Identity Masking with Class Roll Data

coop711

2018-03-14

Data

```
class_roll <- read.xlsx("../data/class_roll0303.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	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

Names

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

Cell Phone Numbers

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

e-mail

email 주소는 `@` 를 사이에 두고 나뉘어지므로 앞의 방법을 그대로 적용할 수 없다. email 주소에서 서비스업체만 그대로 두고 개인 식별이 가능한 이름 부분은 `user_name` 으로 대체

```
email_na <- which(is.na(class_roll$email))
class_roll$email[email_na] <- "NA@NA"
email_split <- sapply(class_roll$email,
                      function(x) unlist(strsplit(x, split = "@")))
# email_split
email_split[1, ] <- "user_name"
class_roll$email <- apply(email_split,
                          MARGIN = 2,
                          paste, collapse = "@")
kable(head(class_roll))
```

dept	id	name	year	email	cell_no
중국학과	20119999	강ㅇㅇ	4	user_name@naver.com (mailto:user_name@naver.com)	010-9164-xxxx
전자공학과	20119999	강ㅇㅇ	4	user_name@hanmail.net (mailto:user_name@hanmail.net)	010-8574-xxxx
컴퓨터공학과	20179999	강ㅇㅇ	1	user_name@naver.com (mailto:user_name@naver.com)	010-6435-xxxx
화학과	20149999	강ㅇㅇ	4	user_name@hanmail.net (mailto:user_name@hanmail.net)	010-2066-xxxx
경영학과	20169999	강ㅇㅇ	2	user_name@naver.com (mailto:user_name@naver.com)	010-8820-xxxx
경제학과	20129999	강ㅇㅇ	3	user_name@gmail.com (mailto:user_name@gmail.com)	010-7499-xxxx