

Text as Data para Ciências Sociais
guiia prático com aplicações

Davi Moreira

2019-10-04

Contents

Prefácio	5
Objetivo	5
Sobre o autor/organizador	5
Licença	6
Agradecimentos	6
1 Introdução	7
1.1 O R e o RStudio	7
1.2 O Pacote <code>txt4cs</code> e outros	9
1.3 Material de apoio	10
2 Text as data: o texto como dado	11
2.1 Panorama da área	11
2.2 Oportunidades	11
2.3 Quadro geral de metodologias	15
2.4 O processo de análise do texto como dado	16
3 R e o Processamento de Linguagem Natural	19
3.1 Encoding - Codificação de caracteres	19
3.2 Encoding para remover acentos	20
4 Strings no R	21
4.1 Strings	21
4.2 O pacote <code>stringr</code>	21
4.3 Regular Expressions no R	21
5 Obtenção de conteúdo	23
5.1 Webscraping	23
5.2 Arquivos .pdf	23
5.3 Twitter	23
5.4 Áudio Transcrição	23
5.5 Imagens	23
6 Processamento dos dados	25

6.1	Tokens	25
6.2	Corpus	25
6.3	Tokens e Corpus	25
6.4	DFM: Matriz de documentos e termos	25
6.5	Stemming	25
6.6	FCM: Matriz de co-ocorrência de termos	25
7	Mineração e estatísticas básicas	27
7.1	Análise de frequência	27
7.2	Nuvem de palavras	27
7.3	tf-idf	27
7.4	Rede de n-grams	27
7.5	Correlação pareada	27
7.6	Diversidade lexical	27
7.7	Similaridade entre documentos/termos	27
7.8	KEYNESS: Análise de Frequência Relativa	27
8	Escalonamento	29
8.1	Wordscore	29
8.2	Wordfish	29
9	Classificação	31
9.1	Método de dicionário: Análise de sentimento	31
9.2	Naive Bayes	31
9.3	LDA: Latent Dirichlet Allocation	31
9.4	STM: Structed Topic Model	31

Prefácio

txt4cs|

A partir da produção de material para o curso *Text as Data: análise automatizada de conteúdo* que ministrei no [MQ-UFMG](#) em 2019 e no artigo que publiquei em coautoria com [Maurício Izumi](#) (Izumi and Moreira, 2018), esse livro tem como propósito difundir nas ciências sociais e humanidades técnicas e métodos de análise automatizada de conteúdo usando a linguagem R.

Objetivo

O principal objetivo do livro é ser tutorial prático de uso e aplicação de técnicas e métodos de análise automatizada de conteúdo na língua portuguesa através da linguagem R .

Sobre o autor/organizador

Davi Moreira é Professor Visitante do departamento de Ciência Política da Universidade Federal de Pernambuco (UFPE). Ph.D. em Ciência Política pela Universidade de São Paulo (USP) e vencedor do Prêmio CAPES de tese 2017 na área de Ciência Política e Relações Internacionais. Atuo nas seguintes áreas: políticas públicas, estudos legislativos, métodos quantitativos em ciências sociais e análise automatizada de conteúdo.

Para mais informações:

- [Página pessoal.](#)
- [Currículo Lattes.](#)
- [Google Scholar.](#)

Licença

Este livro é distribuído de acordo com a licença Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License ([CC BY-NC-SA 4.0](#)).

Agradecimentos

Agradeço aos organizadores do MQ-UFGM 2019 pela oportunidade de ministrar o curso e assim me estimular a empreender esse projeto. Também agradeço aos amigos Manoel Galdino, Rafael Magalhães, Lincon Ribeiro e Umberto Mignozetti pelo apoio e incentivo ao longo de toda minha trajetória como cientista.

Este livro é escrito com o uso do pacote **bookdown** ([Xie, 2019](#)), através do R Markdown e **knitr** ([Xie, 2015](#)).

