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 =
  file_name %>%
  read_csv(col_names = TRUE, na = c("", "NA", "?")) %>%
  select(-1)

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

Table 1: Wybrane pomiary

length	cfin1	cfin2	chel1	chel2	lcop1	lcop2	fbar	recr	cumf	totaln
23.0	0.02778	0.27785	2.46875	NA	2.54787	26.35881	0.356	482831	0.3059879	267380.8
22.5	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879	267380.8
25.0	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879	267380.8
25.5	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879	267380.8
24.0	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879	267380.8
22.0	0.02778	0.27785	2.46875	21.43548	2.54787	NA	0.356	482831	0.3059879	267380.8

Oryginalnie zbiór posiada znaki `?` jako oznaczenie wartości pustej (brakującej). Dzięki wykorzystaniu parametru `na` podczas wywołania funkcji `read_csv` możemy zastąpić znak `?` poprawnym oznaczeniem braku wartości `NA`.

Przetwarzanie brakujących danych

TODO - jakieś wnioskowanie tutaj? Uśrednienie wartości?

Podstawowe statystyki zbioru danych

W zbiorze danych mamy do czynienia z 52582 obserwacjami.