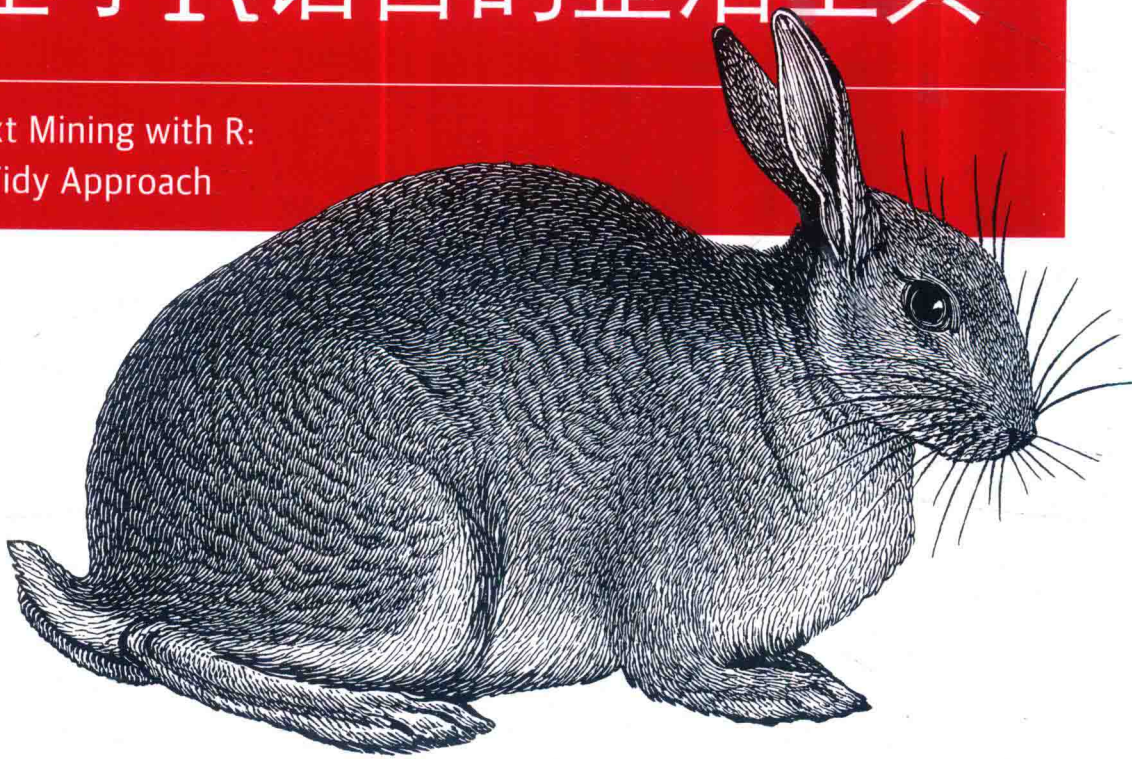
O'REILLY®

HZ BOOKS
华章IT

文本挖掘

基于R语言的整洁工具

Text Mining with R:
A Tidy Approach



Julia Silge David Robinson 著
刘波 罗棻 唐亮贵 译

业出版社
achine Press

文本挖掘：基于 R 语言的整洁工具

Julia Silge
David Robinson 著
刘波 罗棻 唐亮贵 译

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®
O'Reilly Media, Inc. 授权机械工业出版社出版

机械工业出版社

图书在版编目（CIP）数据

文本挖掘：基于R语言的整洁工具 /（美）茱莉亚·斯拉格（Julia Silge），（美）戴维·罗宾逊（David Robinson）著；刘波，罗棻，唐亮贵译．－北京：机械工业出版社，2018.1

(O'Reilly精品图书系列)

