用R语言分析文学文本

Text Analysis with R for Students of Literature

马修·L·乔克斯 著
汪顺玉 赵晴 译
陈萍 校

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
www.sflep.com

用R语言分析文学文本

Text Analysis with R for Students of Literature

马修·L·乔克斯 著
汪顺玉 赵 晴 译
陈 萍 校

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

图书在版编目(CIP)数据

用R语言分析文学文本 /（美）马修·L·乔克斯著；汪顺玉，赵晴译.
—上海：上海外语教育出版社，2018
ISBN 978-7-5446-3099-3

Ⅰ. ①用… Ⅱ. ①马… ②汪… ③赵… Ⅲ. ①程序语言-应用-文学研究
Ⅳ. ①I06

中国版本图书馆CIP数据核字(2018)第021018号

出版发行：上海外语教育出版社
（上海外国语大学内） 邮编：200083
电　　话：021-65425300（总机）
电子邮箱：bookinfo@sflep.com.cn
网　　址：http://www.sflep.com
责任编辑：李健儿

印　　刷：上海同济印刷厂有限公司
开　　本：787×1092　1/16　印张 13.5　字数 214千字
版　　次：2018年6月第1版　2018年6月第1次印刷
印　　数：1 100册

书　　号：ISBN 978-7-5446-3099-3 / H
定　　价：50.00元

本版图书如有印装质量问题，可向本社调换
质量服务热线：4008-213-263　电子邮箱：editorial@sflep.com

译者序

随着数字时代的到来，信息技术正在深度地影响和改造我们的学习、生活、工作与研究方式。各个学科的交叉融合趋势日益明显。计算分析、统计模型、机助软件等方法正在打破传统的学科边界。基于大数据的分析与可视化研究，正在与文学、语言学、历史学、政治学、传播学等传统的人文和社会科学联姻。本书《用 R 语言分析文学文本》(*Text Analysis with R for Students of Literature*)也是这种联姻的成果。

选择翻译此书的动因有三。首先，作者马修·L·乔克斯(Matthew L. Jockers)是美国著名的数字人文专家，从事文学文本量化和可视化方面的开拓性研究，近几年出版了《宏大分析：数字方法与文学史》(*Macroanalysis: Digital methods and literary history*)以及本书。这些研究突破了传统文学研究对单文本或者小样本作品的精读(close reading)的藩篱，开启了基于大数据的文学文本量化和可视化分析模式，倡导远距离阅读(distant reading)。其二，作为多功能分析工具的 R 语言，在经济学科领域、生命科学领域、工程领域等都得到大量的应用，针对这些领域的专门 R 语言运用出版物也算丰富；用 R 语言做语言学科研究的书，也出版了几部，如《R 语言定量语料库语言学》(*Quantitative Corpus Linguistics with R*)、《分析语言数据：实用统计学导论》(*Analyzing Linguistic Data: A practical introduction to statistics*)、《第二语言研究中的 R 统计分析指南》(*A Guide to Doing Statistics in Second Language Research Using R*)等。其三，在文学领域使用 R 语言进行研究的著作还十分罕见。翻译本书的目的是：将基于大数据的文学量化和可视化方法进行引介；相信对于国内还未了解该方法的年轻学者和研究生，应该有一定的拓展研究方法、拓宽研究视野的作用。

本书作者马修·L·乔克斯是美国内布拉斯加大学艺术和科学学院分管研究和合作的副院长,该大学数字人文中心的研究员。他的研究兴趣集中在使用计算方法进行文学研究。本书讲述了R语言的基础知识,可让文学方向的学生在学习R语言的同时免去学习统计知识而带来的困难,作者特别将统计的知识维持在较低限度。本书理论与实践结合紧密,书中使用的软件和文本材料均发布在与本书配套的网站上(http://www.mathewjockers.net/text-analysis-with-r-for-students-of-literature)。本书在分析层次的安排上,也呈现了循序渐进的特点。在掌握了R语言的基本知识后,后面三大部分分别是聚焦词汇层面的微观分析、句子层面的中观分析和篇章内容层面的宏观分析;特别是宏观分析中的聚类分析、分类分析和主题建模分析,可期对文学文本量化研究带来新的视野。

