Nabokov's Favorite Word Is Mauve

纳博科夫最喜欢的词

WHAT THE NUMBERS REVEAL ABOUT THE CLASSICS, BESTSELLERS AND OUR OWN WRITING

Ben Blatt

[美] 本•布拉特 著

杜森 译

文学没有标准答案
数据可以帮我们寻找答案

伟大作家有什么样的创作偏好？
统计学视角下的经典作品有何规律？

用大数据分析文学，探究黄金写作法则

北京联合出版公司 Beijing United Publishing Co.,Ltd.

纳博科夫最喜欢的词

WHAT THE NUMBERS REVEAL ABOUT THE CLASSICS,
BESTSELLERS AND OUR OWN WRITING

Nabokov's
Favorite Word
Is Mauve

[美] 本·布拉特 著　　杜森 译

Ben Blatt

北京联合出版公司 · 旧音

图书在版编目（CIP）数据

纳博科夫最喜欢的词 /（美）本·布拉特著；杜森译 . — 北京：北京联合出版公司，2019.1

ISBN 978-7-5596-2691-2

Ⅰ . ①纳… Ⅱ . ①本… ②杜… Ⅲ . ①文学研究 Ⅳ . ① I0

中国版本图书馆 CIP 数据核字（2018）第 230414 号

纳博科夫最喜欢的词

作　　者：[美]本·布拉特（Ben Blatt）

译　　者：杜　森

出版监制：刘　凯　马春华

选题策划：联合低音

责任编辑：唐乃馨　周　杨

封面设计：7拾3号工作室

内文排版：红点印像

北京联合出版公司出版

（北京市西城区德外大街83号楼9层　100088）

北京联合天畅文化传播公司发行

北京华联印刷有限公司印刷　新华书店经销

字数232千字　889毫米×1194毫米　1/32　10.5印张

2019年1月第1版　2019年1月第1次印刷

ISBN 978-7-5596-2691-2

定价：60.00元

汉密尔顿（Alexander Hamilton）、麦迪逊（James Madison）、还是杰伊（John Jay）？

《联邦党人文集》[①]为美国走向民主打下基础，其中有12篇文章的作者未明，历史学家们为此争论了150多年。尽管这些文章在美国史的语汇中是世人皆知的标志性作品，但每一篇的作者究竟是谁却一直是团疑云。哪一位开国元勋撰写了这些篇章？这个问题激起了无尽的争论，后来甚至成了历史学家聚会时客厅里一个广为流行的猜谜游戏。美国的治国框架建立在这些振奋人心的论述之上，可到底是谁写下了这些文章呢？

答案隐藏在文章的词语中，但要找到这些词语，学者们无

① 《联邦党人文集》被誉为美国宪法的圣经，是有关美国宪法和联邦制度的评论文章合集，共85篇文章，由18世纪80年代三位美国政治家亚历山大·汉密尔顿、詹姆斯·麦迪逊、约翰·杰伊在制定美国宪法的过程中写成。

须精读文本，只要细细地数一下数。他们所要看的只是数字。

疑云始于1787年年末，当时纽约的报纸刊登了一系列鼓吹新宪法的文章，用的笔名是普布利乌斯（Publius，源自古罗马执政官Publius Valerius Publicola）。用一个具有爱国含义的笔名来隐藏自己的身份似乎有点儿可笑。实际上，在当时美国近400万居民中，只有三个人才有资格进入这场关于作者身份争议的角逐。

汉密尔顿、麦迪逊和杰伊撰写了这些文章——这在当时是一个公开的秘密，但三个人都不想站出来承认写过哪些特定文章。他们都有自己的政治抱负，后来分别升任财政部长、总统和最高法院首席大法官，所以他们有充足的理由隐藏自己的作者身份。但他们的过分谨慎留下了难以攻破的疑云，在之后的时日中，同时撩动着历史教授和热心的业余爱好者的神经。

