Načítání dat do R
a
ukládání dat na pevný disk

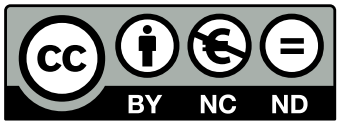
17VSADR – Skriptování a analýza dat v jazyce R

Lubomír Štěpánek[1, 2]

[1]Oddělení biomedicínské statistiky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova v Praze

ČVUT
FBMI

[2]Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze



Obsah

Pracovní složka

- zjištění, která složka je pracovní

```
getwd()
```

- výpis obsahu pracovní složky formou vektoru

```
dir()
```

- nastavení pracovní složky

```
setwd("C:/.../my_working_directory")
```

- nastavení pracovní složky dialogovým oknem

```
setwd(choose.dir())
```

Import a export volného textu

- pomocí funkcí `readLines()` a `writeLines()`
- lze tak nahrát libovolná data, která mají textovou reprezentaci

```
my_html <- readLines(
    con = paste(
        "https://predmety.fbmi.cvut.cz",
        "cs/17VSADR",
        sep = "/"
    ),
    encoding = "UTF-8"
)
```

Import a export volného textu

- uložení a načtení desktopového dokumentu

```
writeLines(                 # ukládám textový dokument
    text = paste(
        "One R to rule them all",
        "one R to find them",
        "one R to bring them all",
        "and in the darkness bind them",
        sep = "\n"   # separátor typu nový řádek
    ),
    con = "my_text.txt"
)

my_loaded_text <- readLines(
    con = "my_text.txt",
    encoding = "UTF-8"
)                           # načítám textový dokument
```

Import a export dat tabulky

- základní funkcí je `read.table()` a `write.table()`
- obě funkce mají řadu wrapperů (`read.csv()` a `write.csv()`, `read.delim()` a `write.delim()` a další)

```
1  write.table(                    # ukládám data.frame
2      x = mtcars,
3      sep = ";",
4      row.names = FALSE,
5      file = "mtcars.csv"  # anebo "mtcars.txt"
6  )
7
8  my_mtcars <- read.table(
9      file = "mtcars.csv",
10     sep = ";",
11     header = TRUE
12 )                               # načítám data.frame
```

Import a export dat tabulky

- funkce `read.table()` má spoustu užitečných argumentů

```
write.table(iris, "iris.txt")
my_iris <- read.table(
    file = "iris.txt",
    sep = " ",
    header = TRUE,
    stringsAsFactors = FALSE,
    nrows = 1,     # načte jen první řádek;
                   # může se hodit pro odhad koerce,
    check.names = FALSE,
                   # vynechá kontrolu korektnosti
                   # popisků sloupců
    colClasses = "character"
                   # přetypuje všechny sloupce
                   # na textové proměnné
)
```

Import a export dat tabulky z MS Excel® (.xlsx)

- vhodný je například balíček openxlsx
  - má výhodu, že narozdíl od balíčku např. xlsx nepotřebuje Java Tool Kit, takže dokáže najednou nahrát více souborů MS Excel®
- uložení tabulky do excelového formátu (.xlsx)

```
browseURL(paste(
  "https://raw.githubusercontent.com/LStepanek",
  "17VSADR_Skriptovani_a_analyza_dat_v_jazyce_R",
  "master/export_dat_do_ms_excelu.R", sep = "/"
))
```

- načtení excelové tabulky

```
my_data<- read.xlsx(
  xlsxFile = "moje_tabulka_je_ted_v_excelu.xlsx",
  sheet = 1,    # anebo jméno listu
  colNames = TRUE
)
```

Import bitmapového obrázku

- vhodný je například balíček `png`, `jpeg` či `raster`

```
1     my_picture <- readJPEG(
2         "__03_landmarky__.jpg"
3     )
```

- výsledkem je array o třech rozměrech
  - svislá souřadnice
  - vodorovná souřadnice
  - barevné kanály

Export konzolového výpisu do souboru

- pomocí `sink()` - `sink()` anebo `capture.output()`

```
(muzi <- rnorm(100, mean = 175, sd = 10))
(zeny <- rnorm(100, mean = 160, sd = 10))

t.test(muzi, zeny)

# výpis z konzole do textového souboru
capture.output(
    t.test(muzi, zeny),
    file = "t_test.txt"
)

# anebo
sink("tohle_je_taky_t_test.txt")
t.test(muzi, zeny)
sink()
```

Export smysluplného R-kového objektu do TEX-ového kódu

- pomocí balíčku `xtable`

```
1  library("xtable")
2
3  my_linear_model <- lm(mpg ~ hp + cyl,
       mtcars)
4
5  print(
6      xtable(my_linear_model,
7             digits = 4),
8      floating = FALSE,
9      tabular.environment = "tabular",
10     hline.after = NULL,
11     include.rownames = TRUE,
12     include.colnames = TRUE
13 )
```

Import „exotických“ souborů do R

- naším přítelem je balíček `foreign`
- podporuje načítání dat z formátů
  - Epi Info
  - Minitab
  - S
  - SAS, SPSS, STAT, Systat, Weka

```
1  library("foreign")
2
3  # import dat z SPSS
4  my_data <- read.spss(
5      file = "du1_30.sav",
6      to.data.frame = TRUE
7  )
```

Intermezzo

- jak importovat data z MS Word®?
- jak načíst do R datovou strukturu `list`?

Stažení HTML obsahu webové stránky do R (webscraping)

- načítání volného textu s kódováním

```
my_html <- readLines(
     con = paste("https://predmety.fbmi.",
                 "cvut.cz/cs/17vsadr", sep = ""),
     encoding = "UTF-8"
       # v úvahu připadá "latin1", "latin2", "ASCII"
)
```

- náhrada pevných mezer vlastním tagem

```
my_text <- gsub(" ", "M_E_Z_E_R_A",
                my_html)
```

- vymazání všech HTML tagů a entit

```
my_text <- gsub("<.*?>", "", my_text)
my_text <- gsub("&.*?;", "", my_text)
my_text <- gsub("M_E_Z_E_R_A", " ", my_text)
       # nahrazuji nezlomitelnou mezeru zlomitelnou
```

Stažení HTML obsahu webové stránky do R (webscraping)

- odstranění white strapes

```
for(i in 1:length(my_text)){
  while(grepl("  ", my_text[i])){
    my_text[i] <- gsub("  ", " ", my_text[i])
  }
}
```

- ponechání jen řádků obsahujících text

```
my_text <- my_text[!my_text %in% c("", " ")]
```

- jednoduchá úloha – extrakce emailů

```
gsub(
  "(.*?)([a-zA-Z\\.]+@[a-zA-Z]+\\.[a-zA-Z]+)(.*)"
    ,
  "\\2",
  my_text[grepl("@", my_text)]
)
```

Odstranění diakritiky ve staženém textu

- někdy se může hodit odstranění diakritiky

```
my_text <- iconv(

    my_text,
    from = "UTF-8",          # nepovinný argument
    to = "ASCII//TRANSLIT"

)
```

Literatura

Děkuji za pozornost!

lubomir.stepanek@lf1.cuni.cz
lubomir.stepanek@fbmi.cvut.cz

▸ GitHub

github.com/LStepanek/17VSADR_Skriptovani_a_analyza_dat_v_jazyce_R