书名原文：Text Mining with R: A Tidy Approach

ISBN 978-7-111-58855-9

I. 文… II. ①茱… ②戴… ③刘… ④罗… ⑤唐… III. 程序语言－程序设计 IV. TP312

中国版本图书馆CIP数据核字（2018）第000948号

北京市版权局著作权合同登记

图字：01-2017-7865号

书　　名 / 文本挖掘：基于R语言的整洁工具

书　　号 / ISBN 978-7-111-58855-9

责任编辑 / 陈佳媛

封面设计 / Karen Montgomery，张健

出版发行 / 机械工业出版社

地　　址 / 北京市西城区百万庄大街22号（邮政编码100037）

印　　刷 / 三河市宏图印务有限公司

开　　本 / 178毫米×233毫米　16开本　10.5印张

版　　次 / 2018年3月第1版　2018年3月第1次印刷

定　　价 / 59.00元（册）

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：(010)88379426；88361066

购书热线：(010)68326294；88379649；68995259

投稿热线：(010)88379604

读者信箱：hzit@hzbook.com

O'Reilly Media, Inc. 介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——Wired

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——CRN

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去 Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

译者序

文本挖掘是一种从文本数据中抽取有价值的信息和知识的计算机处理技术，也是自然语言处理的热门话题。本书主要介绍整洁数据的文本挖掘与分析。整洁数据具有简单且新颖的结构，对其进行分析会更有效、更容易。本书的所有代码都是基于R语言来编写的，采用tidytext软件包以及其他整洁工具来挖掘文件中的有用信息，并用图形展示出来，这对理解文本内容非常有帮助。本书提供了非常有用的真实案例，这会为对文本分析工作感兴趣的人提供有价值的信息。

全书共9章，主要介绍如何使用基于R的整洁工具来进行文本分析。首先介绍了整洁文本的格式，以及如何获取整洁文本数据集；并通过tidytext中的情感数据集来进行情绪分析；接着介绍了如何根据tf-idf统计量来识别特定文档中的重要单词，以及如何利用n-gram来分析文本中的文字网络；之后介绍了如何将整洁文本转换为文档词项矩阵和Corpus对象格式，并给出了主题建模的概念；最后通过整合多种已知的整洁文本挖掘方法，给出了一些研究案例，这些案例涉及Twitter归档文件、NASA数据集以及来自新闻组的即时通信信息。总的来说，本书侧重于分析文学、新闻和社交媒体方面的文本，非常适合从事相关文本挖掘的工作人员和自然语言的初学者阅读。与此同时，使用书中提供的大量针对性编程例子，不但可以提高工程实战能力，而且可以在本书提到的整洁框架上建立自己的分析任务。

翻译本书的过程也是译者不断学习的过程。为了保证专业词汇翻译的准确性，我们在翻译过程中查阅了大量相关资料。但由于时间和能力有限，书中内容难免出现差错。如果你在阅读中发现了问题，欢迎通过电子邮件liubo7971@163.com或luofcn@163.com与我们联系，期待与你一起探讨，共同进步。

译者

2017年11月

目录

前言

如果你从事分析或数据科学方面的工作，那么一定熟知这样一个事实：数据正在以前所未有的速度快速生成（也许这样的话有很多人都讲过）。通常培训分析人士来处理数字的表格或规整的数据。但现在大部分新增的数据都是非结构化的文本，而许多在分析领域工作的人都没有接受过乃至简单接受过处理自然语言方面的训练。

尽管我们熟悉许多数据处理和可视化方法，但是将这些方法应用于文本处理并非易事，所以开发了 tidytext R 包（Silge 和 Robinson，2016）。我们发现采用数据整洁原则可以使许多文本挖掘任务变得更简单、更有效，并且该原则和广泛使用的工具也是相一致的。把文本当作由单个单词构成的数据框的优势在于：（1）有助于轻松地操作、汇总以及展示文本特征；（2）有助于将自然语言处理整合到有效的工作流程中。

本书介绍了如何使用 tidytext 包以及其他基于 R 语言的 tidy 工具来进行文本挖掘。tidytext 包提供的函数相对简单，但如何使用这个包则很重要。因此，本书还提供了真实的、极具吸引力的文本挖掘案例。