你也许会以为，当时的学者和精明政客应能确定作者是谁。毕竟只有三个潜在的候选人，每人都有自己的政治倾向，交流表述的风格也各不相同。如果放在今天，这个问题可能相当于《纽约时报》刊登了一篇匿名社论，执笔者可能是奥巴马、希拉里或桑德斯，也可能是小布什、麦凯恩或特朗普，即使三个人来自同一阵营（前三人是民主党，后三人是共和党），但观点和文笔不可能完全一样。

到了1804年，答案似乎终于浮现。汉密尔顿给他的朋友本森（Egbert Benson）写了一封信，信中列出每一篇文章的作者。当时汉密尔顿正准备与美国副总统伯尔（Aaron Burr）

决斗[①]，他意识到《联邦党人文集》的重要历史意义，也明白自己可能无法从决斗中生还。最终，他决定不让这些答案随他一同逝去。

疑云本应就此告终，全国上下关注此事的人没有理由怀疑汉密尔顿的第一手信息。但 13 年后，麦迪逊在结束他的第二个总统任期后不久，列出了他的著作清单，与汉密尔顿当年所说有出入。其中，汉密尔顿认领的 12 篇文章，麦迪逊声称是他的作品。

此事点燃了群众的新一波热情，历史学家们为此又争吵了一个多世纪。1892 年，洛奇（Henry Cabot Lodge，曾担任参议员）为此问题著文，赞同汉密尔顿的说法，而著名历史学家伯恩（E. G. Bourne）则认为那 12 篇文章的作者是麦迪逊。

大多数历史学家试图根据每篇文章的政治理念进行梳理，确定作者——麦迪逊真的会用那些措辞主张设立中央银行？汉密尔顿会如此直接地支持针对国会的限制？也许这一篇是杰伊写的？

直到两个世纪以后的 1963 年，问题才最终得以解决。两位受人尊敬的教授——哈佛大学的莫斯特勒（Frederick Mosteller）和芝加哥大学的华莱士（David Wallace）——给出了明确的答案。然而，与之前试图解决这个问题的许多教授不同，两人并非历史学家，不以研究早期美国历史的学术工作闻

① 1804 年，汉密尔顿在与政敌杰斐逊的副总统阿伦·伯尔决斗时死去。

名，甚至从未发表过一篇关于历史人物的论文。莫斯特勒和华莱士都是统计学家。

莫斯特勒最为人瞩目的一篇论文是关于“美国职业棒球大联盟总冠军赛”的，他在论文中提出问题：从统计学的角度来看，七场比赛是否能够决出最好的棒球队。在着手研究 12 篇文章著作权的前几年，华莱士也发表过一篇论文，题目是《T 分布和卡方分布的正态近似界限》。听上去很难以置信吧？有人想用概率方程解决历史难题，1963 年的历史学教授大概会认为这是一派胡言。

莫斯特勒和华莱士所用的方法与政治或意识形态无关，他们只是首批利用词频和概率展开研究的统计学家。

他们解决问题的过程在某些方面较为复杂，比如采用了含有阶乘的方程、指数、求和、对数以及 T 分布，但核心方法却简单得惊人：

- 根据确定是汉密尔顿或麦迪逊所写的文章，分别统计某些常用词出现的频率。
- 在需要进行研究的文章里统计相同词语出现的频率。
- 通过比较上述两个频率，确定争议文章的作者。

事后回头看，即便不使用那些玄妙的概率方程，两位统计学家的研究结果似乎也是显而易见的。《联邦党人文集》里麦迪逊的文章中，超过一半使用了“whilst”这个词，但从未用过“while”。相反，汉密尔顿大约三分之一的文章中使用了“while”，但从未用过“whilst”。

