人类语言技术综合讲座

格雷姆·赫斯特　系列主编

社交媒体自然语言处理//第二版//

Natural Language Processing for Social Media（Second Edition）

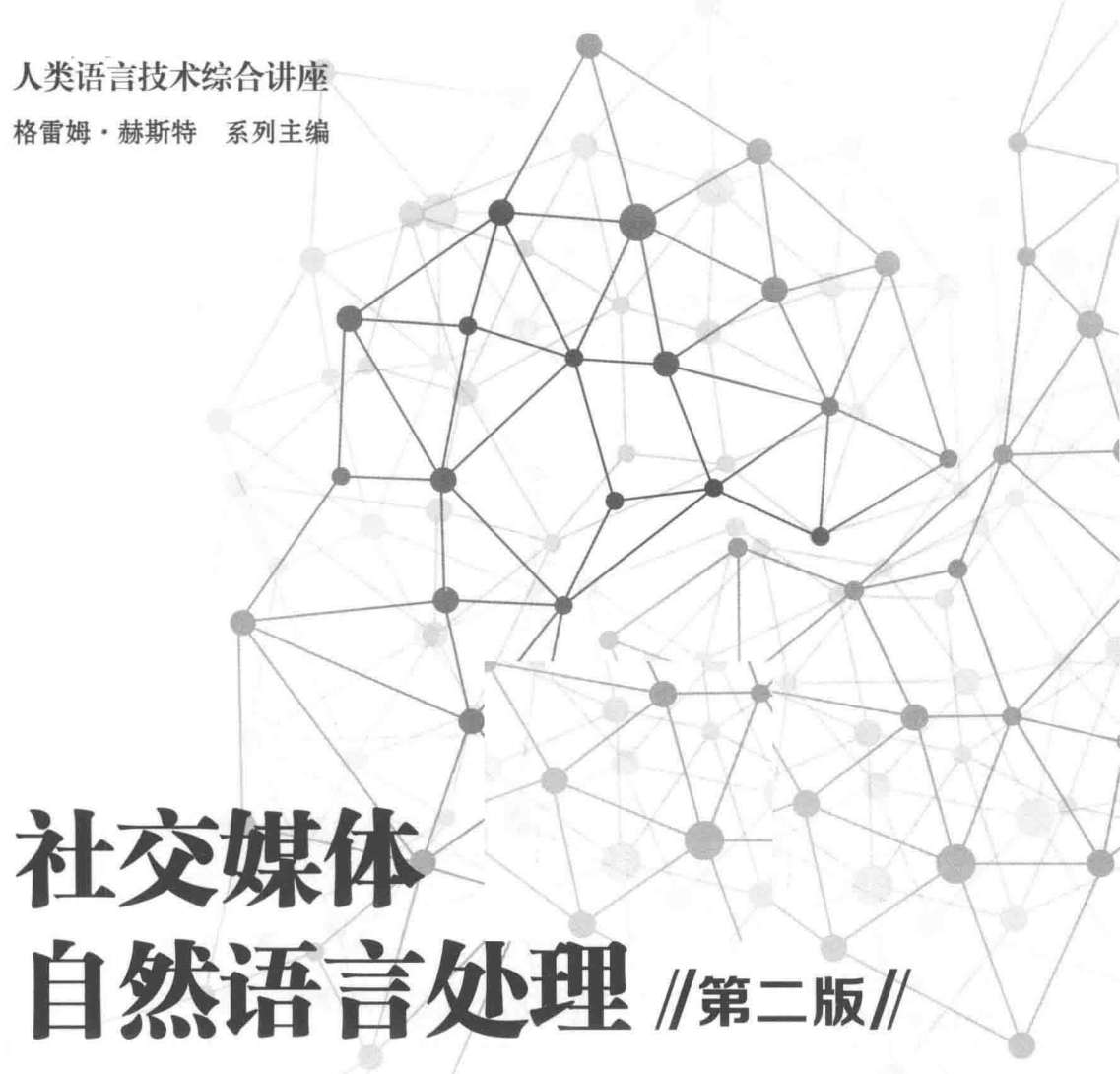
[加] 阿塔夫·法辛达　[加] 戴安娜·英克彭　著

许舟军　焦程波　译

中国宇航出版社

人类语言技术综合讲座

格雷姆·赫斯特　系列主编



社交媒体自然语言处理 //第二版//

Natural Language Processing for Social Media（Second Edition）

[加] 阿塔夫·法辛达　[加] 戴安娜·英克彭　著

许舟军　焦程波　译

中国宇航出版社

·北京·

图书在版编目（CIP）数据

社交媒体自然语言处理 : 第二版 / (加) 阿塔夫·法辛达, (加) 戴安娜·英克彭著 ; 许舟军, 焦程波译. -- 北京 : 中国宇航出版社, 2019.1
书名原文: Natural Language Processing for Social Media, Second Edition
ISBN 978-7-5159-1541-8

