

Реорганизация на данните и обработка на СИМВОЛНИ НИЗОВЕ

Статистическа обработка на данни с R

Пламен Петров и Тодор Балабанов

Център за обучение
Институт по информационни и комуникационни технологии
Българската академия на науките

p.petrov@iit.bas.bg todorb@iinf.bas.bg

31.V.2020

Acknowledgments

These teaching materials are funded by Velbazhd Software LLC and it is partially supported by the Bulgarian Ministry of Education and Science (contract D01–205/23.11.2018) under the National Scientific Program “Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)”, approved by DCM # 577/17.08.2018.

Съдържание

- 1 Обединяване на множества от данни
 - Множества с идентична структура
 - Функция merge
 - Функция join
 - Транспониране на данните
- 2 Сложни сливания на данни и трансформация на форматите
 - Обединение на редове и колони
 - Сложни сливания на данни
 - Реформатиране на данните
- 3 Работа със символни низове
 - Формиране на текст
 - Извличане на текст
- 4 Заключение
 - Дискусия

Обединяване на множества от данни

Обединяване на множества от данни

Обединяване по редове

```
ds1 <- cbind(TV=c("BNT", "bTV", "Nova"), Channel=c(1,2,3),
Rating=c(0.1,0.3,0.2))
ds2 <- data.frame(TV=c("HBO", "VH1", "MTV"),
Channel=c(4,5,6), Rating=c(0.4,0.5,0.6),
stringsAsFactors=FALSE)
ds <- rbind(ds1, ds2)
```

Работа с външни данни

USAID множество от данни

```
setwd( "~ /Desktop" )
download.file(url="https://github.com/TodorBalabanov/Statistical-Data-Processing-with-R/raw/master/data/aid.zip",
destfile=" aid.zip" )
unzip(" aid.zip", exdir=" ./" )
```

Обработка на множество файлове

Зареждане USAID данните в R

```
library( stringr )
for(file in dir("./", pattern = "\\*.csv")) {
  name <- str_sub(string=file, start=12, end=18)
  data <- read.table(file=file.path(".", file), header=TRUE,
    sep="," , stringsAsFactors=FALSE)
  assign(x=name, value=data)
}
```


Подобрено бързодействие

Сливане на данни с join

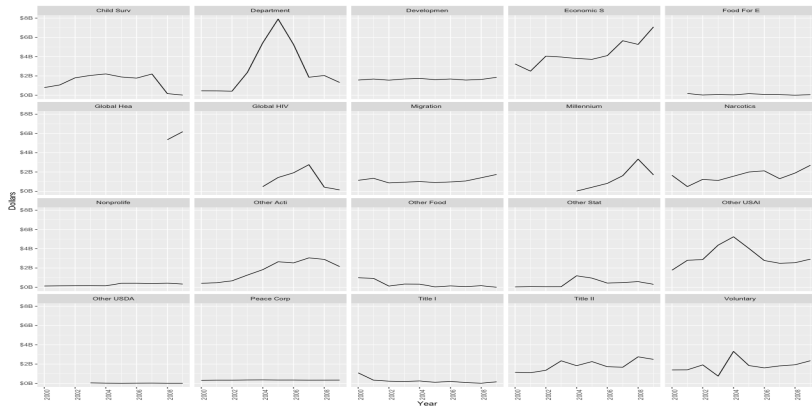
```
library(plyr)
head( join(x=Aid_90s, y=Aid_00s, by=c("Country.Name",
"Program.Name") ))
```

Предварителна обработка

От колони към редове

```
library( reshape2 )
melt00 <- melt(Aid_00s, id.vars=c(" Country.Name",
" Program.Name"), variable.name="Year", value.name=" Dollars")
head(melt00, n=3)
library( scales )
melt00$Year <- as.numeric(str_sub(melt00$Year, start=3, 6))
melt00$Program.Name <- str_sub(melt00$Program.Name,
start=1, end=10)
melt00 <- aggregate(Dollars ~ Program.Name + Year,
data=melt00, sum, na.rm=TRUE)
```


Разход на пари по програми и години



Предварительна обработка

От редове към колони

```
melt00 <- melt(Aid_00s, id.vars=c("Country.Name",
"Program.Name"), variable.name="Year", value.name="Dollars")
cast00 <- dcast(melt00, Country.Name + Program.Name ~Year,
value.var="Dollars")
```

Сложни сливания на данни и трансформация на форматите

Сложни сливания

```
library( ggplot2 )
library( readr )
library( dplyr )
colors <-
as_tibble(read.table("https://raw.githubusercontent.com/TodorBalabanov/
Data-Processing-with-R/master/data/colors.csv", header=TRUE,
sep=","))
left_join(diamonds, colors, by=c('color'='Color')) %>%
select(carat, color, price, Description, Details)
tail(right_join(diamonds, colors, by=c('color'='Color')))
inner_join(diamonds, colors, by=c('color'='Color'))
semi_join(colors, diamonds, by=c('Color'='color'))
anti_join(colors, diamonds, by=c('Color'='color'))
```


Зареждане на данни

Данни за реакциите

```
library( readr )
emotions <-
read_tsv("https://raw.githubusercontent.com/TodorBalabanov/
Statistical-Data-Processing-with-R/master/data/reaction.txt")
```

Трансформации по колони и редове

Свиване от колони в редове

```
library( tidy )
emotions %>% gather(key=Type, value=Measurement, Age, BMI,
  React, Regulate)
gather(emotions, key=Type, value=Measurement, -ID, -Test,
  -Gender)
```


Работа със символни низове

Обработка на неструктурирани документи

Достъп на HTML страници

```
library( XML )
presidents <-
readHTMLTable(" http://www.loc.gov/rr/print/list/057_chron.html",
which=3, as.data.frame=TRUE, skip.rows=1, header=TRUE,
stringsAsFactors=FALSE)
library( stringr )
years <- str_split(string=presidents$YEAR, pattern="-")
ranges <- data.frame( Reduce(rbind, years) )
names( ranges ) <- c("Start", "Stop")
```

Обработка на неструктурирани документи

Анализ на HTML страници

```
presidents <- cbind(presidents, ranges)
presidents$Start <- as.numeric( as.character(presidents$Start) )
presidents$Stop <- as.numeric( as.character(presidents$Stop) )
presidents[str_sub(string=presidents$Start, start=4, end=4) ==
1, c("YEAR", "PRESIDENT", "Start", "Stop")]
```

Обработка по шаблон

Регулярни изрази

```
presidents[ str_detect(string=presidents$PRESIDENT,
pattern=" John" ), c("YEAR", " PRESIDENT", " Start", " Stop" ) ]
sum( str_detect(presidents$PRESIDENT, "john" ) )
sum( str_detect(presidents$PRESIDENT, ignore.case(" John" )) )
```


Обработка по шаблон

Сложни регулярни изрази

```
connction <- "https://github.com/TodorBalabanov/Statistical-
Data-Processing-with-R/raw/master/data/war.rdata"
load( connction )
close( connection )
wars[ str_detect(string=wars, pattern="-") ]
str_split(string=wars, pattern="(ACAEA)|-", n=2)
```

Заклучение

Въпроси и отговори

Благодаря за вниманието!