Análise e Sistematização de Big Data

Leonardo Sangali Barone

April 20, 2017

Análise e Sistematização de Big Data

O objetivo do curso, tal como construímos ao longo dos 4 dias, foi fazer um breve introdução ao R com foco e manipulação de dados e apresentar diversos tópicos relacionados ao termo pouco preciso, porém popular, “Big Data”.

Big Data é um termo impreciso

Umas das grandes dificuldades do curso é lidar com o fato de que há muita coisa sob o guarda-chuva de Big Data. Quando falamos de Big Data, estamos falando alternadamente de:

- Ciência de Dados (roupagem nova para estatística)
- Armazenamento de dados (que é um dos problemas da computação desde o princípio)
- Captura de dados (que é um tema novo)
- Aprendizado de Máquina (que é um uso novo para ferramentas estatísticas velhas)

O plano inicial do curso

Se tivéssemo que mudar o nome do curso, o mais adequado seria "R para análise de dados e Big Data". O plano inicial do curso, antes de conhecer vocês, era trabalhar:

- Uma introdução ao R e à gramática *dplyr*
- Webscrapping
- Aprendizado de máquina

O curso como termina

Conhecendo vocês e ampliando o material do curso, terminamos com os seguintes tópicos

- Uma introdução ao R e à gramática *dplyr* (ampliado)
- Webscrapping
- Dados volumosos e integração R e SGBD
- RevoScaleR e Microsoft R Client
- Aprendizado de máquina
- Aprendizado de máquina com Spark

Todos os tópicos, exceto o primeiro, são parte do que poderíamos chamar de R para Big Data

O prometido não entregue

Faltou entregarmos um tópico menos importante (tanto como solução quanto para vocês):

- Bibliotecas básicas do R para dados volumosos

E faltou exercitarmos aprendizado de máquina com dados do Cadunico.

O que mais poderia haver no curso

- Integração R e Hadoop

Ainda assim, o tópico R e Spark proporciona ferramentas para trabalhar com dados em uma infra-estrutura Hadoop.

Tópico a tópico

Uma introdução ao R e à gramática *dplyr*

Idealmente, teríamos um curso de introdução ao R antes de um curso R para Big Data. Este tópico teve a função de apresentar os fundamentos da linguagem R de forma rápida e contemplando seu desenvolvimento recente pela RStudio (pacote *tidyverse*). Para continar aprendendo, recomendo a leitura de:

- Aquino (2014) R para cientistas sociais,
- Wickham, Hadley, and Garrett Grolemund. 2016. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. "O'Reilly Media, Inc.".

Webscrapping

A captura de dados não estruturados e em volume é um tema de "Big Data". Os tutoriais do curso, que compõem o curso original oferecido pelo MQ-UFMG, mostram os passos iniciais para webscrapping.

Dados volumosos e integração R e SGBD

Em nossa interação, aprensetamos soluções para dados volumosos a integração entre R e outros SGBD. Alguns dos sistemas têm suporte para *dplyr*, como MySQL e PostgreSQL. Outros, porém, como Teradata, exigem conexão ODBC e uso de queries em SQL.

RevoScaleR e Microsoft R Client

No contexto de trabalho do MDS, no qual os dados estão em processo de migração para Teradata, faz bastante sentido utiliza uma ferramenta que integra de maneira eficiente R e Teradata: o Microsoft R Client, para o qual estão disponíveis os métodos e funções do RevoScaleR.

Mesmo sem a necessidade de integração com Teradata, RevoScaleR também é uma solução para trabalhar com dados que extrapolam a RAM e para aprendizado de Máquina.

Aprendizado de Máquina

"Big Data" também é sobre a análise de dados. No curso aprensetamos uma breve introdução ao aprendizado de máquina, tema que, por si só, poderia se objeto de mais algumas dezenas de horas de curso.

Aprendizado de Máquina com Spark e R

Apache Spark é um framework de código aberto para computação e processamento de dados em paralelo e é uma das soluções mais atraentes para aprendizado de máquina com velocidade e para um grande volume de dados. Com o pacote *sparklyr*, da RStudio, conectar R e Spark é bastante simples, pois podemos utiliar os verbos do *dplyr* e as funções básicas de aprendizado de máquina estão implementadas.

Futuro

- Seguir com o aprendizado de R
- Seguir com um dos tópicos
- Explorar o Microsoft R Cliente e integraçao com Teradata
- Explorar o Apache Spark com R