

Ementa

Análise de Big Data usando R

Felipe Nunes*

Fernando Meireles†

1/2017

Segunda versão

Expressões como *big data*, *machine learning* e *data science* tornaram-se relativamente comuns a partir dos anos 2000. Empresas já realizam diariamente pesquisas de larga escala para acompanhar comportamentos individuais e coletivos em redes sociais e *websites*. Na academia também, investigações com dados massivos – arquivos digitais, censos, API, dados abertos – são conduzidas para examinar se variáveis como renda, escolaridade, discursos e votos afetam coisas como expectativa de vida, mobilidade social, popularidade e fragmentação partidária. Nosso objetivo neste curso é fornecer os conceitos e as ferramentas básicas para realizar análises como estas.

O curso está organizado em dois módulos. No primeiro, oferecemos uma introdução a uma das ferramentas mais versáteis e potentes para a análise dados: o ambiente de programação R. Ao longo dele, aprenderemos a aplicá-lo para realizar as tarefas mais comuns numa pesquisa, tais como: coletar, ler, limpar, manipular e documentar os mais diversos tipos de dados. Complementarmente, também cobriremos as noções básicas de *data science* e *big data* de forma aplicada às Ciências Sociais.

Após esta primeira etapa, veremos mais detidamente como o R pode ser usado para contar histórias convincentes com base sólida nos dados de forma rápida, eficiente e reproduzível. Neste segundo módulo, também realizaremos atividades práticas que cobrirão desde o cálculo de estatísticas descritivas básicas até a exportação dos resultados de forma automatizada para *Word*, *L^AT_EX*, *.pdf*, *HTML* ou qualquer outro meio.

Deve ficar claro, contudo, que não esgotaremos as possibilidades do mundo da análise de dados, ou *data science*, com o R (na verdade, seria irrealista prometer isto). Procuraremos apenas mostrar como empregá-lo para reduzir tarefas bastante complexas de análise, acadêmicas ou não, a outras essencialmente simples. Ao fim da disciplina, é esperado que os(as) alunos(as) estejam preparados para realizar suas próprias análises

*E-mail: <felipenunes@ufmg.br>.

†E-mail: <fmeireles@ufmg.br>.

de dados, produzindo estatísticas descritivas simples, visualizações eficientes e inferências replicáveis.

Materiais

A princípio, realizaremos o curso num laboratório de informática. De qualquer modo, sendo possível, é recomendado que os(as) alunos(as) tragam *notebooks*; isto facilita o aprendizado e o uso posterior do R (já que nos permite auxiliar presencialmente a contornar quaisquer erros causados por incompatibilidade de *hardware* ou *software*). No último tópico da ementa, indicamos os procedimentos necessários para ter a última versão funcional do R devidamente instalada.

Todos os *scripts* e leituras usados no curso serão disponibilizados antes da realização das atividades. A quem interessar, uma boa fonte introdutória é o curso *online* do *Datacamp*. Basicamente, ele simula um *console* do R e ensina o básico de forma prática, com dicas e ajudas eventuais, um corretor que notifica erros e um sistema de avaliação progressivo. O curso é gratuito (mas em inglês) e pode ser acessado em:

- www.datacamp.com

Por fim, a melhor fonte em português para solucionar dúvidas pontuais é o *Stackoverflow*, um fórum de programação dedicado ao R (mas não só a ele). Ali, não só é possível encontrar respostas a diversas questões já feitas como também postar outras novas.

- pt.stackoverflow.com

Logística do curso

As aulas serão majoritariamente práticas, precedidas ou sucedidas de breves explicações acompanhadas de *slides* (que eventualmente ficarão à disposição após o término de cada aula). Para acompanhar devidamente a disciplina, é esperado que leituras e exercícios para cada aula sejam realizadas previamente. Como a parte principal da disciplina será orientada, auxílio e esclarecimento de dúvidas serão feitos principalmente em aula (outras questões podem ser discutidas antes ou depois destas, ou por e-mail). Também sugerimos que os(as) alunos(as) trabalhem em duplas ou trios tanto para maximizar o engajamento nas atividades quanto para promover trabalho conjunto na solução das atividades propostas.

Datas e horários das aulas: sextas-feiras, rigorosamente às 14h.