Chapter 1

Introdução

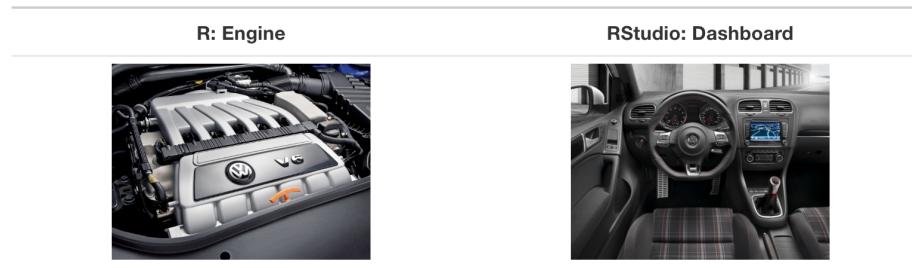
1.1 O R e o RStudio

Com o objetivo de ser um tutorial prático de uso e aplicação de técnicas e métodos de análise automatizada de conteúdo para ciências sociais e humanidades este livro fará uso da linguagem R.

R é uma linguagem de programação e também um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos. Ele pode ser facilmente instalado através do link: <https://cran.r-project.org/>.

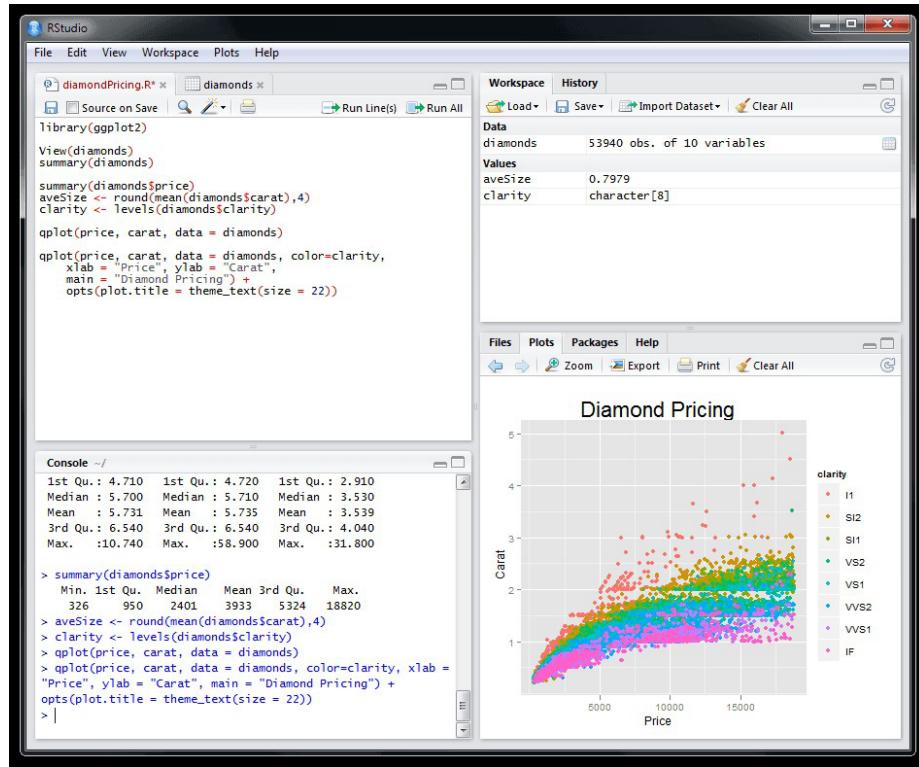
Para auxiliar no desenvolvimento das análises, este livro incentiva o uso do **RStudio**. Trata-se de um software livre de ambiente de desenvolvimento integrado (IDE) para o R¹.

De forma ilustrativa, o R e o RStudio operam como a figura abaixo:



Com o RStudio, você estará diante do seguinte dashboard:

¹IDE, do inglês *Integrated Development Environment*, é um programa de computador que reúne características e ferramentas de apoio ao desenvolvimento de software com o objetivo de agilizar este processo.



Se está começando a usar o R para análise de dados, recomendo o seguinte material:

1. [R for Data Science](#);
2. [Modern Dive - Statistical Inference via Data Science](#);
3. [Curso R](#);
4. [Usando R: Um Guia para Cientistas Políticos](#);

Em caso de dúvidas, use e abuse de fóruns como o [Stackoverflow](#). Para aprimorar seu código e otimizar o desenvolvimento de suas análises, os guias de estilo do [Google](#) e do [RStudio](#) são ótimas referências.



