Identity Masking with Class Roll Data

MYUNG-BIN KWAK

2020-04-04

Data

```
class_roll <- read.xlsx("../data/class_roll0303_deid.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
중국학과	20119999	강가나	4	a@naver.com (mailto:a@naver.com)	010-1164-5954
전자공학과	20119999	강나다	4	b@hanmail.net (mailto:b@hanmail.net)	010-1174-8159
컴퓨터공학과	20179999	강다라	1	c@naver.com (mailto:c@naver.com)	010-1135-5735
화학과	20149999	강라마	4	d@hanmail.net (mailto:d@hanmail.net)	010-1166-8619
경영학과	20169999	강마바	2	e@naver.com (mailto:e@naver.com)	010-1120-6892
경제학과	20129999	강사아	3	f@gmail.com (mailto:f@gmail.com)	010-1199-8710

이름 가리기

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

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

dept	id	name	year	email	cell_no
중국학과	20119999	강ㅇㅇ	4	a@naver.com (mailto:a@naver.com)	010-1164-5954
전자공학과	20119999	강ㅇㅇ	4	b@hanmail.net (mailto:b@hanmail.net)	010-1174-8159
컴퓨터공학과	20179999	강ㅇㅇ	1	c@naver.com (mailto:c@naver.com)	010-1135-5735

dept	id	name	year	email	cell_no
화학과	20149999	강ㅇㅇ	4	d@hanmail.net (mailto:d@hanmail.net)	010-1166-8619
경영학과	20169999	강ㅇㅇ	2	e@naver.com (mailto:e@naver.com)	010-1120-6892
경제학과	20129999	강ㅇㅇ	3	f@gmail.com (mailto:f@gmail.com)	010-1199-8710

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
중국학과	20119999	강ㅇㅇ	4	a@naver.com (mailto:a@naver.com)	010-1164-xxxx
전자공학과	20119999	강ㅇㅇ	4	b@hanmail.net (mailto:b@hanmail.net)	010-1174-xxxx
컴퓨터공학과	20179999	강ㅇㅇ	1	c@naver.com (mailto:c@naver.com)	010-1135-xxxx
화학과	20149999	강ㅇㅇ	4	d@hanmail.net (mailto:d@hanmail.net)	010-1166-xxxx
경영학과	20169999	강ㅇㅇ	2	e@naver.com (mailto:e@naver.com)	010-1120-xxxx
경제학과	20129999	강ㅇㅇ	3	f@gmail.com (mailto:f@gmail.com)	010-1199-xxxx

전공 단위 이름 가리기

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

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

dept	id	name	year	email	cell_no
ㅇㅇ학과	20119999	강ㅇㅇ	4	a@naver.com (mailto:a@naver.com)	010-1164-xxxx
ㅇㅇ학과	20119999	강ㅇㅇ	4	b@hanmail.net (mailto:b@hanmail.net)	010-1174-xxxx
ㅇㅇ학과	20179999	강ㅇㅇ	1	c@naver.com (mailto:c@naver.com)	010-1135-xxxx
ㅇㅇ학과	20149999	강ㅇㅇ	4	d@hanmail.net (mailto:d@hanmail.net)	010-1166-xxxx
ㅇㅇ학과	20169999	강ㅇㅇ	2	e@naver.com (mailto:e@naver.com)	010-1120-xxxx
ㅇㅇ학과	20129999	강ㅇㅇ	3	f@gmail.com (mailto:f@gmail.com)	010-1199-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 표본추출한다.

```
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
34	ㅇㅇ학과	20169999	김ㅇㅇ	2	user_name@naver.com (mailto:user_name@naver.com)	010-1122-xxxx
116	ㅇㅇ학과	20119999	이ㅇㅇ	4	user_name@naver.com (mailto:user_name@naver.com)	010-1152-xxxx
128	ㅇㅇ학과	20179999	전ㅇㅇ	1	user_name@naver.com (mailto:user_name@naver.com)	010-1132-xxxx
45	ㅇㅇ학과	20179999	김ㅇㅇ	1	user_name@naver.com (mailto:user_name@naver.com)	010-1178-xxxx
158	ㅇㅇ학과	20159999	황ㅇㅇ	3	user_name@naver.com (mailto:user_name@naver.com)	010-1156-xxxx
151	ㅇㅇ학과	20149999	최ㅇㅇ	3	user_name@gmail.com (mailto:user_name@gmail.com)	010-1129-xxxx
117	ㅇㅇ학과	20169999	이ㅇㅇ	2	user_name@naver.com (mailto:user_name@naver.com)	010-1150-xxxx
152	ㅇㅇ학과	20149999	최ㅇㅇ	4	user_name@naver.com (mailto:user_name@naver.com)	010-1159-xxxx
57	ㅇㅇ학과	20179999	박ㅇㅇ	1	user_name@naver.com (mailto:user_name@naver.com)	010-1111-xxxx
43	ㅇㅇ학과	20159999	김ㅇㅇ	3	user_name@naver.com (mailto:user_name@naver.com)	010-1186-xxxx
84	ㅇㅇ학과	20149999	안ㅇㅇ	2	user_name@hanmail.net (mailto:user_name@hanmail.net)	010-1161-xxxx
148	ㅇㅇ학과	20139999	최ㅇㅇ	3	user_name@naver.com (mailto:user_name@naver.com)	010-1125-xxxx
114	ㅇㅇ학과	20149999	이ㅇㅇ	4	user_name@naver.com (mailto:user_name@naver.com)	010-1156-xxxx
111	ㅇㅇ학과	20149999	이ㅇㅇ	2	user_name@naver.com (mailto:user_name@naver.com)	010-1150-xxxx
22	ㅇㅇ학과	20139999	김ㅇㅇ	2	user_name@naver.com (mailto:user_name@naver.com)	010-1121-xxxx
26	ㅇㅇ학과	20179999	김ㅇㅇ	1	user_name@naver.com (mailto:user_name@naver.com)	010-1155-xxxx
53	ㅇㅇ학과	20129999	박ㅇㅇ	4	user_name@nate.com (mailto:user_name@nate.com)	010-1178-xxxx
37	ㅇㅇ학과	20179999	김ㅇㅇ	1	user_name@naver.com (mailto:user_name@naver.com)	010-1116-xxxx

	dept	id	name	year	email	cell_no
56	ㅇㅇ학과	20169999	박ㅇㅇ	2	user_name@naver.com (mailto:user_name@naver.com)	010-1197-xxxx
7	ㅇㅇ학과	20149999	고ㅇㅇ	4	user_name@naver.com (mailto:user_name@naver.com)	010-1184-xxxx
54	ㅇㅇ학과	20169999	박ㅇㅇ	2	user_name@naver.com (mailto:user_name@naver.com)	010-1166-xxxx
46	ㅇㅇ학과	20149999	김ㅇㅇ	2	user_name@naver.com (mailto:user_name@naver.com)	010-1173-xxxx
132	ㅇㅇ학과	20149999	정ㅇㅇ	3	user_name@naver.com (mailto:user_name@naver.com)	010-1154-xxxx
120	ㅇㅇ학과	20179999	이ㅇㅇ	1	user_name@naver.com (mailto:user_name@naver.com)	010-1173-xxxx
49	ㅇㅇ학과	20179999	나ㅇㅇ	1	user_name@naver.com (mailto:user_name@naver.com)	010-1112-xxxx

```
write.table(class_roll, file ="../data/class_roll_masked.txt")
save.image("../R/class_roll_170303_data_masked.RData")
```

–>