02. Datové objekty

Vít Gabrhel

R101

2019-09-29

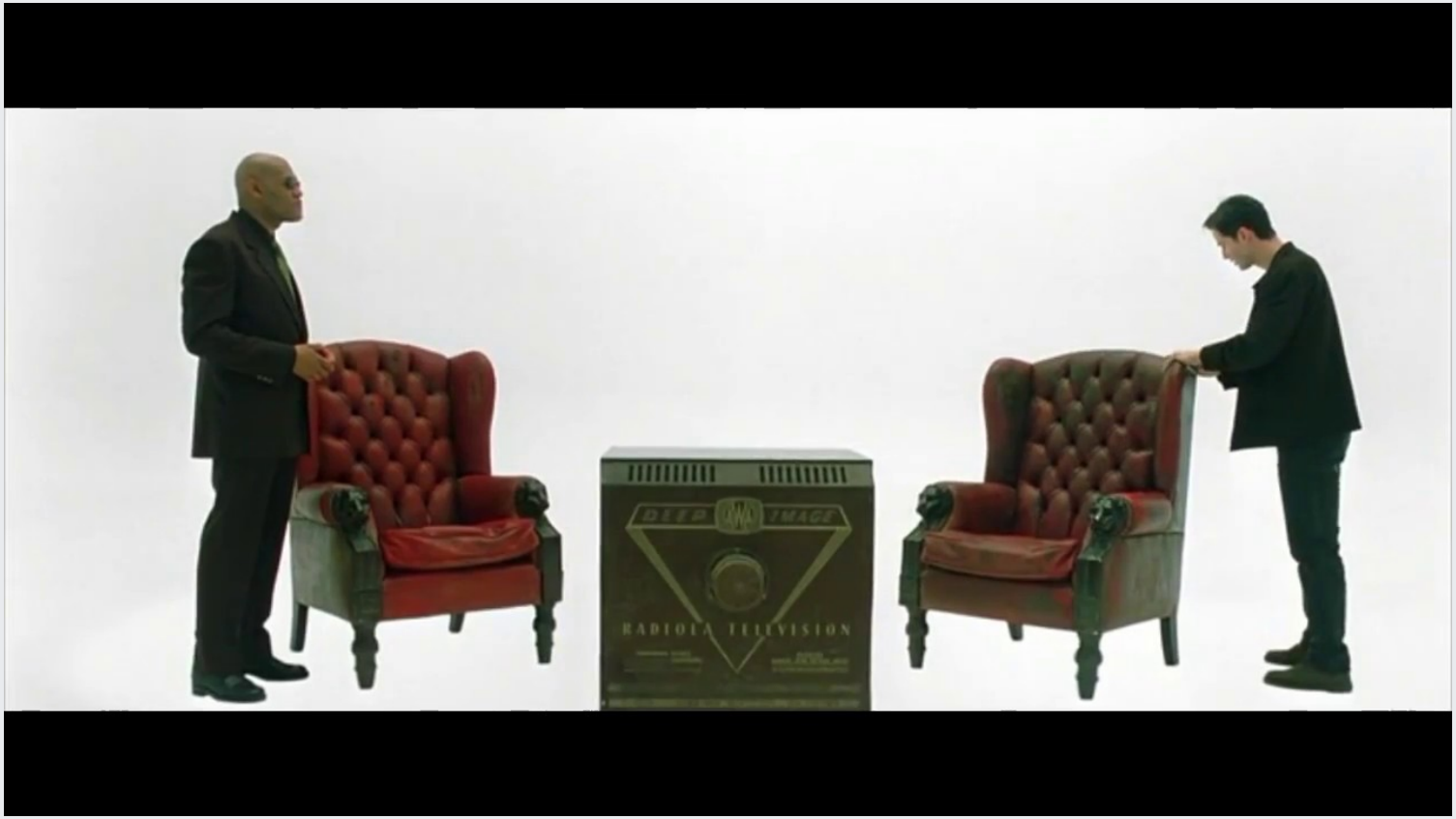
Harmonogram

1. Vector

2. Factor

3. Data Frame

Co je to objekt?



Data

Data - FiveThirtyEight

FiveThirtyEight

Politics Sports Science & Health Economics **Culture** Politics Podcast: How Does Clinton's Assessment Of 2016 Compare W...