莫斯特勒和华莱士并不是只依靠一个词的分析，从统计学上来讲，那样做是不充分的。他们先系统地甄选出几十个基本单词，然后在有争议的文章中统计每个词的使用频率。许多词没有任何政治含义，两位不同作者的使用频率竟然出现明显的不同。比如，麦迪逊用“also”这个词的频率是汉密尔顿的 2 倍，而汉密尔顿使用“according”的频率则比麦迪逊高很多。

莫斯特勒和华莱士采用的方法具有可证伪性[①]。研究结果表明，如果在已知作者身份的文章中使用相同的方法，他们可以准确无误地识别作者。而对于那些有争议的文章，他们得出结论：麦迪逊是 12 篇文章的实际作者[②]。

在总结研究结果时，也许担心惹恼一代又一代苦恼不已的历史学家，两位数学家的立论和措辞十分谨慎，但展示的数字却毫不含糊，两人对自己的统计方法有十足的信心。所有已知作者身份的文章的测试分析都毫无瑕疵，作者未明的文章也与其一致。由此得出最终结论，汉密尔顿所言为虚，那 12 篇文章的作者并不是他。

---

① 指依照某种逻辑或原则得出的结论可以被证明是错的，但可证伪性并不等于已经证伪。正如著名科学哲学家卡尔·波普尔所说的：“所有科学命题都要有可证伪性。”

② “在已知作者身份的文章中准确无误地识别作者”，是指将已知作者身份的文章分成两部分，用一部分进行词频等数据的统计分析，然后用这些统计指标检验另外一部分文章，以此检验结果；“对于那些有争议的文章得出的结论”，是指用上述统计指标来确定未知文章的作者。在统计学中，前一种方法为“样本内预测”（in-sampel test），后一种为“样本外预测”（out-of-sample test）。

经过无数统计和非统计的研究后，莫斯特勒和华莱士的分析结果（麦迪逊是作者）已经成为目前统计学家和历史学家们的共识。他们超前于所处的时代，所做的研究虽然涉及一些复杂公式，但本质上依靠的还是统计分析。如果是今天，通过计算机统计单词和频率是件简单的小事，但在 1963 年，情形却并非如此。

当时统计单词是靠手工完成的。比如，要找出每一篇文章中“upon”出现的次数，他们得一页页、一个个地找出来。为了感受和理解莫斯特勒和华莱士（至少是他们的研究助理）都经历了什么，我打印了一本完整的《联邦党人文集》，开始数“upon”这个词出现的次数。30 分钟后，我只进展到全文的八分之一，在大约 40 页里有 37 个 upon。没过多久，我便眼皮狂跳，脑子发木——upon 在哪里？这种痛苦就像在漫漫人海中寻找某张人脸。

活在 1963 年实在有些辛苦，最后我放弃了，转而采用 21 世纪的技术进行计数：我打开谷歌，搜索“联邦党人文集完整文本”，点进第一个搜索结果进行下载，再用 Microsoft Word 打开文件。两分钟后，我选定部分内容，再使用菜单里“查找”命令，随后发现“upon”出现了 46 次。借助电脑后，不仅在速度上快了 28 分钟，而且结果远比疲惫的肉眼来得准确。

再找一个词语结果也还是一样，一个人浏览一遍《联邦党人文集》全文的时间在 4 小时左右，电脑所需时间几乎可以忽略不计。不管是莎士比亚文集、《圣经》《白鲸》，还是英语

文学集，对当时的莫斯特勒和华莱士来说，进行类似的分析都是无法想象的难题。现在情况就完全不一样了，在电脑上统计某个单词在大部头文本里出现的次数，绝大多数十来岁的青少年皆可轻松完成。

在莫斯特勒和华莱士公布研究结果后的 50 年间，计算机辅助文本处理发展迅速。谷歌在其搜索结果中运用文本分析，以此决定对哪些用户投放哪些广告。目前还有研究人员试图用文本分析进行判断，是什么原因让一条 Twitter 像病毒一样传播。媒体也经常对同类型的内容进行措辞上的细微调整，以期实现页面浏览量的最大化。但是到目前为止，这些科技公司对文本分析的应用还比较单一，它还有更大的可能性。