Avaliação

Realizaremos várias tarefas durante o curso, mas a maior parte delas não serão avaliadas. Como a disciplina é centralmente prática, avaliaremos principalmente a participação e o compromisso com a realização das tarefas dadas em aula. De qualquer forma, teremos duas avaliações principais. A primeira será uma atividade em aula para aferir o conhecimento dos alunos sobre as leituras cobradas e o básico do R. Já a avaliação consistirá na elaboração de um projeto, que deverá ser apresentado no formato de relatório ou artigo, a critério. Para tanto, deverá ser entregue um *script* documentado e reprodutível com uma análise original. Cada código será avaliado individualmente, e as notas serão dadas pelo nível de desenvolvimento das análises - a ideia principal aqui, portanto, é incentivar que cada um procure suas próprias soluções, e não receitas prontas.

O peso de cada item na nota final será o seguinte:

- Engajamento, frequência e realização de atividades em aula - 20%
- Avaliação ao final do Módulo 1 - 30%
- Projeto final - 50%

Código de conduta

Embora a maioria das tarefas demandadas sejam relativamente simples, não serão admitidos códigos ou respostas copiadas de colegas ou da internet nas avaliações. Em outras palavras, plágios acarretarão em perda de nota.

Atendimento a necessidades especiais

Alunos(as) com quaisquer necessidades ou solicitações individuais podem nos procurar diretamente, por e-mail ou pessoalmente, para obterem auxílio. Todos os pedidos serão mantidos em sigilo.

Política de gênero

Cursos de programação e de metodologia frequentemente são o reino dos neandertais: homens são maioria, monopolizam a participação e os computadores. Por conta disso, seguiremos um protocolo muito simples nesta disciplina: alunas têm preferência sobre computadores e participação – o que implica em não interromper colegas, não centralizar as atividades em grupo e priorizar o uso de *mouse* e teclado às mulheres sempre que faltarem computadores.

Plano dos módulos

Módulo 1 Manipulação de dados

1.1 Introdução ao R; 1.2 Introdução ao *big data*; 1.3 Básico: objetos, classes e vetores; 1.4 Básico II: manipulando objetos e vetorização; 1.5 Funções; 1.6 Chamadas, argumentos e ajuda; 1.7 Manipulando `data.frames`; 1.8 Lendo dados de arquivos e da internet (*webscraping*).

Módulo 2 Análise de Dados

2.1 Manipulando `data.frames`; 2.2 Estatísticas descritivas; 2.3 Visualização de dados avançada; 2.4 Modelos lineares e generalizados; 2.5 Reprodutibilidade e documentação de pesquisas; 2.6 Auxílio no projeto final.

Total de aulas: 15.

Cronograma e leituras

Módulo 1

Aula 1 – Apresentação da disciplina. Data: 17/3.

Aula 2 – Introdução ao R e ao Big Data. Data: 24/3.

Objetivo: Oferecer uma introdução ao conceito de Big Data e apresentar o ambiente de programação R.

Jakson Alves de Aquino. *R para cientistas sociais*. EDITUS - Editora da UESC, 2014, cap. 1 e 2.

C D Shikida; Rodrigo Fernandez. *Notas introdutórias em econometria aplicada usando R/Rstudio*. 2016. URL <http://wp.ufpel.edu.br/cdshikida/apostilarstudio/>, cap. 1 e 2.

Poder da linguagem R fascina analistas (*website*).

O que é Data Science? (*website*).

Aula 3 – Introdução ao R II: objetos, classes e vetores. Data: 31/3.

Objetivo: Discutir as potencialidades do R e aprender a criar e manipular vetores de diferentes tipos.

Jakson Alves de Aquino. *R para cientistas sociais*. EDITUS - Editora da UESC, 2014, cap. 3.

Jakson Alves de Aquino. Software livre e desenvolvimento de trabalhos científicos: o r como exemplo a ser seguido. *Revista Política Hoje-ISSN: 0104-7094*, 24(2):75–86, 2015.

Aula 4 – Controle de fluxo, vetorização e funções. Data: 7/4.

Objetivo: Controlar e automatizar a repetição de tarefas no R.

Jakson Alves de Aquino. *R para cientistas sociais*. EDITUS - Editora da UESC, 2014, cap. 3.

