Lecture 07 - 정규표현식 #2

개요

- 복습 겸 과제 풀이
- R에서 정규표현식 사용하기
- 정규표현식을 체계적으로 이해하기
- R에서 쇼핑몰 홈페이지에 접속해서 원하는 데이터 뽑아내기

복습 겸 과제 풀이 (20m)

오만과 편견 중 남자 주인공 “Darcy” 빈도 분포:

```
// 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

// R에서 stem-and-leaf plot 그리기
stem(as.integer(readLines(url("http://s.g15e.com/darcy.txt"))))

// R에서 stem-and-leaf plot 그리기 (풀어서 쓰면…)
conn <- url("http://s.g15e.com/darcy.txt")
lines <- readLines(conn)
close(conn)
nums <- as.integer(lines)
stem(nums)
```

R에서 다 하려면?

```
conn <- url("http://s.g15e.com/pride.txt")
lines <- readLines(conn)
close(conn)
```

```
nums <- grep("Darcy", lines, perl=TRUE)
stem(nums)
```

이렇게 좋은데 왜 CLI에서 할까?

- 처리 방식의 차이
- 13주차에 다룰 예정

정규표현식을 체계적으로 이해하기 (20m)

무슨 분야이건 역사적 배경과 맥락을 알면 더 쉽게, 더 깊게 이해할 수 있음.

정규표현식은 어디에서 튀어나왔나? 어떻게 하면 잘 이해할 수 있나?

- 형식 언어(formal language)
- 유한 상태 기계(finite state machine)

이메일 주소 찾아내기:

```
\S+@\S+(\.\S+)+
```

말로 풀어쓰면?

1. 공백이 아닌 글자가 한 번 이상 나온다
   \S+

2. “@” 문자가 한 번 나온다
   @

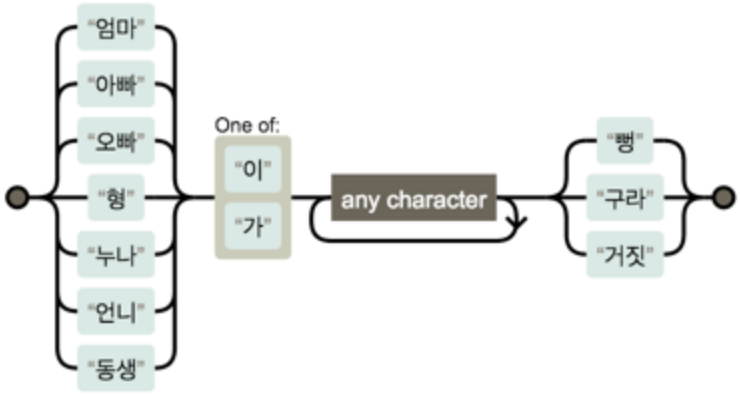
3. 공백이 아닌 글자가 한 번 이상 나온다
   \S+

4. “.” 문자가 한 번 나오고 공백이 아닌 글자가 한 번 이상 나오는게 한 번 이상 반복된다
   (\.\S+)+

더 쉽게 이해하려면? http://regexper.com/

가족 중 거짓말을 가장 많이 하는 사람? 좀 더 정확히 말하면 “거짓말"과 함께 언급되는(co-occurrence) 사람.

```
(엄마|아빠|오빠|형|누나|언니|동생)[이가].+(뻥|구라|거짓)
```



휴식 (10m)

R에서 쇼핑몰 데이터 긁어오기 (60m)

바가지머리 Best 30: http://www.bagazimuri.com/product/list.html?cate_no=150

데이터 긁어오기:

```r
# 문자열 처리 라이브러리 설치
install.packages("stringr")
library(stringr)

# 쇼핑몰 Best 30 페이지 읽어오기
conn <- url("http://www.bagazimuri.com/product/list.html?cate_no=150")
lines <- readLines(conn)
close(conn)
```

상품명 뽑아내기:

```r
# 상품명 추출
# 1. 상품명 추출 패턴
p.name <- "cate_no=150&.+>(.+)</span></a>"
# 2. 이름 부분만 뽑아내기
names <- str_match(lines, p.name)[,2]
# 3. N/A 제거
names <- names[!is.na(names)]
```

판매가 뽑아내기:

```
# 판매가 추출
# 1. 판매가 추출 패턴
p.price <- "판매가.+?([\\d,]+)원"
# 2. 가격 부분만 뽑아내기
prices <- str_match(lines, p.price)[,2]
# 3. N/A 제거
prices <- prices[!is.na(prices)]
# 4. 쉼표 제거
prices <- gsub(",", "", prices)
# 5. 숫자로 변환
prices <- as.integer(prices)
```

잘 되었나 확인해봅시다:

```
summary(prices)
length(prices)
```

실습: 문제를 찾아서 고쳐봅시다

상품 카테고리 칼럼 추가하기:

```
# 상품 카테고리 추출
# 1. 카테고리 추출 패턴
p.category <- ".+-(.+)$"
# 2. 카테고리 부분만 뽑아내기
categories <- str_match(names, p.category)[,2]
```

합쳐서 데이터 프레임으로:

```
df <- data.frame(names, prices)
names(df) <- c('name', 'price')
```

결과물:

	name	price	category
1	오우-가디건	12000	가디건
2	롱디-팬츠	19400	팬츠
3	(bari&u)바스-티셔츠	11900	티셔츠
4	쿨러-팬츠	22000	팬츠
5	(bari&u)하와이-티셔츠	15400	티셔츠

과제: 관심있는 웹 사이트에서 흥미로운 정보를 추출해서 분석이 가능한 형태로 만들어보기