Figure 1.1: Fonte: Empresa Brasil de Comunicação - EBC

1.2 O Pacote `txt4cs` e outros

`txt4cs|`

Este livro conta com o pacote `txt4cs`. Ele traz consigo funções específicas e bases de dados utilizadas nos exemplos apresentados. Um dos acervos de exemplo se refere ao conteúdo proferido em 17 de abril de 2016, dia de aprovação do impeachment da então Presidenta Dilma Rousseff na Câmara dos Deputados.

Para instalação, use os comandos abaixo:

```
if(require(devtools) == F) install.packages('devtools'); require(devtools);
devtools::install_github("davi-moreira/txt4cs-pkg")
require(txt4cs)
```

Ademais, os seguintes pacotes são essenciais para o desenvolvimento da análise

automatizada de conteúdo com o R. Conforme forem necessários, serão apresentados no livro.

```
install.packages("tidyverse")
install.packages("stringr")
install.packages("quanteda")
install.packages("readtext")
install.packages("stringi")
install.packages("tm")
```

1.3 Material de apoio

Este livro não é feito do zero e resulta de inspiração em diferentes fontes. As principais são:

1.3.1 Referências para processamento de sequências de caracteres com o R

1. [Handling and Processing Strings in R](#) e [Handling Strings with R](#)
2. [R Wikibook: Programming and Text Processing](#)
3. [stringr: modern, consistent string processing](#)

1.3.2 Referências em análise de conteúdo com o R:

1. [Quanteta Tutorials](#)
2. [Text Mining with R](#)

Chapter 2

Text as data: o texto como dado

2.1 Panorama da área

A análise de conteúdo possui grande relevância para as ciências sociais. Contudo, sua abordagem manual sempre limitou o volume de documentos sob análise. São raros os projetos como o *Manifesto Research Group* que, desde os anos 1970, analisa a ênfase temática de manifestos partidários ou o *Comparative Agendas Project*, que coleta e analisa dados sobre agendas de políticas públicas em diferentes países.

O avanço tecnológico e científico permitiu que técnicas automatizadas de análise do conteúdo fossem desenvolvidas e aplicadas de forma simples a grandes acervos. Este avanço não foi realizado sem a contribuição das ciências sociais. Só a *Political Analysis*, principal revista de métodos em ciência política, possui dois *special issues* dedicados ao tema (*Special Issue*, *Virtual Issue*).

2.2 Oportunidades

Com o desenvolvimento de métodos para análise automatizada de conteúdo, hoje o leque de oportunidades as ciências sociais é diverso e promissor. Agora, é possível:

- **Analizar grandes acervos** de forma ágil e barata, otimizando o trabalho do pesquisador.
- **Pesquisar novos acervos** para inferir o conteúdo presente e assim guiar pesquisas através de atalhos informacionais.



Figure 2.1: Biblioteca Florestan Fernandes - FFLCH - USP

Figure 2.2: Acervo da CIA: <<https://www.cia.gov/library/readingroom/advanced-search-view>>

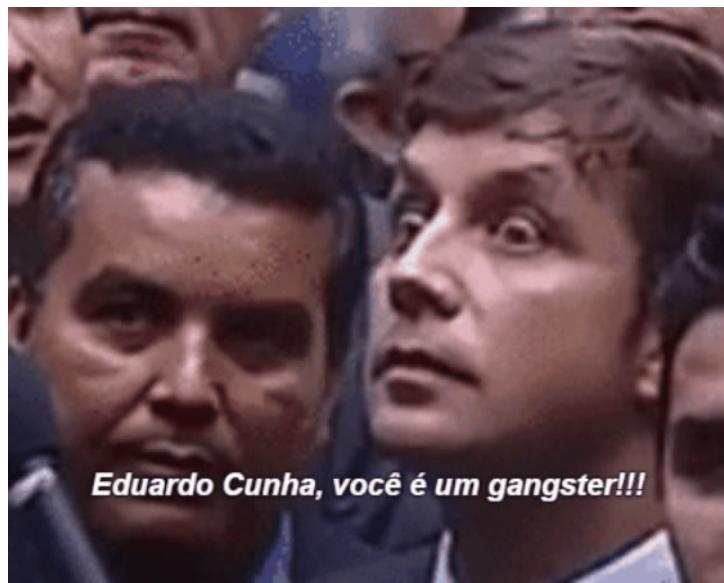


Figure 2.3: Trecho de fala do Deputado Federal Glauber Braga (PSOL-RJ) durante seu voto no processo de impeachment da então Presidenta da República Dilma Rousseff em 2016.