Ⅰ. ①社… Ⅱ. ①阿… ②戴… ③许… ④焦… Ⅲ. ①互联网络－传播媒介－自然语言处理－研究 Ⅳ. ①TP391

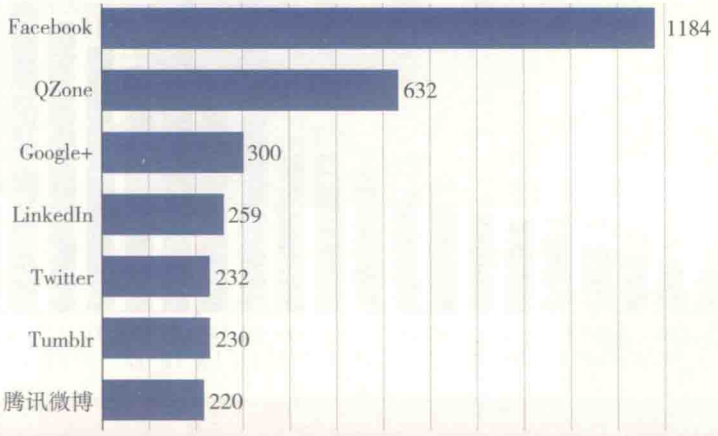
中国版本图书馆CIP数据核字(2018)第245025号

策划编辑 田芳卿
责任编辑 吴媛媛　　**装帧设计** 宇星文化

出　版 发　行	**中国宇航出版社**		
社　址	北京市阜成路8号	邮　编	100830
	(010)60286808		(010)68768548
网　址	www.caphbook.com		
经　销	新华书店		
发行部	(010)60286888		(010)68371900
	(010)60286887		(010)60286804(传真)
零售店	读者服务部		
	(010)68371105		
承　印	三河市君旺印务有限公司		
版　次	2019年1月第1版		2019年1月第1次印刷
规　格	710×1000	开　本	1/16
印　张	16.5　　彩　插　8面	字　数	185千字
书　号	ISBN 978-7-5159-1541-8		
定　价	68.00元		

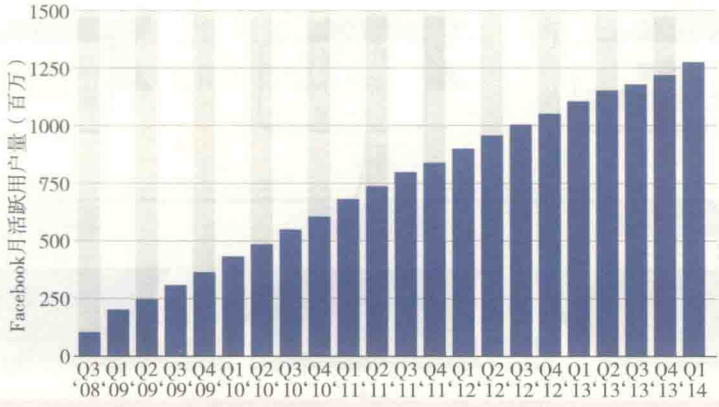
精彩图例

统计门户网站 Statista 公布了 2014 年 1 月基于活跃用户数量的社交网络排名。



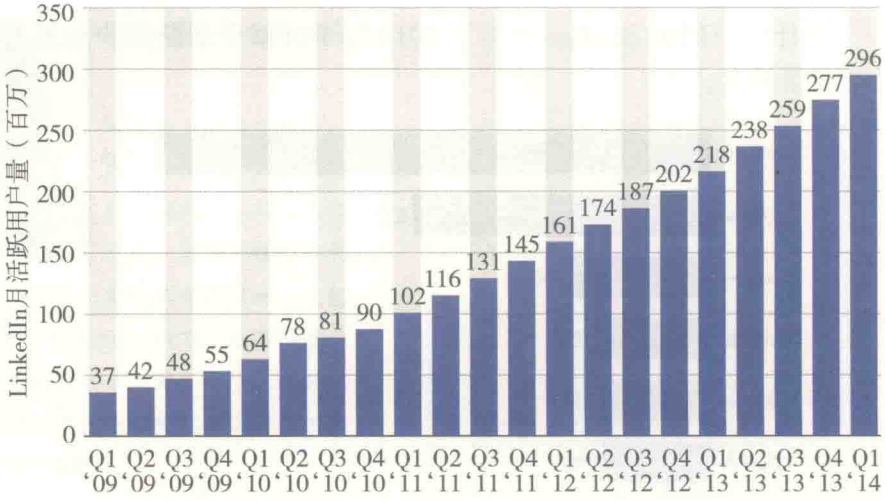
统计门户网站 Statista 提供的 2014 年 1 月社交网站活跃用户数量（百万）排名