Fernando Meireles; Denisson Silva; Beatriz Costa. electionsbr: R functions to download and clean brazilian electoral data. 2016. URL <http://fmeireles.com/files/electionsbr.pdf>

Victor Lemes Landeiro. Introdução ao uso do programa r, págs. 4-5.

Controles de fluxo: *loops* e *apply* (*website*).

Control Flow (*website*).

Aula 5 – Introdução a `data.frames`. Data: 14/4.

Objetivo: Criar e manipular `data.frames`.

Jakson Alves de Aquino. *R para cientistas sociais*. EDITUS - Editora da UESC, 2014, cap. 4 e 5.

Fernando Meireles. rsciolo - an r package to scrape meta-data from scientific articles hosted on scielo. 2016. URL <https://github.com/meirelesff/rSciolo>

Aula 6 – Carregando e trabalhando com dados. Data: 21/4.

Objetivo: Carregar e manipular os mais diversos tipos de dados no R (com `dplyr`).

Jakson Alves de Aquino. *R para cientistas sociais*. EDITUS - Editora da UESC, 2014, cap. 4 e 5.

Como importar qualquer arquivo no R (*website*).

Como importar dados em `.csv` no R (*website*).

Aula 7 – Webscraping e webcrawling (`rvest`). Data: 28/4.

Objetivo: Coletar dados da *web* a partir de páginas estáticas.

Denisson Silva and Fernando Meireles. Ciência política na era do big data: automação na coleta de dados digitais. *Revista Política Hoje*, 24(2):87–102, 2016.

Fernando Meireles. rsciolo - an r package to scrape meta-data from scientific articles hosted on scielo. 2016. URL <https://github.com/meirelesff/rSciolo>

Aula 8 – Webscraping e webcrawling (`selenium`). Data: 5/6.

Objetivo: Coletar dados da *web* de sites dinâmicos.

Denisson Silva and Fernando Meireles. Ciência política na era do big data: automação na coleta de dados digitais. *Revista Política Hoje*, 24(2):87–102, 2016.

Aula 9 – Avaliação (data provisória). Data: 12/6.

Módulo 1

Aula 10 – Introdução à análise de dados. Data: definir.

Aula 11 – Visualização de dados. Data: definir.

Aula 12 – Estatísticas e modelos. Data: definir.

Aula 13 – Exportação e apresentação de resultados. Data: definir.

Aula 14 – Orientação para o projeto final. Data: definir.

Aula 15 – Orientação para o projeto final. Data: definir.

Para instalar o R

Para instalar o R, basta ir ao site do CRAN (*Comprehensive R Archive Network*), que é a rede de fundadores e administradores do *core* da linguagem R, e baixar o *setup* indicado para o seu sistema operacional:

- cran.r-project.org

Feito isto, já é possível usar o R – mas só via *console*, o que não é tão fácil/útil. É por isso que usaremos uma IDE (i.e. Ambiente de Desenvolvimento Integrado) neste curso: especificamente, usaremos o *RStudio*. Para baixá-lo, basta entrar no seguinte site e escolher a opção mais adequada para o seu sistema operacional:

- www.rstudio.com

Referências

Jakson Alves de Aquino. *R para cientistas sociais*. EDITUS - Editora da UESC, 2014.

Fernando Meireles; Denisson Silva; Beatriz Costa. electionsbr: R functions to download and clean brazilian electoral data. 2016. URL <http://fmeireles.com/files/electionsbr.pdf>.

Jakson Alves de Aquino. Software livre e desenvolvimento de trabalhos científicos: o r como exemplo a ser seguido. *Revista Política Hoje-ISSN: 0104-7094*, 24(2):75–86, 2015.

C D Shikida; Rodrigo Fernandez. *Notas introdutórias em econometria aplicada usando R/Rstudio*. 2016. URL <http://wp.ufpel.edu.br/cdshikida/apostilarstudio/>.

Victor Lemes Landeiro. Introdução ao uso do programa r.

Fernando Meireles. rscielo - an r package to scrape meta-data from scientific articles hosted on scielo. 2016. URL <https://github.com/meirelesff/rScielo>.

Denisson Silva and Fernando Meireles. Ciência política na era do big data: automação na coleta de dados digitais. *Revista Política Hoje*, 24(2):87–102, 2016.