- Analisar processos políticos contemporâneos.
- Redes sociais.
- Fake news!

Jair M. Bolsonaro 
@jairbolsonaro

O que é ciência política?

09:26 · 06/03/2019 · Twitter for iPhone

- **Olhar o passado com as lentes do presente.** Questões que antes não podiam ser enunciadas agora podem ser respondidas! Processos políticos conhecidos podem ganhar novas interpretações através do uso de métodos e técnicas contemporâneas de análise automatizada de conteúdo.



Figure 2.4: Foto de Pedro Ladeira, Folha de São Paulo, maio de 2019.



Figure 2.5: Liberdade Guiando o Povo - Eugène Delacroix - 1830

- **Contribuir socialmente: Retórica Parlamentar** - Projeto experimental desenvolvido no primeiro Hackathon da Câmara dos Deputados em 2013 por Davi Moreira, Manoel Galdino e Luis Carli. Posteriormente incubado pelo Laboratório Hacker da Câmara dos Deputados.



2.3 Quadro geral de metodologias

Dada a complexidade da linguagem, o processo de geração, produção e seleção de dados que resultam na comunicação humana é ainda um mistério para a ciência (Izumi and Moreira, 2018; Grimmer and Stewart, 2013). Logo, modelos estatísticos desenvolvidos falham na tarefa de prover um relato preciso do processo de geração de dados utilizados na produção de conteúdo e, principalmente, em seu significado.

Os modelos de análise de conteúdo, portanto, não devem ser avaliados pelo quanto explicam do processo de geração dos dados. Transformar palavras em números não substitui a leitura cuidadosa e atenta de documentos. Reconhecendo que “métodos de análise automatizada de conteúdo são modelos incorretos de linguagem” (Grimmer and Stewart, 2013, p. 2), a performance de qualquer método automatizado não é garantida sem a consideração de ao menos quatro princípios:

1. Todos os modelos quantitativos de análise de conteúdo estão errados, mas alguns são úteis;
2. Métodos quantitativos de análise de conteúdo amplificam a capacidade humana, mas não a substitui;

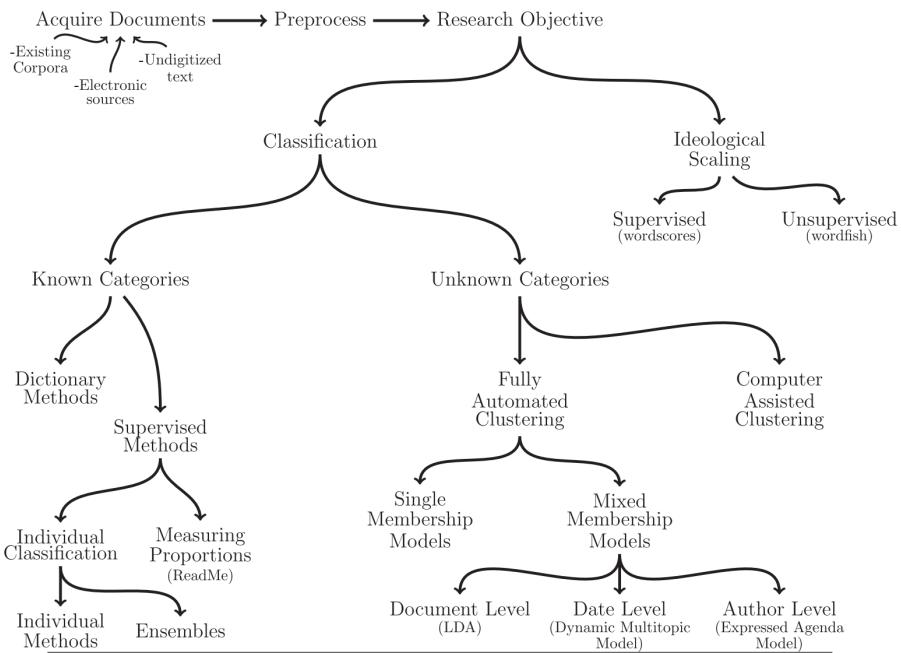


Figure 2.6: Quadro geral de metodologias para análise automatizada de conteúdo (Grimmer e Stewart, 2013)