Facebook 用户在 2011 年年底达到 8.45 亿，2013 年年底累计达到 12.28 亿。



统计门户网站 Statista 提供的 2008 年第三季度到 2014 年第一季度 Facebook 月活跃用户量（百万）

LinkedIn 用户在 2013 年年底也达到了 2.77 亿，它在 2011 年年底只有 1.45 亿用户。

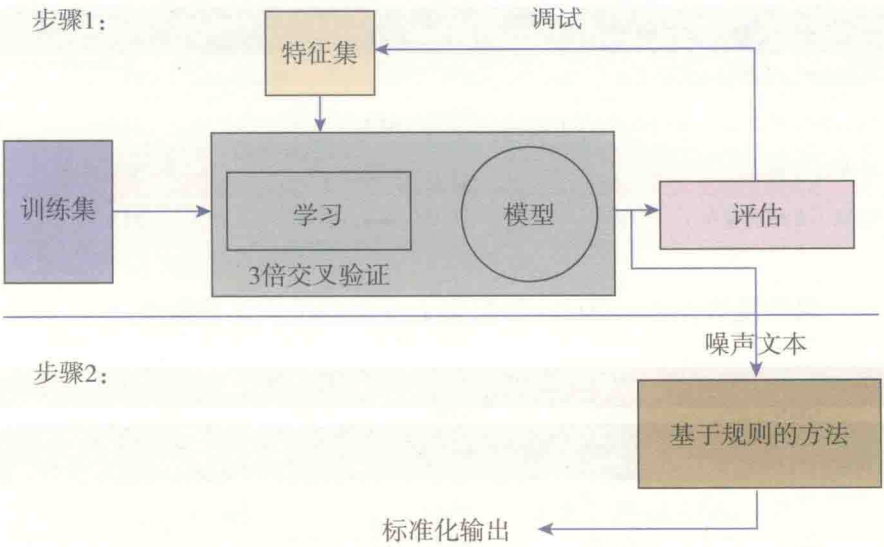


统计门户网站 Statista 提供的 2009 年第三季度到 2014 年第一季度 LinkedIn 月活跃用户量（百万）

社交媒体语义分析（SASM）分析数据的过程可由数据可视化方法来完成。

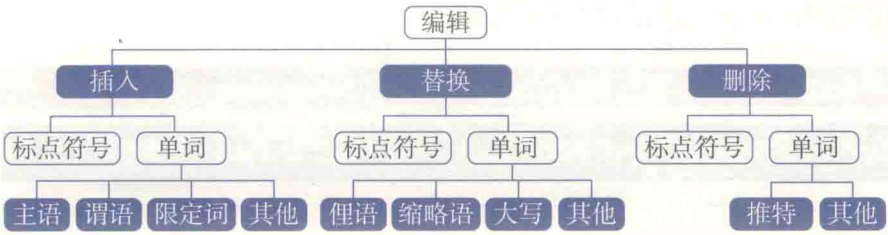
应用自然语言处理工具将数据转换成情报的社交媒体语义分析框架

Akhtar et al.[2015] 提出了一种将推文标准化的混合方法，该方法分两个步骤进行。研究人员训练了一个监督学习模型，使用 3 倍交叉验证来确定最佳特征集。



推文标准化的方法。水平线将两个步骤分隔开来（检测将要进行标准化的文本并应用标准化规则）[Akhtar et al., 2015]

Baldwin and Li [2015] 设计了一种使用标准化编辑的分类法，该方法按照 3 个粒度级别对编辑进行分类，其结果表明：该分类法的针对性应用是标准化的有效方法。



标准化编辑的分类法 [Baldwin and Li, 2015]

Liu and Inkpen [2015] 的 DeepNN 模型给出了最好的结果。我们很惊喜地发现基于 SVM 和朴素贝叶斯的简单模型性能很好。

Eisenstein 数据集上用户位置检测分类准确率 [Liu and Inkpen, 2015]