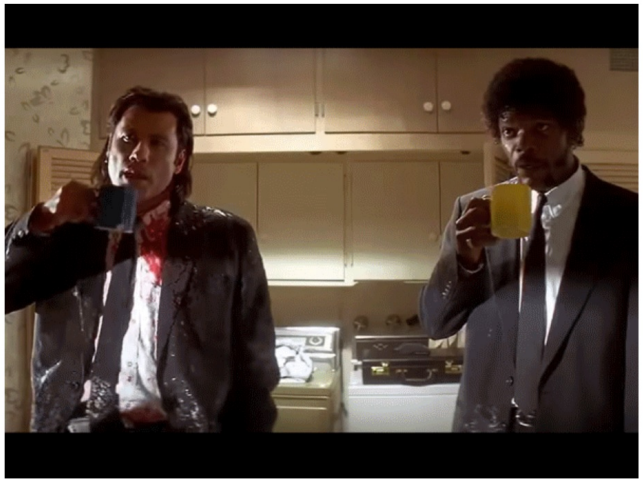
DEC. 9, 2015 AT 10:29 AM

A Complete Catalog Of Every Time Someone Cursed Or Bled Out In A Quentin Tarantino Movie

By Oliver Roeder

Filed under Word Count

Get the data on GitHub



Quentin Tarantino, John Travolta and Samuel L. Jackson in "Pulp Fiction."

RECOMMENDED

Students At Most Colleges Don't Pick 'Useless' Majors

The GOP Establishment Got What It Wanted (Sorta) In Alabama's Senate Primary

Al Gore's New Movie Exposes The Big Flaw In Online Movie Ratings

Trump Approval Ratings

UPDATED 15 HOURS AGO

56.0% Disapprove

38.7% Approve

See all approval polls

Vector

Vector

Vector je **jednoduchý** datový **objekt** o různé délce obsahující hodnoty.

```
c("Reservoir Dogs", "Pulp Fiction", "Inglorious Basterds")
```

```
## [1] "Reservoir Dogs"      "Pulp Fiction"        "Inglorious Basterds"
```

```
c(421, 469, 51)
```

```
## [1] 421 469  51
```

```
c(421, "Reservoir Dogs", "death", FALSE, 10)
```

```
## [1] "421"           "Reservoir Dogs" "death"          "FALSE"
## [5] "10"
```

Vector

Vytvoření a pojmenování vektorového objektu

Počet cursing words dle filmů

```
Words_Movie = c(421, 469, 57, 51)
```

Co je co aneb pojmenování vektorů

```
names(Words_Movie) = c("Reservoir Dogs",
                       "Pulp Fiction",
                       "Kill Bill 1",
                       "Inglorious Basterds")
```

Vector

Výběr hodnot(y) z vektoru

```
Words_Movie[c(1, 4)]
```

```
##      Reservoir Dogs Inglorious Basterds
##                 421                  51
```

```
Words_Movie[c("Reservoir Dogs", "Inglorious Basterds")]
```

```
##      Reservoir Dogs Inglorious Basterds
##                 421                  51
```

Vector

Vektorová aritmetika

Sčítání vektorů

```
Hell = c(12, 5, 3, 4)
Goddamn = c(10, 28, 7, 8)
Spirituality = Hell + Goddamn
```

Součet hodnot ve vektoru

```
Words_N <- sum(Spirituality)
Words_N
```

```
## [1] 77
```

Vector

Logické operátory

```
< for less than
> for greater than
<= for less than or equal to
>= for greater than or equal to
== for equal to each other
!= not equal to each other
```

Ve kterých filmech padlo více **cursing words**, než byl jejich průměrný počet za osm filmů?

```
Words_Movie > mean(Words_Movie)
```

```
##      Reservoir Dogs        Pulp Fiction         Kill Bill 1
##                TRUE                TRUE               FALSE
## Inglorious Basterds
##               FALSE
```

Vector

Logické operátory

```
names(Hell) = c("Reservoir Dogs", "Pulp Fiction",
                "Kill Bill 1", "Inglorious Basterds")

names(Goddamn) = c("Reservoir Dogs", "Pulp Fiction",
                   "Kill Bill 1", "Inglorious Basterds")

Hell[c(1, 4)] > Goddamn[c(1, 4)]
```

```
##      Reservoir Dogs Inglorious Basterds
##                TRUE                FALSE
```

```
Hell[c("Reservoir Dogs", "Inglorious Basterds")] !=
  Goddamn[c("Reservoir Dogs", "Inglorious Basterds")]
```

```
##      Reservoir Dogs Inglorious Basterds
##                TRUE                TRUE
```

```
names(Spirituality) = c("Reservoir Dogs", "Pulp Fiction",
                        "Kill Bill 1", "Inglorious Basterds")
```