本书由我的同事赵晴女士进行了初译,陈萍女士进行了校译。本人对全部校译后的译文进行了通读和统稿。在翻译的过程中,为了校对方便,陈璇老师带领我的研究生周雅、胡琴琴、李诗品、张琼芳对译稿进行了段对段的英汉双语对齐处理。为了保证R语言按照书中所述运行无误,我还请了计算机专业研究生张其龙对全书的R代码进行了运行。绝大多数代码运行得到的数据与书中所写一致。发现了两个小误差,与作者本人进行核对并得到他的认可(见附录E)。

我要感谢原书作者为本书的中文版写序;感谢上海外语教育出版社出版此书,特别感谢责任编辑李健儿先生对本书细致、全面和专业的编校。最后,由于译者水平有限,译文中难免存在纰漏,恳请读者指正并不吝赐教。

汪顺玉
记于重庆邮电大学
2017年12月
williamwsy@hotmail.com

序　言

本书介绍使用公开源代码编程语言 R 进行计算文本分析的方法。与其他类似的书籍不同，本书的目的既不是介绍如何用 R 对语言数据进行统计分析[1]，也不是介绍如何用 R 进行定量的语料库语言学分析[2]，而是旨在为文学领域学生和研究者——甚至广义的人文研究者，提供一些分析文本的定量和计算方法。本书力求做到简明扼要。R 语言是一套复杂的程序，没有任何一本的书能把它完全解释清楚。本书的重点是让技术不再乏味。更重要的是，让技术发挥作用并能立即带来回报！这里的“回报”不是掌握了一种编程语言所获得的满足感，而是你能即刻将自己学到的东西付诸实践。你可以立即着手分析和处理文本，而且每一章将带领你学会一种新的技术或分析方法。

计算的方法能为我们提供的信息，是用传统的仔细阅读和综合分析文本这些定性方法所不能得到的。这些信息既可归属于宏观层面也可归属于微观层面。如果你读完这本书，对这些核心技术有了一个基本认识，并对这些技术带来的种种可能性有了一个整体了解，那这本书的目的就达到了。当你把这本书搁置一边，开始创建你自己的项目时，真正的学习就算开始。我的目标是为你提供足够的背景知识，这样你就能自如地开

[1] Baayen, H. A. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge UP, 2008.

[2] Gries, Stefan Th. *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York: Routledge, 2009.

展自己的项目，而且能在这个过程中持续提升自己。

我是一个计算人文学者。当论及我工作的时候，人们常常问我这个问题：你倡导的方法是否真的能为人文学科带来新知？我的回答坚定而确切：能。当然，这个坚定的“能”必须有所保留——不是文本分析揭示的一切都称得上开创性的发现。大量计算工作的目的在于检验、摒弃或再次证实我们认为已经掌握的知识。在一次讲座中，我以《白鲸》这部小说为例，分析了19世纪小说文学风格的宏观模型。我演示了《白鲸》在一千多部19世纪的美国小说中如何呈现统计学意义上的异常。这时，有一位同事举手质疑：学者们已经知道《白鲸》是一部脱离常规的小说，为什么还要用计算的方法证明一个大家都知道的问题？

我同事的问题问得好，令人脑洞大开。这个问题也揭示了人文学科的传统，同时也极具讽刺意味——这个学科一直以来赞同这么一个观点：文学争论永无止境。我们真的知道《白鲸》是一部脱离常规的小说吗？或许《白鲸》只是在与其他二三十部美国小说（一直以来我们常拿它们与《白鲸》比较）的比较中显得异常而已。我举《白鲸》的例子不是为了假装声称对这部小说在美国文学传统中的地位有了新发现，而是为此提供一种新的证据，带来一个新的视角，以加固现有的假设。