模型	准确率（%）（4 个区域）	准确率（%）（49 个州）
Geo topic 模型 [Eisenstein et al., 2010]	58.0	24.0
DeepNN 模型 [Liu and Inkpen, 2015]	**61.1**	**34.8**
朴素贝叶斯	54.8	30.1
SVM（支持向量机）	56.4	27.5

四种模型在 Eisenstein 数据集上预测的平均误差距离。

Eisenstein 数据集上预测的平均误差距离 [Liu and Inkpen, 2015]

模型	平均误差距离（km）
[Liu and Inkpen, 2015]	**855.9**
[Priedhorsky et al., 2014]	870.0
[Roller et al., 2012]	897.0
[Eisenstein et al., 2010]	900.0

Han et al. [2014] 的模型包含广泛的特性工程，比其他模型的性能更好。DeepNN 模型尽管有计算限制，但是使用了少量特征，也比 Roller et al. [2012] 的结果更好。

Roller 数据集上的用户位置预测结果 [Liu and Inkpen, 2015]

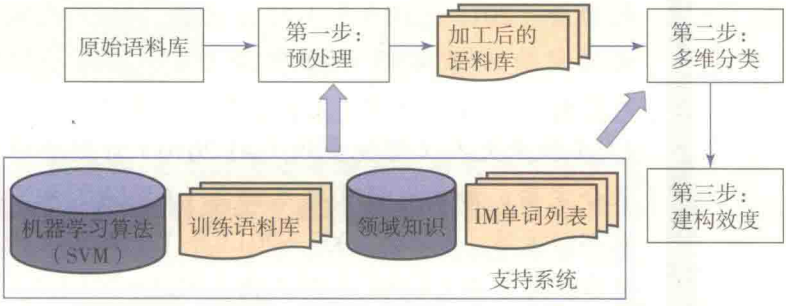
模型	平均误差（km）	中值误差（km）	准确率（%）
[Roller et al., 2012]	860	463	34.6
[Han et al., 2014]	–	260	45.0
[Liu and Inkpen, 2015]	733	377	24.2

下表展示了 Aman 数据集的一个子集上每个分类的详细结果（至少包含一个明确情感单词的句子）。

Ghazi et al. [2014] 的情绪分类结果

		精确率	召回率	F-measure
SVM+词袋 准确率 50.72%	幸福	0.59	0.67	0.63
	悲伤	0.38	0.45	0.41
	生气	0.40	0.31	0.35
	惊喜	0.41	**0.33**	0.37
	厌恶	0.51	0.43	0.47
	恐惧	0.55	0.50	0.52
	非情绪	0.49	0.48	0.48
SVM+其他特征 准确率 **58.88%**	幸福	**0.68**	**0.78**	**0.73**
	悲伤	**0.49**	**0.58**	**0.53**
	生气	**0.66**	**0.48**	**0.56**
	惊喜	**0.61**	0.31	**0.41**
	厌恶	**0.43**	**0.38**	**0.40**
	恐惧	**0.67**	**0.63**	**0.65**
	非情绪	**0.51**	**0.53**	**0.52**

SVM 分类器能够以 70% ~ 75%的准确度（分类过程是一系列 5 个不同的二元分类器）预测句子所属的 IM 维度。



基于 SVM 的用于印象管理（Impression Management）的文本挖掘程序 [Schniederjans et al., 2013]

作者简介

阿塔夫·法辛达

阿塔夫·法辛达博士（安娜）是南加利福尼亚大学数据科学研究院（DSI）的研究助理，也是南加利福尼亚大学维特比工程学院计算机科学系的教师。曾获得蒙特利尔大学计算机科学专业博士学位，2005 年获得巴黎索邦大学博士学位，主要研究方向为自动法律文件摘要。

法辛达博士是自然语言处理科技公司的创始人兼 CEO，专门从事自然语言处理、法律决策摘要、机器翻译和社交媒体分析。她曾担任加拿大人工智能协会（2013—2015 年）行业主席；2013 年加拿大里贾纳 AI/GI/CRV 大会联合主席；2014 年加拿大蒙特利尔 AI/GI/CRV 大会主席；加拿大渥太华第 23 届加拿大人工智能会议（AI 2010）计划委员会联合主席；加拿大语言技术协会（AILIA）语言技术部门主席（AILIA 2009—2013 年）；加拿大语言技术研究中心（LTRC）副总裁（2012—2014 年）；加拿大自然科学与工程研究理事会（NSERC）、计算机科学联络委员会（自 2014 年起）和国际化标