大纲

本书首先介绍整洁文本格式，一些有关 dplyr、tidyr 和 tidytext 包的使用方法则按如下过程来介绍：

- 第 1 章概述了整洁文本格式和 unnest_tokens() 函数，同时介绍了 gutenbergr 和 janeaustenr 包，这些包提供了与文学相关的文本数据集，本书会使用这些数据集来进行介绍。
- 第 2 章介绍了如何使用 tidytext 中的 sentiments 数据集以及 dplyr 包中的 inner_join() 函数来对整洁文本数据集进行情感分析。

- 第 3 章介绍了 tf-idf 统计量（词项频率乘以逆文档频率），它可用来识别特定文档中特别重要的词项。
- 第 4 章介绍了 n-gram 以及如何使用 widyr 包和 ggraph 包来分析文本中的文字网络。

文本在分析的所有阶段并不是整洁的，能够在整洁和不整洁格式之间进行转换就显得非常重要。

- 第 5 章介绍了通过 tm 包和 quanteda 包来使文档 – 词项矩阵和 Corpus 对象变整洁的方法，以及如何将整洁文本数据集转换为文档 – 词项矩阵和 Corpus 对象格式。
- 第 6 章介绍了主题建模的概念，并使用 tidy() 方法对 topicmodels 包的输出进行解释和可视化。

通过整合多种已知的整洁文本挖掘方法，还给出了几个研究案例：

- 第 7 章通过作者自己的 Twitter 档案展示了整洁文本分析的应用。例如，Dave 和 Julia 的 Twitter 习惯有什么不同？
- 第 8 章通过查看超过 32 000 个 NASA 数据集（可用于 JSON 格式）中的关键字与标题、描述字段的关系来探索元数据。
- 第 9 章分析不同新闻组（与政治、曲棍球、技术、无神论等有关的主题）的即时通信消息数据集来了解新闻组中共同的模式。

本书不包括的主题

本书对整洁文本挖掘框架进行了介绍，并给出了一系列的示例，但对于全面研究自然语言处理领域而言，这些依然不够。CRAN Task View on Natural Language Processing（*https://cran.rproject.org/view=NaturalLanguageProcessing*）提供了其他使用 R 进行计算语言学研究的详细信息。根据个人需求，你可能还想在以下方面进一步研究：

聚类、分类和预测

文本机器学习是一个广泛的话题，可以轻松地找到很多与之相关的内容。第 6 章将介绍一种无监督聚类（主题建模）方法，但是还有更多其他的机器学习方法可以用来处理文本。

词嵌入

当前流行的一种文本分析方法是将单词映射为向量，以便能检查单词之间的语言关系并对文本进行分类。尽管这些单词表示并不像我们理解的那样整洁，但已经可以在机器学习方法中得以广泛应用。

更复杂的词条化

tidytext 包通过信任词条化包（Mullen，2016）来进行标记，其本身使用统一的界面并包括各种词条化方法，但是在具体的应用程序中还有许多其他的词条化方法。

除英文以外的其他语言

一些用户已经成功地将 tidytext 应用于除英语以外的其他语言的文本挖掘，但是本书不涵盖这方面的例子。

关于本书

本书重点介绍实际软件示例和数据展示，几乎没有公式，但是有大量的代码。我们重点关注在分析文学、新闻和社交媒体时的深入理解。

本书不需要读者具有文本挖掘知识，而专业语言学家和文本分析师可能会认为本书的示例比较初级，但我们相信，他们也可以在这个框架上建立自己的分析。

本书假设读者至少熟悉 R 中的 dplyr、ggplot2 和 %>%（管道）运算符，并且对如何应用这些工具进行文本数据挖掘感兴趣。对于没有这种专业背景的读者，推荐阅读 Hadley Wickham 和 Garrett Grolemund（O'Reilly）的《R for Data Science》一书。若读者有一点背景并对整洁文本感兴趣，即使是 R 初学者也可以理解和使用本书的示例。

如果你正在阅读本书的纸质版本，那么图像会以灰度而不是彩色的形式呈现。要查看彩色版本的图像，请参阅本书的 GitHub 页面（*http://github.com/dgrtwo/tidytext-mining*）。

