

Identity Masking with Class Roll Data

coop711

2018-03-15

Data

```
class_roll <- read.xlsx("../data/class_roll10303.xlsx",
                        sheetIndex = 1,
                        startRow = 2,
                        endRow = 162,
                        colIndex = c(3:7, 9),
                        colClasses = rep("character", 6),
                        encoding = "UTF-8",
                        stringsAsFactors = FALSE)
names(class_roll) <- c("dept", "id", "name", "year", "email", "cell_no")
```

학번 가리기

학번은 입학연도를 나타내는 첫 네자리와 개인 식별번호로 구성되어 있다. 여기서, 개인식별번호를 “9999”로 가려보자. `substr()` 을 이용하면 학번의 개인정보를 가리는 일은 한 줄의 코드로 가능하다.

```
substr(class_roll$id, start = 5, stop = 8) <- "9999"
kable(head(class_roll))
```

dept	id	name	year	email	cell_no
중국어학과	20119999	강경민	4	ssilmido@naver.com (mailto:ssilmido@naver.com)	010-9164-5954
전자공학과	20119999	강경윤	4	33169kang@hanmail.net (mailto:33169kang@hanmail.net)	010-8574-8159
컴퓨터공학과	20179999	강보경	1	kbk9818@naver.com (mailto:kbk9818@naver.com)	010-6435-5735
화학과	20149999	강소연	4	crown_girl@hanmail.net (mailto:crown_girl@hanmail.net)	010-2066-8619
경영학과	20169999	강예은	2	yeeun423@naver.com (mailto:yeeun423@naver.com)	010-8820-6892
경제학과	20129999	강정우	3	jeongugang@gmail.com (mailto:jeongugang@gmail.com)	010-7499-8710

이름 가리기

`substring()` 을 이용하면 각 이름의 2번째 글자 이후를 모두 “○○”으로 대체할 수 있다.

```
substring(class_roll$name, 2) <- "○○"
kable(head(class_roll))
```

dept	id	name	year	email	cell_no
중국어학과	20119999	강○○	4	ssilmido@naver.com (mailto:ssilmido@naver.com)	010-9164-5954
전자공학과	20119999	강○○	4	33169kang@hanmail.net (mailto:33169kang@hanmail.net)	010-8574-8159
컴퓨터공학과	20179999	강○○	1	kbk9818@naver.com (mailto:kbk9818@naver.com)	010-6435-5735
화학과	20149999	강○○	4	crown_girl@hanmail.net (mailto:crown_girl@hanmail.net)	010-2066-8619
경영학과	20169999	강○○	2	yeeun423@naver.com (mailto:yeeun423@naver.com)	010-8820-6892
경제학과	20129999	강○○	3	jeongugang@gmail.com (mailto:jeongugang@gmail.com)	010-7499-8710

전화번호 가리기

모바일 폰 번호의 끝 네 자리를 “xxxx” 로 대체한다. 정상적으로 번호가 나올 경우 열번째 글자부터 열세번째글자에 해당한다.

```
substring(class_roll$cell_no, 10, 13) <- "xxxx"
kable(head(class_roll))
```

dept	id	name	year	email	cell_no
중국어학과	20119999	강○○	4	ssilmido@naver.com (mailto:ssilmido@naver.com)	010-9164-xxxx
전자공학과	20119999	강○○	4	33169kang@hanmail.net (mailto:33169kang@hanmail.net)	010-8574-xxxx
컴퓨터공학과	20179999	강○○	1	kbk9818@naver.com (mailto:kbk9818@naver.com)	010-6435-xxxx
화학과	20149999	강○○	4	crown_girl@hanmail.net (mailto:crown_girl@hanmail.net)	010-2066-xxxx
경영학과	20169999	강○○	2	yeeun423@naver.com (mailto:yeeun423@naver.com)	010-8820-xxxx
경제학과	20129999	강○○	3	jeongugang@gmail.com (mailto:jeongugang@gmail.com)	010-7499-xxxx