Vector

Logické operátory

Klelo se v Pulp Fiction ve více než 50 případech?

```
PulpFiction_Celkem <- Spirituality[c(2)] > 50
PulpFiction_Celkem
```

```
## Pulp Fiction
##        FALSE
```

Zaznívalo ve filmech více slovo "Hell" nebo "Goddamn"?

```
Hell < Goddamn
```

```
##      Reservoir Dogs        Pulp Fiction         Kill Bill 1
##               FALSE                TRUE                TRUE
## Inglorious Basterds
##                TRUE
```

Factor

Factor

```
Filmy = c("Kill Bill 1", "Reservoir Dogs", "Inglorious Basterds", "Pulp Fiction")
class(Filmy)
```

```
## [1] "character"
```

Nominální kategorie

```
Factor_Filmy = as.factor(Filmy)
class(Factor_Filmy)
```

```
## [1] "factor"
```

```
levels(Factor_Filmy) <- c("Reservoir Dogs", "Pulp Fiction",
                          "Kill Bill 1", "Inglorious Basterds")
```

Ordinalizace

```
Factor_Filmy <- factor(Filmy, order = TRUE,
                  levels = c("Reservoir Dogs", "Pulp Fiction",
                             "Kill Bill 1", "Inglorious Basterds")
)
```

Data Frame

Data Frame

Data Frame je matice tak, jak ji chápeme při analýze dat

- A data frame has the **variables** of a data set as **columns** and the **observations** as **rows**

O cursing words v Tarantinových filmech už něco víme. Co ale počet mrtvých?

Budeme se věnovat **Pulp Fiction**, **Inglorious Basterds** a **Django Unchained** spolu s počtem zesnulých postav. A Přidáme k tomu známý počet **cursing words** v příslušných filmech:

```
Pulp_Fiction = c(7, 469)
Inglorious_Basterds = c(48, 58)
Django_Unchained = c(47, 262)
```

```
Filmy <- data.frame(Pulp_Fiction, Inglorious_Basterds, Django_Unchained)
View(Filmy) # otevře náhled na matici přímo v RStudio
```

Data Frame

Manipulace s řádky/sloupci

```
colnames(Filmy) <- c("Pulp Fiction", "Inglorious Basterds", "Django Unchained")
rownames(Filmy) <- c("Deaths", "Words")
View(Filmy)
```

```
rowSums(Filmy)
```

```
## Deaths  Words
##    102    789
```

```
colSums(Filmy)
```

```
##        Pulp Fiction Inglorious Basterds    Django Unchained
##                 476                 106                 309
```

Data Frame

Jak do matice přidat sloupec / řádek?

- Skrze příkaz **cbind() / rbind()**

Filmy si rozdělíme z hlediska období tvorby (**90s**, **00s** a **10s**) s kódy "0", "1" a "2":

```
Period = c(0, 1, 2)

Filmy_Period <- rbind(Filmy, Period)
rownames(Filmy_Period) <- c("Deaths", "Words", "Period")
```

Jak příkazem zjistit aktivní objekty?

```
ls()
```

```
##  [1] "Django_Unchained"    "Factor_Filmy"        "Filmy"
##  [4] "Filmy_Period"        "Goddamn"             "Hell"
##  [7] "Inglorious_Basterds" "Period"              "Pulp_Fiction"
## [10] "PulpFiction_Celkem"  "Spirituality"        "Words_Movie"
## [13] "Words_N"
```

Data Frame

Jak vybrat konkrétní prvky z matice?

- Similar to vectors, you can use the square brackets [ ] to select one or multiple elements from a data frame.
- Whereas vectors have one dimension, data frames have two dimensions. You should therefore use a comma to separate that what to select from the rows from that what you want to select from the columns. For example:
  - `Filmy_Period[1,2]` selects the element at the first row and second column.
  - `Filmy_Period[1:3,2:3]` results in a matrix with the data on the rows 1, 2, 3 and columns 2 and 3.
- If you want to select all elements of a row or a column, no number is needed before or after the comma, respectively:
  - `Filmy_Period[,1]` selects all elements of the first column.
  - `Filmy_Period[1,]` selects all elements of the first row.