准组织（ISO）加拿大咨询委员会的成员。她还是加拿大蒙特利尔大学的兼职教授（2009—2015 年），蒙特利尔理工学院工程学院讲师（2012—2014 年），英国胡弗汉顿大学计算语言学研究组客座教授和荣誉研究员（2010—2012 年）。

法辛达博士还参与艺术活动，在创新技术和信息与通信技术类别中赢得了主题为“以平衡的生活方式取得成功”的 Femmessor-Montréal 比赛。她的绘画作品发表在图书《一千零一夜》（2014 年版）中。她发表了 50 多篇会议和期刊论文，撰写了 3 本图书，IGI 出版社出版的 *Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding* 一书中名为“文件摘要中的社交网络整合”一章也是由她撰写的。

戴安娜·英克彭

戴安娜·英克彭博士是加拿大渥太华大学电气工程和计算机科学学院教授，2003 年在多伦多大学计算机科学系获得博士学位，1994 年在罗马尼亚克卢日 - 纳波卡科技大学计算机科学系获得工程学学士学位，次年获得硕士学位。她的研究兴趣和专长是自然语言处理和人工智能，特别是将词

汇语义学应用于近义词和细微差别词、单词和文本相似性、基于情绪和情感的文本分类、自然语音的信息检索、信息提取，以及从社交媒体中检测精神健康问题的迹象。

英克彭博士是第29届佛罗里达人工智能研究学会会议（FLAIRS 2016，佛罗里达州基拉戈，2016年5月）、第28届加拿大人工智能会议（AI 2015，新斯科舍哈利法克斯，2015年6月）以及信息管理和大数据国际研讨会（SimBig 2015，秘鲁库斯科，2015年9月）的受邀讲员。英克彭博士在第25届加拿大人工智能会议（AI 2012，加拿大多伦多，2012年5月）、关于自然语言处理和知识工程的第七届IEEE国际会议（IEEE自然语言处理-KE'11，日本德岛，2011年11月）和关于自然语言处理和知识工程的第六届IEEE国际会议（IEEE自然语言处理-KE'10，中国北京，2010年8月）担任计划委员会联合主席。从2010年9月到2013年8月，她被任命为英国胡弗汉顿大学的计算语言学客座教授。

她曾经带领并将继续带领许多科研项目，这些项目获得来自自然科学和工程研究理事会（NSETC）、加拿大社会科学和人文研究委员会（SSHRC）和安大略卓越中心（OCE）的资助。这些项目也包括和来自渥太华、多伦多和蒙特利尔地区的公司的工业合作。她发表了超过30篇期刊论文、100篇会议论文，并为9本书籍撰写过章节。她是其研究领域很多会议的计划委员会成员，很多期刊的审稿人，并且是《计算智能》和《自然语言工程》期刊的副主编。

这些工作由计算语言学协会（如SemEval任务）或国家标准与技术研究院文本检索会议（TREC）和文本分析会议（TAC）组织开展。在最后一章，我们讨论了自然语言处理这个快速发展的学科的重要性以及未来10年移动技术、云计算、虚拟现实、社交网络不断变化背景下自然语言处理技术（NLP）的巨大潜力。

在第二版中，我们增加了第一版提及的工作、应用的最新进展情况，对新方法及成果也进行了讨论。随着社交媒体数据规模和自动处理需求不断增加，使用社交媒体数据的研究项目和出版物数量持续增长。在第一版300多条引用参考的基础上，第二版增加了85条新的引用参考。除了更新每个章节，我们在4.5节“媒体监测”部分添加了一个新的应用（数字营销）。同时增加了医疗保健应用部分，该部分内容延伸讨论了通过社交媒体检测精神病体征的最新研究进展。

关键词

社交媒体、社交网络、自然语言处理、社交计算、大数据、语义分析

献给我的丈夫马苏德（Massoud）和我的女儿蒂娜（Tina）、阿曼达（Amanda），我的女儿们是妈妈所期望的最好的孩子：她们快乐、可爱、充满乐趣。

——阿塔夫·法辛达（Atefeh Farzindar）

献给我的丈夫尼库（Nicu），在他的陪伴下，我可以翻越任何高山，还要将此书献给我们的宝贝女儿尼科莱塔（Nicoleta）。

——黛安娜·英克彭（Diana Inkpen）

前言

本书介绍了自然语言处理（NLP）在社交媒体数据语义分析上的最新理论和实证研究。随着该领域的持续发展，第二版针对第一版提及的任务和应用的方法及结果增加了最新信息。

在过去的几年中，在线社交网站给个人、团体、社区之间的交流途径带来了革命性的变化，同时改变了人们的日常习惯。用户生成的空前规模的多样化信息，以及用户之间的交互网络，为理解社交行为、构建社会智能系统提供了新的机会。