전공 단위 이름 가리기

전공 단위 이름은 “학과”, “과”, “학”, “전공” 등 매우 다양한 명칭이 있으므로 `gsub()` 함수의 정규표현(regular expression)을 활용하여 “○○학과” 와 같은 방식으로 이름을 가릴 수 있다.

```
class_roll$dept <- sub("^.+$", "○○학과", class_roll$dept)
kable(head(class_roll))
```

dept	id	name	year	email	cell_no
------	----	------	------	-------	---------

dept	id	name	year	email	cell_no
○○학과	20119999	강○○	4	ssilmido@naver.com (mailto:ssilmido@naver.com)	010-9164-xxxx
○○학과	20119999	강○○	4	33169kang@hanmail.net (mailto:33169kang@hanmail.net)	010-8574-xxxx
○○학과	20179999	강○○	1	kbk9818@naver.com (mailto:kbk9818@naver.com)	010-6435-xxxx
○○학과	20149999	강○○	4	crown_girl@hanmail.net (mailto:crown_girl@hanmail.net)	010-2066-xxxx
○○학과	20169999	강○○	2	yeeun423@naver.com (mailto:yeeun423@naver.com)	010-8820-xxxx
○○학과	20129999	강○○	3	jeongugang@gmail.com (mailto:jeongugang@gmail.com)	010-7499-xxxx

e-mail 가리기

email 주소는 @ 를 사이에 두고 나뉘어진다. gsub() 함수와 정규표현(regular expression)을 활용하면 email 주소에서 서비스업체명은 그대로 두고 개인 식별이 가능한 이름 부분을 user_name 으로 대체할 수 있다. 160명 중 20명만 랜덤하게 표본 추출한다.

```
class_roll$email <- sub("^.+@", "user_name@", class_roll$email)
kable(class_roll[sample(1:nrow(class_roll), size = 25), ])
```

dept	id	name	year	email	cell_no
67	○○학과	배○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-3779-xxxx
158	○○학과	황○○	3	user_name@naver.com (mailto:user_name@naver.com)	010-5656-xxxx
77	○○학과	송○○	3	user_name@naver.com (mailto:user_name@naver.com)	010-7711-xxxx
123	○○학과	임○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-4049-xxxx
92	○○학과	유○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-3654-xxxx
137	○○학과	조○○	3	user_name@naver.com (mailto:user_name@naver.com)	010-5559-xxxx
151	○○학과	최○○	3	user_name@gmail.com (mailto:user_name@gmail.com)	010-9529-xxxx
47	○○학과	김○○	3	user_name@naver.com (mailto:user_name@naver.com)	010-5848-xxxx
100	○○학과	이○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-6351-xxxx

dept	id	name	year	email	cell_no
33	○○학과	김○○	4	user_name@gmail.com (mailto:user_name@gmail.com)	010-7209-xxxx
139	○○학과	조○○	2	user_name@naver.com (mailto:user_name@naver.com)	010-9222-xxxx
104	○○학과	이○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-6291-xxxx
154	○○학과	최○○	2	user_name@naver.com (mailto:user_name@naver.com)	010-7637-xxxx
43	○○학과	김○○	3	user_name@naver.com (mailto:user_name@naver.com)	010-4886-xxxx
42	○○학과	김○○	4	user_name@nate.com (mailto:user_name@nate.com)	010-5191-xxxx
4	○○학과	강○○	4	user_name@hanmail.net (mailto:user_name@hanmail.net)	010-2066-xxxx
54	○○학과	박○○	2	user_name@naver.com (mailto:user_name@naver.com)	010-6866-xxxx
38	○○학과	김○○	2	user_name@naver.com (mailto:user_name@naver.com)	010-5457-xxxx
76	○○학과	송○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-8914-xxxx
91	○○학과	유○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-2186-xxxx
143	○○학과	조○○	3	user_name@naver.com (mailto:user_name@naver.com)	010-3528-xxxx
18	○○학과	김○○	3	user_name@naver.com (mailto:user_name@naver.com)	010-3694-xxxx
101	○○학과	이○○	2	user_name@naver.com (mailto:user_name@naver.com)	010-4964-xxxx
57	○○학과	박○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-5311-xxxx
125	○○학과	장○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-8276-xxxx