3. Não há um método global para a análise automatizada de conteúdo;
4. Validar, validar, validar.

A escolha do modelo, da família de modelos ou de eventuais combinações a serem utilizadas é resultado dos objetivos almejados. Há uma variedade de modelos disponíveis e nenhum deles se sobrepõe aos demais.

Além de estatísticas e outras informações que podem ser obtidas através da mineração do texto enquanto dados, nesse livro será dado foco aos métodods de escalonamento e classificação de conteúdo. Assim, como indicado pelo quadro de Grimmer e Stewart (2013) métodos de análise supervisionada e não supervisionada serão abordados.

2.4 O processo de análise do texto como dado

O processo de trabalho para análise quantitativa de texto é muito similar a qualquer tipo de fluxo de trabalho para análise de dados em geral. Como indicado no livro [Text Mining with R: a tidy approach](#) (Silge and Robinson, 2017), o seguinte fluxograma será adotado nesse livro:

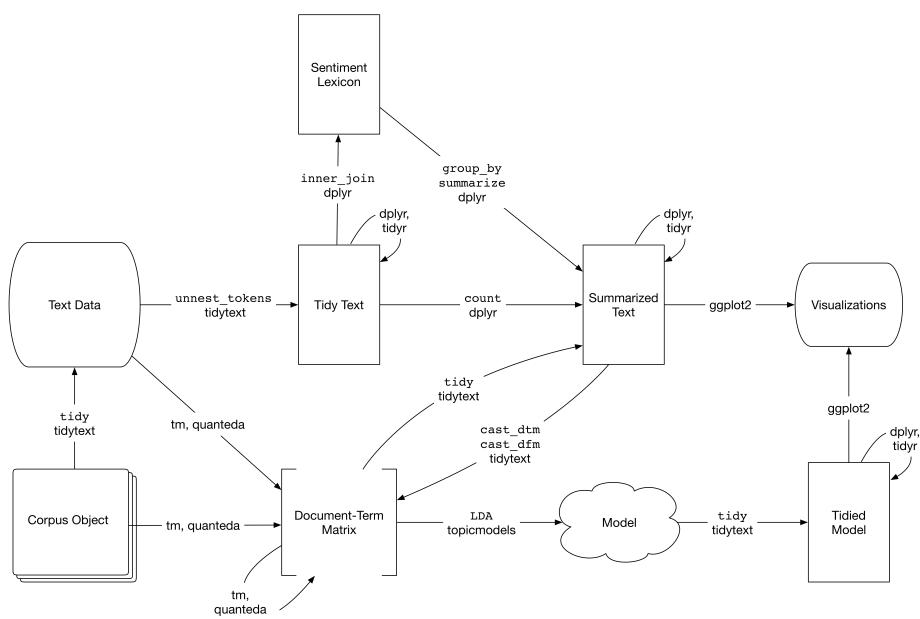


Figure 2.7: Fonte: Text Mining with R

Chapter 3

R e o Processamento de Linguagem Natural

O [processamento de linguagem natural \(NLP\)](#) é um subcampo da ciência da computação relacionado às interações entre computadores e a linguagem humana. O R dispõe de uma [série de pacotes](#) dedicados a essa área e apresenta grande potencial ao conectar o processamento de linguagem natural a todo seu arcabouço de pacotes estatísticos¹.

3.1 Encoding - Codificação de caracteres

Um repertório de caracteres é representado por algum tipo de sistema de codificação ([Wiki](#)). Exemplo comum de sistema de codificação é o código Morse que codifica as letras do alfabeto latino e os numerais como sequências de pulsos elétricos de longa e curta duração. Outro exemplo é o sistema de codificação [UTF-8](#), capaz de codificar todos os 1.112.064 pontos de código válidos em Unicode usando até 8 bits.