很多社交网络、社交网络挖掘研究都是基于图论展开的。这种思路是合理的，因为社交结构是由社交参与者集合、社交参与者之间的二元关系组合组成。我们认为，面向社交网络的结构信息扩散图挖掘方法或社交网络影响力传播图挖掘方法，需要与社交媒体内容分析结合使用。这为使用社交互动产生的可获取的公开信息的新的应用提供了机会。应用改进的传统自然语言处理方法，可以部分解决主要针对社交媒体发布消息的内容分析问题。当我们收到一个少于10个字符(包

含表情和心情符号）的文本，我们可以理解甚至回应。虽然自然语言处理方法不能处理此类文本，但社交媒体数据存在逻辑信息，基于这种逻辑信息，两个人才能沟通。同样的逻辑在世界上占据主导地位，全人类可以使用它与其他人共享和交流信息。这是自然语言处理面临的一种新的挑战性语言。

我们相信需要新理论、算法开展社交媒体数据语义分析，同时需要一种新的大数据处理方法。本书提及的语义分析是指可能与社交网络结构相结合的、进行了语义增强的社交媒体信息语言处理。事实上，我们使用这个术语在一个更广义层面来表示能进行社交媒体文本和元数据智能处理的应用。一些应用可以访问超大规模的数据。为此，算法需要调整以适应数据的在线处理，不必非以存储所有数据（再处理）的形式处理数据（大数据）。

这种情况促使我们提出两个教程：EMNLP 2015[①] 大会上的《社交媒体文本分析应用》和第29届加拿大人工智能会议(AI 2016）上的《社交媒体自然语言处理》[②]。我们还组织了多个主题研讨会：社交网络中的语义分析（SASM 2012）[③]、社交媒体中的语言分析（LASM 2013[④]、LASM 2014[⑤]）以及计算语

① http://www.emnlp2015.org/tutorials/3/3_OptionalAttachment.pdf
https://www.cs.cmu.edu/~ark/EMNLP-2015/proceedings/EMNLP-Tutorials/pdf/EMNLP-Tutorials06.pdf

② http://aigicrv.org/2016/

③ https://aclweb.org/anthology/W/W12/#2100

④ https://aclweb.org/anthology/W/W13/#1100

⑤ https://aclweb.org/anthology/W/W14/#1300

言学协会（如 ACL、EACL 和 NAACL-HLT）[①] 组织的会议。

我们的目标是广泛呈现语言分析研究及其成果，为自然语言处理、计算语言学、社会语言学、心理语言学等领域提供参考。我们的研讨会邀请所有与社交媒体语言分析相关的原创研究参与，包括以下主题：

- 人们在社交媒体上讨论什么？
- 他们如何表达自己？
- 他们为什么在社交媒体上发布内容？
- 语言和社交网络属性如何相互作用？
- 面向社交媒体分析的自然语言处理技术。
- 辅助理解社交数据的语义 Web / 本体 / 域模型。
- 通过语言分析来表征参与者。
- 语言、社交媒体和人类行为。

还有其他几个相关的主题研讨会，例如与 2012—2016 年国际万维网大会合作的理解微博（#Microposts）[②] 系列研讨会。这些研讨会特别侧重于易发布的非正式短文本（如推文、脸书共享信息、Instagram 类型共享信息、Google+信息）。另外还有自 2013 年开始举办的社交媒体自然语言处理系列研讨会（SocialNLP），包括与 EACL 2017[③] 合作举办的 SocialNLP 2017 以及 IEEE BigData 2017。[④]

---

① http://www.aclweb.org/

② http://microposts2016.seas.upenn.edu/

③ http://eacl2017.org/

④ http://cci.drexel.edu/bigdata/bigdata2017/

本书的目标读者是对开发自动化社交媒体文本分析工具和应用感兴趣的研究者。我们假定读者拥有自然语言处理和机器学习的基础知识，希望本书能帮助读者更好地理解计算语言学和社交媒体分析，特别是文本挖掘技术和专为社交媒体文本设计的自然语言处理应用，如摘要、地点检测、情感和情绪分析、话题检测和机器翻译。

阿塔夫·法辛达

戴安娜·英克彭

2017 年 12 月