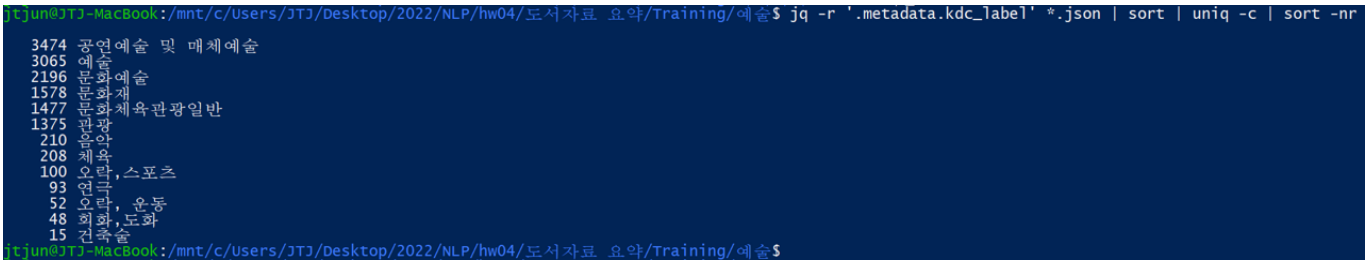

```
sudo apt-get install jq
```

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

D. 실습

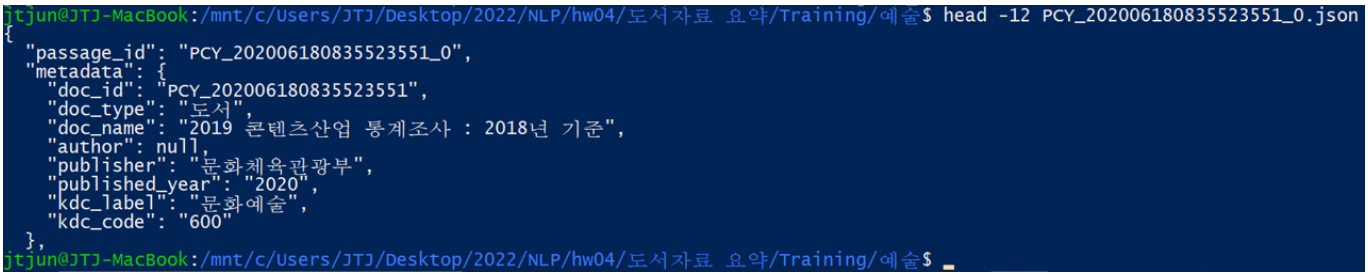
```
jq -r '.metadata.kdc_label' *.json | sort | uniq -c | sort -nr
```

위 명령어를 실행한 결과는 다음과 같다.



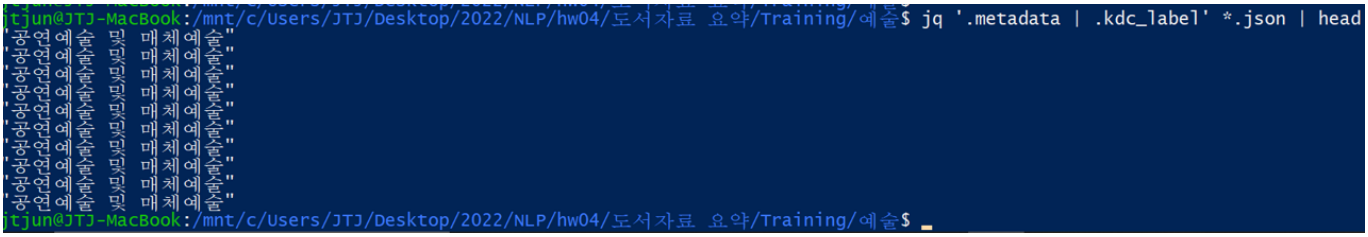

```
head -12 PCY_202006180835523551_0.json
```

위 명령어를 실행한 결과는 다음과 같다.