本书约定

本书使用以下惯例：

斜体（Italic）

表示新的术语、网址、电子邮件地址、文件名和文件扩展名。

等宽字体（`Constant width`）

用于程序清单，以及段落中引用的程序元素，如变量或函数名称、数据库、数据类型、环境变量、语句和关键字。

等宽粗体（**`Constant width bold`**）

展示用户应直接输入的命令或其他文字。

等宽斜体（*Constant width italic*）

表示应使用用户提供的值来替换或由上下文确定的值。

表示提示或建议。

表示普通注释。

表示警告或注意。

使用代码示例

本书在大部分分析的过程中都给出了代码，但出于篇幅考虑，如果生成图形的代码已经出现过，则不再提供类似的代码。相信读者可以学习并延伸本书示例，另外本书代码可以在 GitHub 公共库中找到。

本书旨在帮助读者完成工作，一般来讲，读者可以在程序和文档中使用本书提供的示例代码。除非对代码的重要部分进行加工出版，否则不需要与我们联系。例如，使用本书中多个代码块开发程序不需要经过我们许可，但出售或发行 O'Reilly 书籍示例的 CD-ROM 则需要许可，引用本书和示例代码来回答问题不需要许可，将本书中重要的示例代码合并到产品文档则需要许可。

如果你引用了本书中的内容，我们希望你能注明出处，包括标题、作者、出版商和 ISBN。例如："Text Mining with R by Julia Silge and David Robinson (O'Reilly). Copyright 2017 Julia Silge and David Robinson, 978-1-491-98165-8"。

如果你认为本书代码示例或上述许可不合理，请随时通过 permissions@oreilly.com 与我们联系。

Safari® 在线图书

Safari 是一个为企业、政府、教育和个人提供的会员制培训、参考平台。

会员可以访问数以千计的书籍、培训视频、学习路径、互动教程以及来自 250 多个出

版社策划的播放列表，包括O'Reilly Media、Harvard Business Review、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Adobe、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett，以及其他在线技术。

更多信息请访问：*http://oreilly.com/safari*。

联系我们

对于本书，如果有任何意见或疑问，请按照以下地址联系本书出版商。

美国：

O'Reilly Media，Inc.
1005 Gravenstein Highway North
Sebastopol，CA 95472

中国：

北京市西城区西直门南大街2号成铭大厦C座807室（100035）
奥莱利技术咨询（北京）有限公司

要询问技术问题或对本书提出建议，请发送电子邮件至：

bookquestions@oreilly.com

要获得更多关于我们的书籍、会议、资源中心和O'Reilly网络的信息，请参见我们的网站：

http://www.oreilly.com
http://www.oreilly.com.cn

我们在Facebook上的主页：*http://facebook.com/oreilly*

我们在Twitter上的主页：*http://twitter.com/oreillymedia*

我们在YouTube上的主页：*http://www.youtube.com/oreillymedia*

致谢

非常感谢为推进本项目做出贡献、提供帮助和观点的人，这里特别感谢几个人和组织。

感谢Oliver Keyes和Gabriela de Queiroz对tidytext的贡献、Lincoln Mullen在tokenizers

软件包方面的工作、Kenneth Benoit 在 quanteda 软件包方面的工作，Thomas Pedersen 在 ggraph 软件包方面的工作，以及 Hadley Wickham 在制定数据整洁原则和构建整洁工具方面的工作。还要感谢 Karthik Ram 和 rOpenSci，他们在项目之初进行召集，感谢 NASA Datanauts 项目成员，感谢你们在项目期间对 Julia 提供的机会和支持。

仔细、彻底的技术审查大大提高了本书的质量。特别感谢 Mara Averick、Carolyn Clayton、Simon Jackson、Sean Kross 和 Lincoln Mullen，感谢你们在技术评论中投入的时间和精力。

本书是以公开方式进行撰写的，有几个人通过提出要求或问题的方式提供了建议。特别感谢那些通过 GitHub 做出贡献的人：@ainilaha、Brian G. Barkley、Jon Calder、@eijoac、Marc Ferradou、Jonathan Gilligan、Matthew Henderson、Simon Jackson、@jedgore、@kanishkamisra、Josiah Parry、@ suyi19890508、Stephen Turner，以及 Yihui Xie。

