Eksploracja Masywnych Danych - Analiza danych

Kajetan Zimniak & Bartosz Górka

27 October, 2019

Contents

Podsumowanie analizy

TODO

Wykorzystane biblioteki

- `knitr`
- `dplyr`
- `tidyverse`

Ustawienie ziarna generatora

Celem zapewnienia powtarzalności operacji losowania, a co za tym idzie powtarzalności wyników przy każdym uruchomieniu raportu na tych samych danych zastosowano ziarno generatora o wartości `102019`.

```
set.seed(102019)
```

Wczytanie danych z pliku

Dane zamieszczone na stronie przedmiotu w postaci pliku CSV pobieramy wyłącznie w sytuacji braku pliku w katalogu roboczym. Pozwala to nam na ograniczenie niepotrzebnego transferu danych, jeżeli plik już istnieje.

```
file_name = "sledzie.csv"
source_url = "http://www.cs.put.poznan.pl/alabijak/emd/projekt/sledzie.csv"

if (!file.exists(file_name)) {
  download.file(source_url, destfile = file_name, method = "wget")
}
```

Po ewentualnym pobraniu wczytujemy dane do pamięci.

```
library('knitr')
library('dplyr')
library('tidyverse')

content = read_csv(file_name, col_names = TRUE)

kable(head(content[0:11], n = 6), align = 'c', caption = 'Wybrane pomiary')
```

Table 1: Wybrane pomiary

X1	length	cfin1	cfin2	chel1	chel2	lcop1	lcop2	fbar	recr	cumf
0	23.0	0.02778	0.27785	2.46875	?	2.54787	26.35881	0.356	482831	0.3059879
1	22.5	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879
2	25.0	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879
3	25.5	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879
4	24.0	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879
5	22.0	0.02778	0.27785	2.46875	21.43548	2.54787	?	0.356	482831	0.3059879

Jak możemy zaobserwować, część danych nie została uzupełniona. Zamiast wartości `NA` mamy do czynienia z wartościami `?`.

Przetwarzanie brakujących danych

```
content[content == "?"] <- NA
```

Po zamianie znaku `?` na wartość `NA` musimy przeprowadzić zmianę typu danych kolumn.

```
cols = colnames(content)
content[cols] <- sapply(content[cols], as.numeric)
```

Po tych prostych operacjach wszystkie nasze kolumny mają odopowiedni typ danych, a wartości brakujące zostały zastąpione przez `NA`.