Data Frame

Jaký byl průměrný počet mrtvých ve sledovaných filmech?

```
Mean_Dead = as.numeric(Filmy_Period[1,])
mean(Mean_Dead)
```

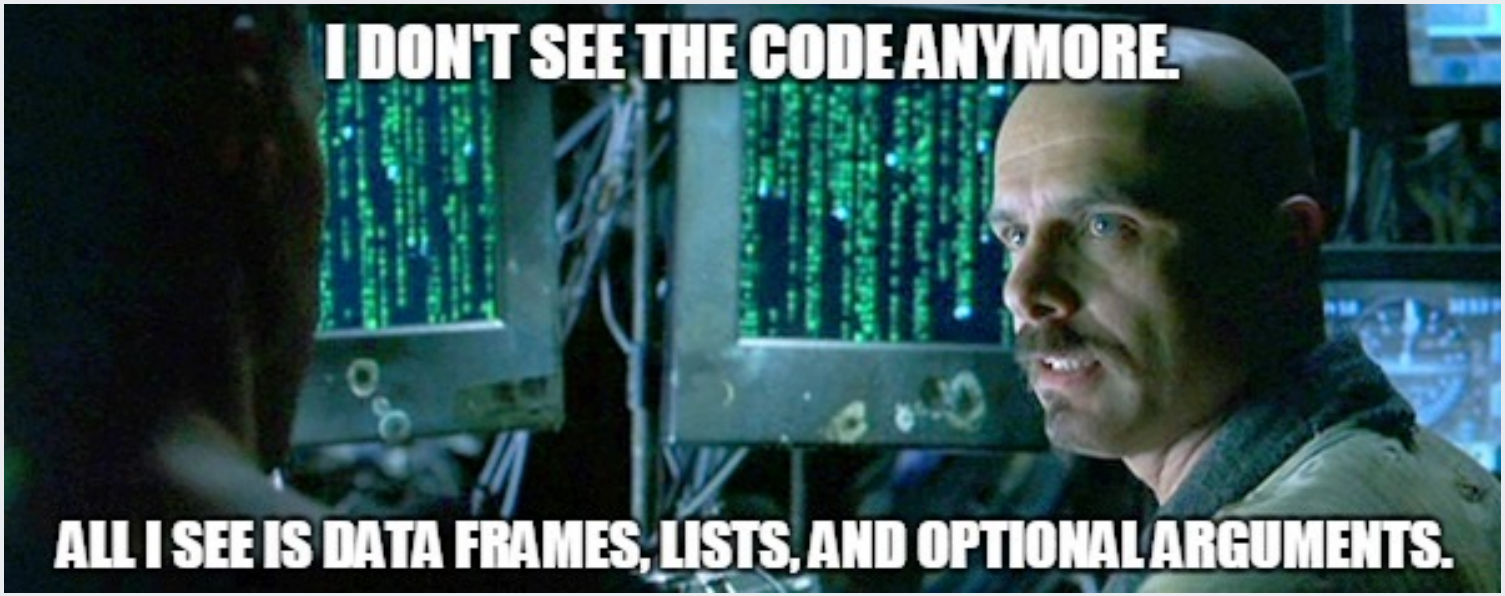
```
## [1] 34
```

Jaký je Tarantino index (tj. počet mrtvých na počet nadávek) pro Inglorious Basterds?

```
Dead_Curse = data.frame(Filmy_Period[1:2,2])
Dead_Curse[2,1]/Dead_Curse[1,1]
```

```
## [1] 1.208333
```

Intermezzo


I DON'T SEE THE CODE ANYMORE.
ALL I SEE IS DATA FRAMES, LISTS, AND OPTIONAL ARGUMENTS.

Data Frame

Vyvolání Data Frame z R

```
data()
data(USArrests)
View(USArrests)
??USArrests
```

Jak se zorientovat v Data Frame?

```
head() # show the first observations of a data frame
tail() # prints out the last observations in your data set
str() # struktura dat
```

Data Frame

Výběr z prvků

```
USArrests[1:3,3]
USArrests[1:3,"UrbanPop"]
USArrests$UrbanPop
USArrests[1, "UrbanPop"]
```

Subsoubory

```
subset(USArrests, UrbanPop < 50)
```

Seřazování

```
order(USArrests$Murder)
```