O R fornece funções para lidar com diferentes sistemas de codificação. Isso é útil se você lida com arquivos de texto que foram criados com outro sistema operacional e especialmente se o idioma não for o inglês e tiver muitos acentos e caracteres específicos. Por exemplo, o esquema de codificação padrão no Linux é [UTF-8](#), enquanto o esquema de codificação padrão no Windows é [Latin1](#).

A função `Encoding()` retorna a codificação de uma sequência de caracteres. Por sua vez, a função `iconv()` é usado para converter a codificação. Vejamos um exemplo de identificação do encoding de uma sequência de caracteres:

¹Este capítulo tem inspiração nesse [Wikibook](#).

```
chr <- "olê, olê, olê, olá, Lula, Lula"
Encoding(chr) <- "UTF-8"
```

```
Encoding(chr)
```

```
## [1] "UTF-8"
```

Utilizando o resultado do código do bloco acima, vamos agora converter o sistema de codificação para [Latin1](#):

```
chr <- iconv(chr, from = "UTF-8", to = "latin1")
Encoding(chr)
```

```
## [1] "latin1"
```

Para conhecer a lista de sistemas de codificação de seu computador, use a função `iconvlist()`.

3.2 Encoding para remover acentos

Conhecer o sistema de codificação e como utilizá-lo é útil se você lida com arquivos de texto criados com outro sistema operacional e/ou em idiomas que utilizam acentos e caracteres específicos. A depender da análise que deseja fazer, pode ser do seu interesse remover os acentos de uma sequência de caracteres. Nesse caso, vejamos um exemplo com o uso do pacote `stringi`:

```
library(stringi)
chr <- "olê, olê, olê, olá, Lula, Lula"
stri_trans_general(chr, "Latin-ASCII")
```

```
## [1] "ole, ole, ole, ola, Lula, Lula"
```

No exemplo acima, removemos os acentos da sequência de caracteres utilizando o American Standard Code for Information Interchange - [ASCII](#).

Se desejar uma solução caseira, o pacote `txt4cs`, que acompanha o livro, possui a função `remove_accent()`. Vejamos sua aplicação:

```
require(txt4cs)
chr <- "olê, olê, olê, olá, Lula, Lula"
remove_accent(chr)
```

```
## [1] "ole, ole, ole, ola, Lula, Lula"
```

Chapter 4

Strings no R

EM CONSTRUÇÃO...

4.1 Strings

4.2 O pacote stringr

4.3 Regular Expressions no R

Chapter 5

Obtenção de conteúdo

EM CONSTRUÇÃO...

5.1 Webscraping

5.2 Arquivos .pdf

5.3 Twitter

5.4 Áudio Transcrição

5.5 Imagens

Chapter 6

Processamento dos dados

EM CONSTRUÇÃO...

6.1 Tokens

6.2 Corpus

6.3 Tokens e Corpus

6.4 DFM: Matriz de documentos e termos

6.5 Stemming

6.6 FCM: Matriz de co-ocorrência de termos

Chapter 7

Mineração e estatísticas básicas

EM CONSTRUÇÃO...

7.1 Análise de frequência

7.2 Nuvem de palavras

7.3 tf-idf

7.4 Rede de n-grams

7.5 Correlação pareada

7.6 Diversidade lexical

7.7 Similaridade entre documentos/termos

7.8 KEYNESS: Análise de Frequência Relativa

Chapter 8

Escalonamento

EM CONSTRUÇÃO...

8.1 Wordscore

8.2 Wordfish

Chapter 9

Classificação

EM CONSTRUÇÃO...

- 9.1 Método de dicionário: Análise de sentimento**
- 9.2 Naive Bayes**
- 9.3 LDA: Latent Dirichlet Allocation**
- 9.4 STM: Structed Topic Model**

Bibliography

- Grimmer, J. and Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, page mps028.
- Izumi, M. Y. and Moreira, D. C. (2018). O texto como dado: desafios e oportunidades para as ciências sociais. *REVISTA BRASILEIRA DE INFORMAÇÃO BIBLIOGRÁFICA EM CIÊNCIAS SOCIAIS - BIB*, 2(86):138–174.
- Silge, J. and Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media, Beijing ; Boston, edição: 1 edition.
- Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.
- Xie, Y. (2019). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.12.