Web Data Mining com R: processamento de dados [no R]

Fabrício Jailson Barth

Faculdade BandTec e VAGAS Tecnologia

Junho de 2013

Sumário

- Projeto R
- O que são **dados**?
- Raw data versus dado tratado.
- Representação de dados no R.

Projeto R

- http://www.r-project.org/
- R Studio - http://www.rstudio.com/
- É free
- É a linguagem de programação mais popular para análise de dados
- Script é melhor que clicar e arastar:
  - ⋆ É mais fácil de comunicar.
  - ⋆ Reproduzível.
  - ⋆ É necessário pensar mais sobre o problema.
- Existe uma quantia grande de pacotes disponíveis

Definição de dados

"Data are values of qualitative or quantitative variables, belonging to a set of items."

http://en.wikipedia.org/wiki/Data

"Data are values of qualitative or quantitative variables, belonging to a **set of items**."

Set of items: conjunto de itens (objetos) de interesse.

"Data are values of qualitative or quantitative **variables**, belonging to a set of items."

variables: uma medida ou uma característica de um item.

"Data are values of **qualitative** or **quantitative** variables, belonging to a set of items."

qualitative: cidade de origem, sexo, fez ou não tratamento.

quantitative: peso, altura, pressão do sangue.

Raw data versus dados processados

Raw data

- Fonte original dos dados
- Geralmente difícil para fazer algum tipo de análise

http://en.wikipedia.org/wiki/Raw_Data

Dados processados

- Dados que estão prontos para serem analisados
- O processamento pode incluir *merging*, *subsetting*, *transforming*, etc...
- Todas as etapas devem ser registradas

http://en.wikipedia.org/wiki/Compute_data_processing

Dados brutos

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Dados brutos

consideração o projeto da aprendizagem que pensa como didaticamente os cursos devem ser projetados com o uso da tecnologia adequada. Isso inclui levar em conta os aspectos sociais e culturais envolvidos. Deixo abaixo algumas indicações de leitura que tratam isso. Assim, acho que dizer que tecnologia deve ser usada de forma responsável, não é discutir MOOCs. Outro ponto importante é destacar que os MOOCs aparecem no contexto da educação aberta e Ciencia aberta e inclui REAs, que costumavam ser chamados de objetos de aprendizem e agora discutem-se as licenças, as perspectivas de reutilização e de localização; os periódicos abertos que reagem aos altos valores de assinaturas dos periódicos tradicionais, as novas formas de publicação incluindo blogs; a educação hibrida; os ambientes pessoais de aprendizagem, etc. No geral

Exemplo de dado processado

Table 1: Exemplo de tabela com as transações dos usuários

usuário	$categoria_1$	$categoria_2$	$categoria_3$	$\cdots$	$categoria_m$
$user_1$	0	2	0	$\cdots$	1
$user_2$	1	1	0	$\cdots$	0
$user_3$	2	0	1	$\cdots$	0
$user_4$	0	1	0	$\cdots$	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$user_n$	1	1	0	$\cdots$	1

Tiny data

- Cada variável (atributo) forma uma coluna.
- Cada observação (exemplo) forma uma linha.
- Cada tabela ou arquivo armazena dados sobre uma observação (i.e., pessoas / hospitais)
- http://vita.had.co.nz/papers/tidy-data.pdf

Big or small - you need the right data

Representação de dados no R

Tipos de dados importantes no R

- Classes: Character, Numeric, Integer, Logical
- Objetos: Vector, Matrices, Data frames, List, Factors, Missing Values
- Operadores: Subsetting, Logical Subsetting

Character

```
nome = "maria"
class(nome)
```

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

```
nome
```

```
## [1] "maria"
```

Numeric

```
peso = 76.2
class(peso)
```

```
## [1] "numeric"
```

```
peso
```

```
## [1] 76.2
```

Integer

```
qtdFilhos = 1L
class(qtdFilhos)
```

```
## [1] "integer"
```

```
qtdFilhos
```

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

Logical

```
temCarro = TRUE
class(temCarro)
```

```
## [1] "logical"
```

```
temCarro
```

```
## [1] TRUE
```