如果一种新的证据恰好证实了我们用更主观的方法得到的知识，那么这种新证据难道不应该被视为是一件好事吗？如果火星探测器最近发回了更多的证据，证明这个星球以前有生命存在，那么这个新证据就尤为重要。尽管它不会比第一次在火星上发现微生物或冰更令人震惊或激动，但它毕竟是一个重要证据，而且会为我们最终了解这个星球添砖加瓦。所以，文学研究的计算方法能提供补充证据，这是一件好事。

本书介绍的方法还能提出矛盾证据，这些证据挑战了我们传统中凭印象或凭经验得出的理论。从这个意义上说，这些方法为我们提供了证伪（卡尔·波普尔和后实证主义提出的调和严格实证主义和严格相对主义之间矛盾的概念）的机会。但正因为这些方法能反驳我们现有的理论，我们一定不能陷于数字游戏中，只重视可以被检验的想法；有些解释可以通过计算或定量的方法加以验证，但有些不行。我认为这也是一件好事。

最后,这些方法能为我们带来全新的发现。计算文本分析能为我们带来用肉眼无法发现的东西[1]。用计算方法,帕特里克·尤奥拉(Patrick Juola)最近发现罗琳(J.K. Rowling)是《布谷鸟的呼唤》一书(这部作品是她用罗比特·加尔布雷斯[Robert Galbraith]的笔名写的)的真正作者。自然,我认为尤奥拉的发现也是一件好事。

这就是关于文本分析我想要说的。在我另一本书中,我更热衷于为我的方法辩护[2],但本书的目的不是为该方法辩护,而是分享这些方法。

马修·L·乔克斯
(Matthew L. Jockers)
2014年1月于
内布拉斯加州林肯市

[1] 见 Flanders, Julia. "Detailism, Digital Texts, and the Problem of Pedantry." *TEXT Technology*, 2:2005, 41-47

[2] Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History*. UIUC Press, 2013.

中文版序

在《用R语言分析文学文本》第1版成书之际，我就希望其成为人文学界和文学研究领域的学生和学者学习计算文本分析的入门之作。本书正式出版之后，赢得了广泛的读者群，尤其对使用英语的许多学生和学者而言，已经达成其预期目的。

我特别感激重庆邮电大学汪顺玉教授和他的团队，不辞辛劳地将本书翻译成中文；特别高兴地获悉上海外语教育出版社将出版此书中文版，为其赢得新的读者群。

希望本书能给中国读者带去无数英语读者一样的阅读感受——即计算文本分析的入门导引；同时，本人乐见中国的文学领域和人文学界涌现喜人的研究新成果。

马修·L·乔克斯

内布拉斯加州林肯市

2017年3月

Preface to the Chinese Edition

When I completed the first edition of *Text Analysis with R for Students of Literature*, my hope was that the book would offer students and scholars in the humanities and literary studies a starting point for learning computational text analysis. Since publication the book has found a wide audience, and many students and scholars in the English-speaking world have found that the book has lived up to its mission.

I am incredibly grateful to Wang Shunyu and his team at Chongqing University of Posts and Telecommunications for their work translating this book into Chinese, and I am incredibly pleased that Shanghai Foreign Language Education Press is offering this translation for a new audience.

I hope that the book will provide for Chinese readers the same entry point that it has provided for so many English readers, and I look forward to seeing exciting new discoveries from Chinese students and scholars of literature and the humanities.