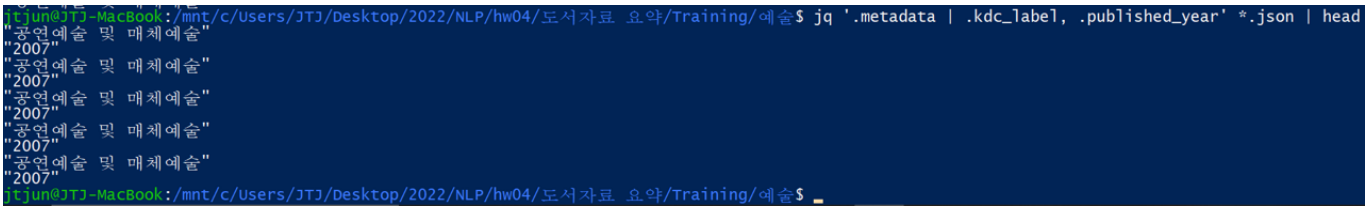

```
jq '.metadata | .kdc_label' *.json | head
```

위 명령어를 실행한 결과는 다음과 같다.



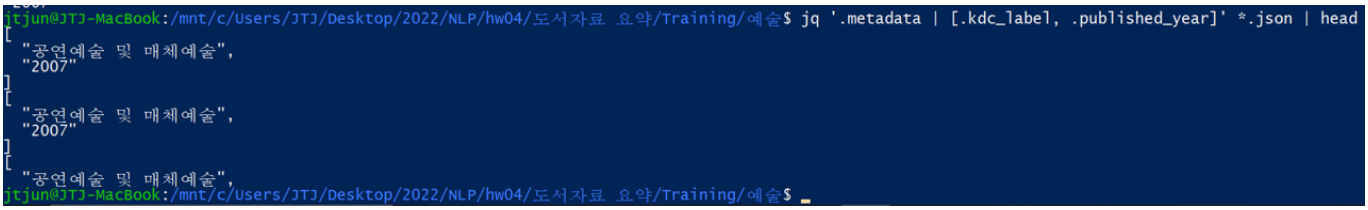

```
jq '.metadata | .kdc_label, .published_year' *.json | head
```

위 명령어를 실행한 결과는 다음과 같다.




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

위 명령어를 실행한 결과는 다음과 같다.

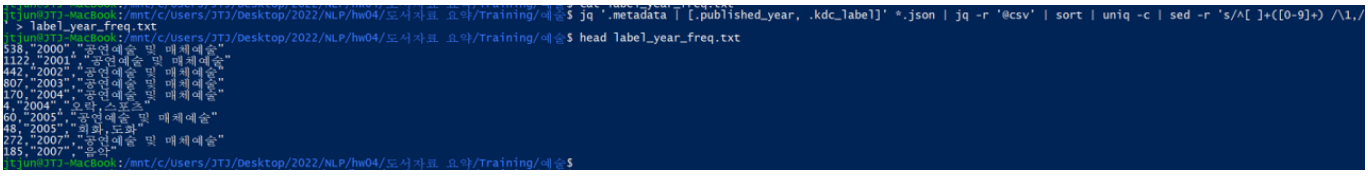



```
jq '.metadata | [.doc_id, .publisher, .published_year, .kdc_label]' *.json | jq -r
'@tsv' | head
```

위 명령어를 실행한 결과는 다음과 같다.


```
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.




```
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,공연예술 및 매체예술
706,,문화예술
538,2000,공연예술 및 매체예술
452,2016,문화체육관광일반
447,2010,예술
447,2019,문화재
442,2002,공연예술 및 매체예술
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  건축술
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
60    2005  공연예술 및 매체예술
272   2007  공연예술 및 매체예술
19    2009  공연예술 및 매체예술
44    2010  공연예술 및 매체예술
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.