Vectors

Um conjunto de valores da mesma classe.

```
pesos = c(76.2, 80.3, 90, 117.4)
pesos
```

```
## [1] 76.2 80.3 90 117.4
```

```
nomes = c("maria", "carlos", "pedro")
nomes
```

```
## [1] "maria" "carlos" "pedro"
```

Lists

Um conjunto de valores que pode ser heterogêneo.

```
pesosV = c(76.2, 80.3, 90, 117.4)
nomesV = c("maria", "carlos", "pedro", "antônio")

myList <- list(pesos = pesosV, nomes = nomesV)
myList

## $pesos
## [1]  76.2  80.3  90.0 117.4
##
## $nomes
## [1] "maria"   "carlos"  "pedro"   "antônio"
```

Lists

Um conjunto de valores que pode ser heterogêneo.

```
pesosV = c(76.2, 80.3, 90, 117.4)
nomesV = c("maria", "carlos", "pedro", "antônio")

myList <- list(pesos = pesosV, nomes = nomesV)
myList

## $pesos
## [1]  76.2  80.3  90.0 117.4
##
## $nomes
## [1] "maria"   "carlos"  "pedro"   "antônio"
```

Matrizes

Vetores com múltiplas dimensões.

```
myMatrix = matrix(c(1, 2, 3, 4), byrow = T, nrow = 2)
myMatrix
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

Data frames

Múltiplos vetores de classes diferentes, mas com o mesmo tamanho.

```
vector1 = c(188.2, 181.3, 193.4)
vector2 = c("jeff", "roger", "andrew", "brian")
myDataFrame = data.frame(heights = vector1,
                         firstNames = vector2)
```

```
## Error: arguments imply differing number of rows: 3, 4
```

```
myDataFrame
```

```
## Error: object 'myDataFrame' not found
```

Data frames

```
> vector1 = c(188.2, 181.3, 193.4)
> vector2 = c("jeff", "roger", "andrew")
> myDataFrame = data.frame(heights = vector1,
                           firstNames = vector2)
> myDataFrame

  heights firstNames
1   188.2       jeff
2   181.3      roger
3   193.4     andrew
```

Factors

Variáveis qualitativas que podem ser incluídas no modelo.

```
smoker = c("yes", "no", "yes", "yes")
smokerFactor = as.factor(smoker)
smokerFactor
```

```
## [1] yes no yes yes
## Levels: no yes
```

Missing values

No R os valores faltantes são codificados como NA

```
vector1 <- c(188.2, 181.3, 193.4, NA)
vector1
```

```
## [1] 188.2 181.3 193.4 NA
```

```
is.na(vector1)
```

```
## [1] FALSE FALSE FALSE TRUE
```

Subsetting

```
vector1 = c(188.2, 181.3, 193.4, 192.3)
vector2 = c("jeff", "roger", "andrew", "brian")
myDataFrame = data.frame(heights = vector1,
                         firstNames = vector2)
vector1[1]
```

```
## [1] 188.2
```

```
vector1[c(1, 2, 4)]
```

```
## [1] 188.2 181.3 192.3
```

Subsetting

```
myDataFrame[1, 1:2]
```

```
## heights firstNames
## 1 188.2 jeff
```

```
myDataFrame$firstNames
```

```
## [1] jeff roger andrew brian
## Levels: andrew brian jeff roger
```

Logical subsetting

```
myDataFrame[myDataFrame$firstNames == "jeff", ]
```

```
## heights firstNames
## 1 188.2 jeff
```

```
myDataFrame[heights < 190, ]
```

```
## heights firstNames
## 1 188.2 jeff
## 2 181.3 roger
## 4 192.3 brian
```

Obtendo dados

Dados locais (*toy examples*)

```
help(data)
data()
data(iris)
```

Earthquake data (*dados reais*)

```
fileUrl <-
"http://earthquake.usgs.gov/earthquakes
      /catalogs/eqs7day-M1.txt"

download.file(fileUrl,destfile="./data/earthquakeData.csv",
              method="curl")
dateDownloaded <- date()

eData <- read.csv("./data/earthquakeData.csv")
```

https://explore.data.gov/Geography-and-Environment/Worldwide-M1-Earthquakes-Past-7-Days/7tag-iwnu