2023학년도 2학기 언어데이터과학

제4강 awk 기초

박수민

서울대학교 인문대학 언어학과

2023년 9월 13일 수요일

지난 시간에 배운 것

1. Linux 명령 기초

오늘의 목표

- awk 언어를 사용하여 고대 로마 요리책을 레시피 단위로 분할할 수 있다.

오늘의 도구

- VS Code > Terminal

지난 시간에 한 일

- 텍스트 전체에서 “PEPPER” 검색하기
- 텍스트 전체에서 “PEPPER”가 사용된 횟수 세기
- 텍스트 전체에서 “PEPPER”와 “CUMIN”이 함께 사용된 경우 찾기

한계

`grep`은 기본적으로 행 단위로 작동한다.

시나리오 1

PEPPER와 HONEY가 함께 쓰인 경우를 찾고 싶다!

```
grep ”PEPPER” recipes.txt | grep ”HONEY”
```

예시: 찾을 수 있는 경우

```
[33] WINE SAUCE FOR TRUFFLES
   _ŒNOGARUM _[1]_ IN TUBERA_

PEPPER, LOVAGE, CORIANDER, RUE, BROTH, HONEY AND A LITTLE OIL.

ANOTHER WAY: THYME, SATURY, PEPPER, LOVAGE, HONEY, BROTH AND OIL.
```

시나리오 1

PEPPER와 HONEY가 함께 쓰인 경우를 찾고 싶다!

```
1  grep ”PEPPER” recipes.txt | grep ”HONEY”
```

예시: 찾을 수 없는 경우

```
1  [35] HYPOTRIMA [1]
2     _HYPOTRIMA_
3
4  [Tor. HYPOTRIMA, MEANING IN LATIN A PERFECT MESS OF POTAGE, REQUIRES
5  THIS]: PEPPER, LOVAGE, DRY MINT, PIGNOLIA NUTS, RAISINS, DATE WINE,
6  SWEET CHEESE, HONEY, VINEGAR, BROTH, WINE, OIL, MUST OR REDUCED MUST
7  [2]
```

“PEPPER”와 “HONEY”가 각기 다른 행에 있으므로 grep으로 검색할 수 없다!

시나리오 2

PEPPER가 들어간 요리가 모두 몇 개인지 알고 싶다!

```
1  grep -o "PEPPER" recipes.txt | uniq -c
```

예시: 중복되어 집계되는 경우

```
1  [61] LUCANIAN SAUSAGE
2     _LUCANICÆ_
3
4  LUCANIAN SAUSAGE [or meat pudding] ARE MADE SIMILAR TO THE ABOVE:
5  CRUSH PEPPER, CUMIN, SAVORY, RUE, PARSLEY, CONDIMENT, LAUREL BERRIES
6  AND BROTH; MIX WITH FINELY CHOPPED [fresh Pork] AND POUND WELL WITH
7  BROTH. TO THIS MIXTURE, BEING RICH, ADD WHOLE PEPPER AND NUTS. WHEN
8  FILLING CASINGS CAREFULLY PUSH THE MEAT THROUGH. HANG SAUSAGE UP TO
9  SMOKE.
```

"PEPPER"가 한 레시피에서 두 개 행에 등장하므로 사용 빈도가 1회가 아닌 2회가 된다!

기타 시나리오

- PEPPER가 들어간 요리의 이름을 알고 싶다.
- PEPPER가 들어간 요리의 레시피를 알고 싶다.
- ...

이 데이터에서 정보를 더 많이 뽑아내려면

텍스트를 요리 단위로 분할해야 한다.

어떻게 분할하는가?

AWK 프로그램의 전제

- 텍스트를 레코드(record)의 연쇄로 처리한다.
- 하나의 레코드는 한 개 이상의 필드(field)로 이루어져 있다.
  - 모든 레코드의 필드 개수가 같을 필요는 없다.

Record 1	Field 1	Field 2		
Record 2	Field 1	Field 2	Field 3	
Record 3	Field 1	Field 2	Field 3	
Record 4	Field 1	Field 2	Field 3	Field 4
Record 5	Field 1	Field 2	Field 3	

example1.txt

```
1 1 강은수 2020-10101 영어영문학과
2 2 조재영 2021-12121 언어학과
3 3 이한솔 2022-12211 수리과학부
4 ...
```

위의 텍스트는 오른쪽의 레코드와 필드로 해석할 수 있다.

예시

- Record 1
  - Field 1: 1
  - Field 2: 강은수
  - Field 3: 2020-10101
  - Field 4: 영어영문학과
- Record 2
  - Field 1: 2
  - Field 2: 조재영
  - Field 3: 2021-12121
  - Field 4: 언어학과
- Record 3
  - Field 1: 3
  - Field 2: ...

AWK 활용

학생들의 이름만 출력하기
⇒ 각 레코드에서 2번 필드($2)를 출력하기

예시

```
$ awk 'print $2' example1.txt
강은수
조재영
이한솔
```

기본적으로 한 행을 한 레코드로 처리한다.

example1.txt

```
1 1 강은수 2020-10101 영어영문학과
2 2 조재영 2021-12121 언어학과
3 3 이한솔 2022-12211 수리과학부
4 ...
```

example2.txt

```
1  1
2  강은수
3  2020-10101
4  영어영문학과
5
6  2
7  조재영
8  2021-12121
9  언어학과
10
11 3
12 이한솔
13 2022-12211
14 수리과학부
15 ...
```

값들이 다른 방법으로 배치되어 있다면 어떨까?

AWK 활용

학생들의 이름만 출력하기
⇒ 각 레코드에서 2번 필드($2)를 출력하기

레코드의 경계(RS)와 필드의 경계(FS)를 구체적으로 명시해 준다.

예시

```
$ awk 'BEGIN{RS="\n\n";FS="\n"} {print $2}' example2.txt
강은수
조재영
이한솔
```

같은 작업을 Python에서 한다고 생각해 보자...

AWK 프로그래밍으로 Apicius 텍스트를 요리 단위로 분할해 보자.

실습 코드

https://github.com/suparklingmin/LingDataSci2023/blob/main/notes/04-20230913.MD

이번 시간에 배운 것

1. awk 명령어를 활용하여 텍스트의 내부 구조 탐색하기