虽然莫斯特勒和华莱士只是用统计方法研究了一个著作权问题，但实验获得的成功却产生了深远的影响。作家们确实有各自的风格，而且是可以进行预测的。事实证明，留下个人风格印记的不仅仅是 18 世纪的政客，所有书籍的作者——无论广受欢迎、远近闻名，还是名声不显、遭受恶评——都在数十年的写作中不断重复自己的遣词造句和行文方式。

莫斯泰勒和华莱士提出的问题和做出的解答虽有一定局限性，但文本分析确实可以回答各种各样的问题，那些让一代代作家和读者感到疑惑的问题：相比其他作家，海明威真的更少使用副词吗？书籍的阅读难易程度对其受欢迎程度有什么样的影响？男性和女性作家的写作方式有何不同？作家提出的创作建议有用吗？他们自己会遵循那些建议吗？除了一些明

显不同的拼法问题，还有什么方法可以用来区分美国小说家和英国小说家？从纳博科夫到 E.L. 詹姆斯（E. L. James）[①]，我们喜欢的作家喜欢用的词是什么？

虽然学术界已经开始研究成功作家的写作模式，但仍有许多问题有待探索。对普通读者、主修文学的大学生以及野心勃勃的作家来说，这些问题既有趣又有用。你可能不关心泊松分布[②]，也不在乎解读语言的程序，但你也许想知道自己最喜爱的作家是如何写作的，以及这对你来说可能意味着什么。

用数据分析来研究写作不仅妙趣横生，还能提供丰富的信息，有时也会非常搞笑。此外，我们也能借此了解平时阅读的作家，思考我们自己写作时使用的词句，这一切正是本书要深入探讨的。在这本书里，每一章都专注于一个文学新问题的研究。

这些研究并不会复杂到令人痛苦的地步。实际上，只要真正有价值，研究无须也不应那般复杂。关于经典文学或现代畅销书的许多有趣问题是可以通过统计的透视镜来观察的，但针对这些问题的统计分析尚未形成体系。本书将用一种崭新的方法来攻克这些简单而独特的问题。这是一本关于“文字”的书，但却是用“数字”写成的。

---

① E.L. 詹姆斯（E.L.James），英国畅销书女作家，小说《五十度灰》（*Fifty Shades of Grey*）的作者。

② 泊松分布（Poisson Distribution），统计与概率学里常见的离散概率分布。

目录 CONTENTS

Part 4

Use Sparingly

简洁“地”用词

通往地狱之路铺满了副词。

——斯蒂芬·金

在有关文学的传说中，史上最好的故事之一只有区区六个单词：For sale: baby shoes, never worn（售：婴儿鞋，全新）。这是“少即是多”的最佳范例，人们经常将其归功于海明威。

这几个单词是否真为海明威所写已无从考证，但有一点可以确定——这个故事写于1991年。作家和读者都愿意相信它是海明威这位诺贝尔奖获得者写的。这也正常，因为海明威一直以文辞简洁而闻名，最起码这部最短的短篇小说与他的风格很像。

简洁风格是海明威有意的选择。他曾在给编辑的一封信中写道：“葛底斯堡演说[①]如此之短，实非偶然。写作和飞行、数学、物理学一样有章可循，遵循的法则不可动摇。”海明威坚信，作品应尽可能精简，只留最核心的部分，多余的文辞只会损害作品。

持此信念的并非只有海明威一人。高中课堂上、各种各样的写作指南里都能接触到同样的观点。任何一个人（只要他的英语老师要求严格）都知道副词是违反简洁原则的罪魁祸首。

① 美国南北战争期间，宾夕法尼亚州的葛底斯堡小镇发生了一场残酷战役。林肯在战后发表了《葛底斯堡演说》，以纪念战死的将士。