Desafio 13

Rafael Gomes Carneiro - RA185462

Laboratório 3: SQLite com Polars

Author: Benilton S Carvalho

Objetivo

Dados relacionais são uma constante no exercício da profissão do estatístico. Esta estratégia permite uma representação mais efetiva de dados estruturados, oferecendo a possibilidade de análises computacionalmente mais eficientes. Neste laboratório, trabalharemos com dados relacionais a partir de uma base de dados SQLite.

Os Dados

O banco de dados para esta atividade é o "IMDb Movie Data", que possui informações sobre filmes, atores, diretores, gêneros e outros. Para os arquivos indicados abaixo, considere que o símbolo `\N` representa valores faltantes. Observe com cuidado a extensão dos arquivos para uma indicação do formato do mesmo.

1. `title.basics0.tsv.gz` (Informações Básicas dos Filmes)

- Link: https://drive.google.com/file/d/1iYqAGTtIhRLK4ycFK41hYaWK3iKfOdT2/view?usp=sharing

Coluna	Descrição
`tconst`	Identificador único do título (ex: tt1234567)
`titleType`	Tipo do título (ex: movie, short, tvSeries)
`primaryTitle`	Título principal
`originalTitle`	Título original

Coluna	Descrição
`isAdult`	Indica se é conteúdo adulto (0: não, 1: sim)
`startYear`	Ano de lançamento/início
`endYear`	Ano de término (para séries)
`runtimeMinutes`	Duração em minutos
`genres`	Gêneros separados por vírgula (ex: Action, Comedy)

2. `title.ratings.tsv.gz` (Avaliações dos Filmes)

- Link: https://drive.google.com/file/d/1kZuj1lnXkPrNURzwYuc4FXTM7Pvknr3o/view?usp=sharing

Coluna	Descrição
`tconst`	Identificador único do título (ex: tt1234567)
`averageRating`	Nota média dos usuários (escala de 1 a 10)
`numVotes`	Número de votos recebidos

3. `title.principals0.tsv.gz` (Elenco e Equipe Técnica)

- Link: https://drive.google.com/file/d/1oLR2_mFyRHEiKlqFKt4QDVYlGH8LC4mn/view?usp=sharing

Coluna	Descrição
`tconst`	Identificador único do título (ex: tt1234567)
`nconst`	Identificador único da pessoa (ex: nm1234567)
`category`	Categoria de trabalho da pessoa (ex: actor, director)
`job`	Função específica desempenhada (para não-atores)
`characters`	Personagens interpretados (para atores)

4. `movies.sqlite3` (Banco de Dados em SQLite)

- Link: https://drive.google.com/file/d/1l-rnMw2bsmbZ-e5h9SGydqGmv9fNAtWs/view?usp=sharing

Atividade

1. Crie um banco de dados SQLite utilizando os 3 arquivos acima. O banco de dados deve conter as seguintes tabelas: `basics`, `ratings` e `principals`
2. (**Utilizando SQL**, responda): Quais são os 5 filmes com as maiores notas (`averageRating`)? Apresente uma solução capaz de desempatar os filmes baseando-se no número de votos recebidos.
3. (**Utilizando SQL**, responda): Qual é o gênero mais frequente entre os filmes com nota maior que 8?
4. (**Utilizando SQL**, responda): Quais são os 3 atores/atrizes que mais participaram de filmes com nota maior que 7.5?

```
library(readr)
library(DBI)
library(RSQLite)
library(dplyr)
```

```
Anexando pacote: 'dplyr'
```

```
Os seguintes objetos são mascarados por 'package:stats':

    filter, lag
```

```
Os seguintes objetos são mascarados por 'package:base':

    intersect, setdiff, setequal, union
```

```
library(tictoc)
```

```
library(readr)

# Callback para o arquivo basics
processa_basics <- function(x, pos) {
  x <- subset(x, titleType == "movie")
  return(nrow(x))
}

basics <- read_tsv_chunked(
```

```
  file = "../dados/title.basics0.tsv",
  callback = SideEffectChunkCallback$new(processa_basics),
  chunk_size = 100000,
  na = "\\N",
  show_col_types = FALSE
)

# Callback para ratings
processa_ratings <- function(x, pos) {
  # apenas contar linhas do chunk, sem filtro
  return(nrow(x))
}

ratings <- read_tsv_chunked(
  file = "../dados/title.ratings.tsv",
  callback = SideEffectChunkCallback$new(processa_ratings),
  chunk_size = 100000,
  na = "\\N",
  show_col_types = FALSE
)

# Callback para principals
processa_principals <- function(x, pos) {
  # apenas contar linhas do chunk, sem filtro
  return(nrow(x))
}

principals <- read_tsv_chunked(
  file = "../dados/title.principals0.tsv",
  callback = SideEffectChunkCallback$new(processa_principals),
  chunk_size = 100000,
  na = "\\N",
  show_col_types = FALSE
)
```

```
dim(basics)
```

```
NULL
```

```
dim(ratings)
```

```
NULL
```

```r
dim(principals)
```

```
NULL
```