最后，我们想把本书献给各自的爱人——Robert 和 Dana，千言万语汇成一句发自肺腑的谢谢。

第 1 章

整洁文本格式

使用整洁数据原则是一种更容易、更有效的数据处理方式，这在处理文本时也是如此。Hadley Wickham（Wickham，2014）认为整洁数据的结构为：

- 每个变量是一列
- 每次观察是一行
- 每次观察的结果会构成一张表

因此，可将整洁的文本格式定义为表的每行都有一个词条（token）。词条是一个有意义的文本单元，例如在分析时感兴趣的单词，而词条化是将文本分解为词条的过程。这种每行一个词条（one-token-per-row）的结构与当前分析文本时采用字符串或文档 – 词项（document-term）矩阵的存储方式形成对比。对于整洁文本挖掘，存储在每行的词条通常是一个单词，但也可以是 n-gram、句子或段落。tidytext 包能通过常用文本单元来进行词条化的功能，并将其转换为每行一个词条的格式。

整洁数据集允许使用一套“简洁”工具进行操作，包括诸如 dplyr（Wickham 和 Francois，2016），tidyr（Wickham，2016），ggplot2（Wickham，2009）和 broom（Robinson，2017）等流行包。通过保证输入和输出为整洁表格的形式，用户在这些包之间的转换很容易。这些简洁工具能扩展到许多文本分析和研究中。

同时，tidytext 软件包并不期望用户在分析过程中始终保证文本数据是整洁的。该软件包基于文本挖掘 R 包，例如 tm（Feinerer 等人，2008）和 quanteda（Benoit 和 Nulty，2016），它包括 `tidy()` 对象（参见 broom 包）的功能。这个包可以使用诸如 dplyr 和其他整洁工具的工作流，即导入、过滤和处理文本，将数据转换为机器学习应用中的文档 – 词项矩阵，最后可用 ggplot2 将模型重新转换成整洁形式进行解释和可视化。

比较整洁文本结构与其他数据结构

如上所述，我们将整洁文本格式定义为每行一个词条形式的表。以这种方式构建文本数据是符合整洁数据原则的，可以通过一组一致的工具来进行操作。值得将其与经常在文本挖掘方法使用的文本存储方式进行比较：

字符串（String）

当然，文本可以作为字符串（即，字符向量）存储在 R 内，通常可以先将这种数据读入内存中。

语料（Corpus）

这些类型的对象通常含有原始字符串，同时还包含标注这些字符串的元数据和详细信息。

文档 – 词项矩阵（Document-term matrix）

这是一个描述文档集合（如语料库）的稀疏矩阵，该矩阵的行表示一个文档，列表示词项，矩阵的值通常是数字或 tf-idf 值（参见第 3 章）。

本书第 5 章还会继续探究语料和文档词 – 项矩阵，现在先了解将文本转换为整洁格式的基础知识。

unnest_tokens 函数

Emily Dickinson 写了一些可爱的文字。

```
text <- c("Because I could not stop for Death -",
          "He kindly stopped for me -",
          "The Carriage held but just Ourselves -",
          "and Immortality")

text
```

```
## [1] "Because I could not stop for Death -"    "He kindly stopped
for me -"
## [3] "The Carriage held but just Ourselves -" "and Immortality"
```

这是一个我们可能想要分析的典型字符向量。为了将其变成一个整洁文本数据集，首先需要将其放入一个数据框（data frame）中。

```
library(dplyr)
text_df <- data_frame(line = 1:4, text = text)

text_df
## # A tibble: 4 × 2
##    line                                  text
```

```
##   <int>                                    <chr>
## 1     1   Because I could not stop for Death -
## 2     2              He kindly stopped for me -
## 3     3 The Carriage held but just Ourselves -
## 4     4                           and Immortality
```

这意味着数据框会作为一个 tibble 输出？`tibble` 是 R 中新的数据框类（class），在 `dplyr` 和 `tibble` 包中有效，它打印方便，不会将字符串转换为元素，也不使用行的名字。`tibble` 能很好地支持整洁工具。