Matthew L. Jockers
Lincoln, NE
March 2017

致 谢

多年来，我用多种工具和编程语言教授文本分析课程。先后教学生用 Perl、Python、php、Java 甚至 XSLT 进行文本解析，用 Excel 分析结果数据。大约在 2005 年，或许是看了 Claudal Engal 和 Dianiela Witten 关于如何教初学者语言的建议而受到启发，我不再使用 Excel，转而开始使用 R 语言。在那之后的很长时间，我仍然编写了很多 Perl、Python、php 的文本分析的代码，并把结果导入 R 语言进行分析。2008 年，我发现这样的模式不具有可持续性，因为我不得不花很多时间把数据从一个环境转移到另一个环境。因钟情于使用 R 语言，我决定忍痛割爱，放弃其他语言。自从使用 R 语言以来，我就几乎没有再求诸其他工具了。

幸运的是，同我转换到 R 语言一样，我的很多同行也开始使用 R 语言。差不多与我开始使用 R 起同步，由 R 语言的程序员和研发人员组成的网上社区也不断扩大。目前的在线 R 语言帮助资源异常丰富。如果没有这些资源，我根本完成不了这本书。一些优秀的 R 语言程序员开发了不少非常有用的软件包。本书只提到少数几个软件包（毕竟，这本书只是一本初学者指南）。但是如果没有像 Stefan Th. Gries、Harald Baayen 和 Hadley Wickham 的贡献，这本书和 R 语言社区一定没有这么丰富。我欣喜于这个网上社区的实用和友好。早年间，情况可不是这样的。因此，我想感谢所有开发软件包并对 R 语言项目编写代码的人；我还要感谢所有就 R 语言平台，尤其是对 R 语言帮助列表和 `stackoverflow.com` 网站，给我提出建议的人。

这本书源于我在斯坦福大学任教时留给学生的一系列练习题：那些学生看到的大部分内容比较原始、粗糙。本书能够出版，有太多的人我想表示感谢。在此，一并感谢我所有教过的学生及我正在教的学生，感谢他

们的耐心和意见反馈。这本书不管有什么瑕疵和缺点，但正是因为有你们，书稿质量已经得到大幅提升。

2011 年，我先把所教班级学生的材料编成文稿，然后同我的几个同事分享了这些零零碎碎的文本。Stéfan Sinclair 在 McGill 学校任教时，运行程序测试了这些文本，从而初步形成了本书的第一稿。他和他的学生还提供了宝贵的反馈意见。Maxim Romanov 在 2013 年初阅读了大部分初稿，给了我鼓励和建议，并最终说服我将初稿交由 LaTeX 公司 进行专业水平的排版。这最终让我接触到了 Sweave 和 Knitr 软件包。这两个 R 软件包使我可以将初稿嵌入并运行 R 代码。因此，在这里我无比感激 Maxim，以及开发 Sweave 和 Knitr 软件包的人，还有 Friedrich Leisch 和谢益辉。我感谢那些下载我在 2013 年 8 月贴在个人网页上的初稿草稿的人。他们给我了反馈意见和提供更正意见，名单特别载明在随后的“赐正者”页面里。最后，我还要感谢我 14 岁大的儿子对这本书中的每一行编码都了如指掌。尽管这本书不是浅显易懂到每个人都看得懂，但我的确知道，机敏的中学生都不需要太多指导就能看得明白的。

赐正者

自本书初稿于2013年8月被上传分享至本人网站后，下载次数逾千。如下所列读者对初稿提供了宝贵的反馈意见，他们的赐正让我的终稿更显完善，在此谨一一谢过。具体赐正意见均存档于如下网址：http://www.matthewjockers.net/text-analysis-with-r-for-studentss-of-literature/。最慷慨的赐正者是Charles Shirley，他提供了133条意见。编码方面最重要的赐正意见来自Carmen McCue，她在第四章里发现了一个颇为低级的小漏洞。再次对诸位深表感谢！

1. Brotnov, Mikal
2. Francom, Jerid
3. Hawk, Brandon
4. Huber, Alexander
5. Johnson, Paul
6. Kumari, Ashanka
7. Laudun, John
8. Maenner, Matthew J.
9. McCue, Carmen
10. McMullen, Kevin
11. Pentecost, Stephen
12. Shirley, Charles
13. Tedrow, Kimberley
14. Wehrwein, Austin
15. Wolff, Mark
16. Xie, Yihui

目 录

附 录

第一部分
微观分析