Lecture 06 - 정규표현식 #1

개요

- 간략 복습
- 파일과 유니코드
- CLI에서 정규표현식 사용하기
- R에서 정규표현식 사용하기

간략 복습 (30m)

CLI와 R:

- **R**: 프로그래밍 언어의 한 종류. 찰흙과 유사. 강력하지만 상대적으로 손이 많이 가는 편
- **CLI**: 작은 도구 모음. 레고 블럭과 유사. 프로그래밍 언어에 비해 자유도가 떨어지지만 손이 덜가는 편
- **R에서 CLI를 호출**:찰흙 작품 안에 레고 블럭을 넣는 것과 유사
- **CLI에서 R을 이용**:레고 조립을 하는데 원하는 블럭이 없어서 만드는 것과 유사
- 꼭 둘 다 배워야할까? 두 개는 시작에 불과. R, CLI, Python, Excel, Tableau, Google Analytics 등 수많은 도구가 존재. 쓸 줄 아는 도구가 많을수록 삶이 편해짐

AWS EC2 접속하기 (윈도)

1. 시작메뉴에서 PuTTY 실행
2. “fastcamp” 더블클릭

AWS EC2 접속하기 (맥)

1. Terminal 앱에서 다음 명령을 입력:

```
ssh -i ~/.ssh/fastcamp.pem ubuntu@원격컴퓨터주소
```

CLI 간단 실습:

```
// 명령행 인자(command-line parameter)를 표준 출력으로 보내기
echo Hello

// 출력 전환 기호를 써서 표준 출력의 내용을 파일로 저장하기
echo Hello > greeting.txt

// 파일이 생성되었는지 확인하기
```

```
ls

// 방금 생성한 파일 내용 출력하기
cat greeting.txt

// > 기호 대신 >> 기호를 써서 덧붙이기
echo Hello, again >> greeting.txt
echo Goodbye >> greeting.txt
echo Goodbye, again >> greeting.txt

// 파일 내용 다시 확인하기
cat greeting.txt

// again 이라는 문자열이 나오는 줄만 걸러내기 (파이프)
cat greeting.txt | grep again

// again 이라는 문자열이 나오는 줄이 총 몇 줄인지 세어보기
cat greeting.txt | grep again | wc -l

// 오만과 편견 다운로드
curl http://s.g15e.com/pride.txt > pride.txt

// Darcy가 나오는 문장들 출력하기
cat pride.txt | grep Darcy

// Darcy가 몇 번째 줄에 많이 나올까?
cat pride.txt | grep -n Darcy

// 줄번호만 뽑아내기
cat pride.txt | grep -n Darcy | cut -f1 -d:

// 저장
cat pride.txt | grep -n Darcy | cut -f1 -d: > darcy.txt
```

과제

- http://s.g15e.com/darcy.txt 는 “Darcy”가 나오는 줄번호만 담겨있는 텍스트 문서입니다.

- R로 Stem-and-leaf plot을 그려서 전체 소설 내에서 Darcy가 언급되는 빈도의 분포를 시각화해봅시다:

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

휴식 (10m)

파일과 유니코드 (10m)

컴퓨터는 0과 1밖에 모른다던데… 문서를 어떻게 저장하지?

- 0과 1로 세상의 모든 자연수 나타내기: 10진수, 4진수, 2진수
- 자연수로 문자를 나타내기: 아스키(ASCII)와 유니코드(Unicode) 문자셋
- 줄바꿈도 문자다
- 파일에 담긴 것은 그저 긴 0과 1의 조합. 파일은 직사각형 모양이 아니다
- 모니터에 그려지는 그림도 마찬가지

정규표현식 기초 (50m)

일단 예시 파일 다운로드:

```
// 저장
curl http://s.g15e.com/tweets.txt > tweets.txt

// 내용 살펴보기
head tweets.txt
tail tweets.txt
```

실습: “엄마”를 언급하는 트윗을 출력해보기

그런데, 엄마 또는 아빠를 언급하는 트윗을 함께 출력하려면 어떻게 할까?

```
// 엄마 또는 아빠를 언급하는 트윗
cat tweets.txt | grep -P "(엄마|아빠)"
```

여기에 쓰인 문법 “(엄마|아빠)”이 정규표현식

정규표현식이란?

- 문자열의 패턴을 기술하는 작은 언어
- 역사가 오래되어 여러 사투리가 존재함. Perl이라는 언어에서 구현한 사투리가 가장 강력함. grep 명령의 -P 옵션은 Perl 버전의 정규표현식 모드를 쓰겠다는 의미

Groups and Ranges	
.	Any character except new line (\n)
(a|b)	a or b
(...)	Group
(?:...)	Passive (non-capturing) group
[abc]	Range (a or b or c)
[^abc]	Not (a or b or c)
[a-q]	Lower case letter from a to q
[A-Q]	Upper case letter from A to Q
[0-7]	Digit from 0 to 7
\x	Group/subpattern number "x"

Ranges are inclusive.

Character Classes	
\c	Control character
\s	White space
\S	Not white space
\d	Digit
\D	Not digit
\w	Word
\W	Not word
\x	Hexadecimal digit
\O	Octal digit

Quantifiers			
*	0 or more	{3}	Exactly 3
+	1 or more	{3,}	3 or more
?	0 or 1	{3,5}	3, 4 or 5

Add a ? to a quantifier to make it ungreedy.

참고자료

- 정규표현식 사전
  http://www.nextree.co.kr/p4327/
- 정규표현식 커닝 페이퍼
  http://www.cheatography.com/davechild/cheat-sheets/regular-expressions/
- Stem-and-leaf display
  https://en.wikipedia.org/wiki/Stem-and-leaf_display