Aquisição de Arquivos e Datasets

Sumarizando Dados

Delermando Branquinho Filho

Revendo os dados

```
reviews = read.csv("./data/reviews.csv"); solutions <- read.csv("./data/solutions.csv")
head(reviews,2)
```

```
##   id solution_id reviewer_id      start       stop time_left accept
## 1  1           3          27 1304095698 1304095758      1754      1
## 2  2           4          22 1304095188 1304095206      2306      1
```

```
head(solutions,2)
```

```
##   id problem_id subject_id      start       stop time_left answer
## 1  1        156         29 1304095119 1304095169      2343      B
## 2  2        269         25 1304095119 1304095183      2329      C
```

Merging data - merge()

- Merges data frames
- Important parameters: *x,y,by,by.x,by.y,all*

```
names(reviews)
```

```
## [1] "id"          "solution_id" "reviewer_id" "start"       "stop"
## [6] "time_left"   "accept"
```

```
names(solutions)
```

```
## [1] "id"         "problem_id" "subject_id" "start"      "stop"
## [6] "time_left"  "answer"
```

```
mergedData = merge(reviews,solutions,by.x="solution_id",by.y="id",all=TRUE)
head(mergedData)
```

```
##   solution_id id reviewer_id    start.x     stop.x time_left.x accept
## 1           1  4          26 1304095267 1304095423        2089      1
## 2           2  6          29 1304095471 1304095513        1999      1
## 3           3  1          27 1304095698 1304095758        1754      1
## 4           4  2          22 1304095188 1304095206        2306      1
## 5           5  3          28 1304095276 1304095320        2192      1
## 6           6 16          22 1304095303 1304095471        2041      1
##   problem_id subject_id    start.y     stop.y time_left.y answer
## 1        156         29 1304095119 1304095169        2343      B
## 2        269         25 1304095119 1304095183        2329      C
## 3         34         22 1304095127 1304095146        2366      C
## 4         19         23 1304095127 1304095150        2362      D
## 5        605         26 1304095127 1304095167        2345      A
## 6        384         27 1304095131 1304095270        2242      C
```

Padrão - mesclar todos os nomes de colunas comuns

```
intersect(names(solutions),names(reviews))
```

```
## [1] "id"        "start"     "stop"      "time_left"
```

```
mergedData2 = merge(reviews,solutions,all=TRUE)
head(mergedData2)
```

```
##   id      start       stop time_left solution_id reviewer_id accept
## 1  1 1304095119 1304095169      2343          NA          NA     NA
## 2  1 1304095698 1304095758      1754           3          27      1
## 3  2 1304095119 1304095183      2329          NA          NA     NA
## 4  2 1304095188 1304095206      2306           4          22      1
## 5  3 1304095127 1304095146      2366          NA          NA     NA
## 6  3 1304095276 1304095320      2192           5          28      1
##   problem_id subject_id answer
## 1        156         29      B
## 2         NA         NA   <NA>
## 3        269         25      C
## 4         NA         NA   <NA>
## 5         34         22      C
## 6         NA         NA   <NA>
```

Usando join no pacote plyr

Mais rápido, mas menos completo - padrão para esquerda join, veja o arquivo de ajuda para more

```
library("plyr")
```

```
## Warning: package 'plyr' was built under R version 3.3.3
```

```
df1 = data.frame(id=sample(1:10),x=rnorm(10))
df2 = data.frame(id=sample(1:10),y=rnorm(10))
arrange(join(df1,df2),id)
```

```
## Joining by: id
```

```
##    id          x           y
## 1   1 -1.0296155  1.04798503
## 2   2  0.7612982  0.08426767
## 3   3  2.5676680  0.86801186
## 4   4  0.2861006  0.78669380
## 5   5  1.7676736 -0.03848672
## 6   6  0.7452961  1.60619796
## 7   7 -0.3926016  1.01391093
## 8   8 -0.3113189 -1.21680056
## 9   9  1.5600582  1.29007666
## 10 10 -0.3771526  0.89730695
```

Se você tiver vários dataframes

```
df1 = data.frame(id=sample(1:10),x=rnorm(10))
df2 = data.frame(id=sample(1:10),y=rnorm(10))
df3 = data.frame(id=sample(1:10),z=rnorm(10))
dfList = list(df1,df2,df3)
join_all(dfList)
```

```
## Joining by: id
## Joining by: id
```

```
##    id          x          y           z
## 1   9 -1.1167648  1.0917366 -0.05292708
## 2   7 -1.5267326 -0.0803516 -0.41229930
## 3   6 -1.5643775  0.1365449 -0.33242135
## 4  10 -0.4719342 -1.0570118  0.59419997
## 5   3  1.0447728  0.5653956  0.84288006
## 6   8 -0.5326668  1.1719425 -2.25697500
## 7   1 -0.4990145  0.9795686 -0.66012351
## 8   4 -0.2268142 -0.2634191 -1.61723037
## 9   5  0.7980957 -0.6917604 -0.40247321
## 10  2 -0.2853184 -2.7923189  0.06364175
```