注意这种包含文本的数据框架与整洁文本分析不兼容。我们不能过滤频繁出现的单词或计数，因为每行都由多个组合的单词构成。为了得到每个文档每行词条（*one token per document per row*）的形式，需要将其转换为数据框架。

词条是一个有意义的文本单元，通常是人们需要进一步分析的单词，词条化是将文本分解为词条化的过程。

在第一个例子中，只有一个文档（一首诗），但接下来就会研究多个文档的例子。

在整洁文本框架中，需要将文本分为单个的词条（即词条化过程），并将其转换为整洁的数据结构。因此需要使用 `tidytext` 的 `unnest_tokens()` 函数。

```
library(tidytext)

text_df %>%
  unnest_tokens(word, text)
```

```
## # A tibble: 20 × 2
##     line    word
##    <int>   <chr>
## 1      1 because
## 2      1       i
## 3      1   could
## 4      1     not
## 5      1    stop
## 6      1     for
## 7      1   death
## 8      2      he
## 9      2  kindly
## 10     2 stopped
## # ... with 10 more rows
```

`unnest_tokens` 使用的两个基本参数是列名。第一个参数为输出结果的列名，在函数执行前，文本（在这种情况下是词）还没有放入到该列中；第二个参数为输入列（在这种情况下为文本）。请记住，上面的 `text_df` 有一个名为 `text` 的列，该列包含了需要分析的数据。

使用 unnest_tokens 后，为保证新数据框中每行只有一个词条（word），必须拆分每行；unnest_tokens() 函数默认是对单个词进行词条化，如上所示。注意：

- 其他列会被保留，例如每个词的行号。
- 标点符号已被删除。
- 默认情况下，unnest_tokens() 将词条转换为小写，这使其更容易与其他数据集进行比较或组合（使用 to_lower = FALSE 参数可以关闭该功能）。

这种格式的文本数据可以使用标准的整洁工具集来操作、处理和可视化文本，这个过程采用的工具分别是 dplyr、tidyr 和 ggplot2（如图 1-1 所示）。

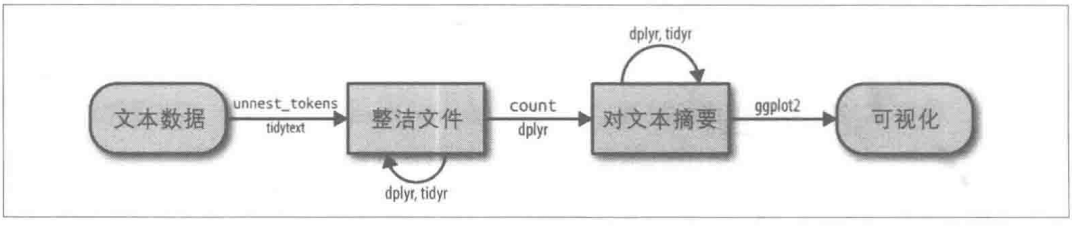


图 1-1：使用整洁数据原则进行文本分析的流程图。本章将介绍如何使用这些工具汇总和可视化文本

整理 Jane Austen 的作品

janeaustenr（*https://cran.r-project.org/package=janeaustenr*）包中有 Jane Austen 的六本小说的电子文档（Silge，2016），在研究这些文本之前，将这些小说转换成整洁格式。janeaustenr 包能提供 one-row-per-line 格式的文本，本文中的一行类似于纸制书中的一行。这里会使用 mutate() 函数来得到一个新的 linenumber，它保存了原始数据中的行号，使用关键词“chapter”（使用正则表达式）来查找所有章节的位置。

```
library(janeaustenr)
library(dplyr)
library(stringr)

original_books <- austen_books() %>%
  group_by(book) %>%
  mutate(linenumber = row_number(),
         chapter = cumsum(str_detect(text, regex("^chapter [\\divxlc]",
                                                 ignore_case = TRUE)))) %>%
  ungroup()

original_books
## # A tibble: 73,422 × 4
##                    text                book linenumber chapter
```