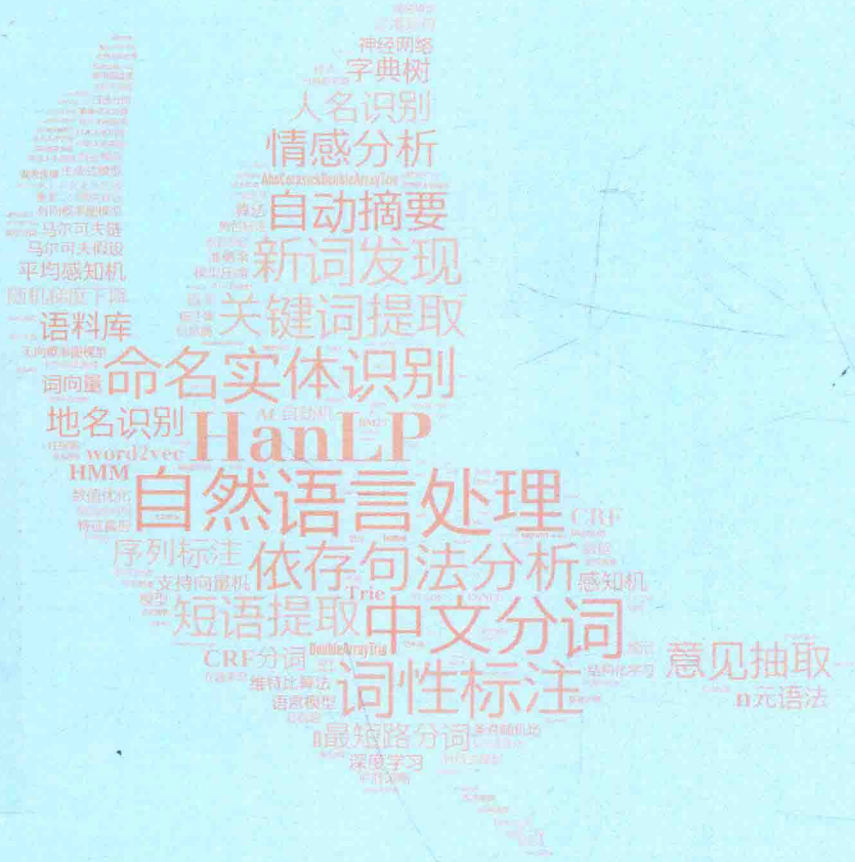
TURING 图灵原创

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

自然语言处理入门



何晗（@hankcs）著

中国工信出版集团
人民邮电出版社 POSTS & TELECOM PRESS

TURING 图灵原创

自然语言处理入门

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

何晗（@hankcs）著

人民邮电出版社
北京

图书在版编目（CIP）数据

自然语言处理入门 / 何晗著. -- 北京 : 人民邮电出版社, 2019.10
（图灵原创）
ISBN 978-7-115-51976-4

Ⅰ. ①自… Ⅱ. ①何… Ⅲ. ①自然语言处理－研究 Ⅳ. ①TP391

中国版本图书馆CIP数据核字(2019)第193305号

内 容 提 要

这是一本自然语言处理入门书。

HanLP作者何晗汇集多年经验，从基本概念出发，逐步介绍中文分词、词性标注、命名实体识别、信息抽取、文本聚类、文本分类、句法分析这几个热门问题的算法原理与工程实现。书中通过对多种算法的讲解，比较了它们的优缺点和适用场景，同时详细演示生产级成熟代码，助读者真正将自然语言处理应用在生产环境中。

随着对本书内容的学习，你将从普通程序员晋级为机器学习工程师，最后进化到自然语言处理工程师。

◆著　　何　晗
责任编辑　王军花
责任印制　周昇亮
◆人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷
◆开本：800×1000　1/16
印张：24
字数：472千字　　2019年10月第1版
印数：1-4 000册　　2019年10月北京第1次印刷

定价：99.00元

读者服务热线：(010)51095183转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字20170147号

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

推荐序

自然语言处理（NLP）的目标是使计算机能够像人类一样理解语言。人类语言是一个复杂的符号系统，人们可以通过不同方式传达信息，比如文字、语音、手势、信号等，而所传达的信息也可能因为用词或语调的微妙不同而大相径庭。完全通过机器来理解人类语言目前还是一个很困难的任务。所幸的是近年来自然语言处理作为一门学科发展迅速，得到了越来越广泛的应用。在使用神经网络技术之前，NLP 的研究经历了从规则到统计的过程，而图像、语音、文本是信息记载的不同载体，这些正是深度学习（Deep Learning）的运用范围，目前深度学习在 NLP 中也取得了很好的结果。

NLP 发展迅速，进入这个领域的初学者也越来越多。这个领域所需要的知识比较繁杂，掌握难度较大，因此，大家对于阅读相对轻松的入门资料是有很大需求的，而这在 NLP 领域是个缺口。

何晗所著的《自然语言处理入门》是汉语自然语言处理方面实用性很强的一本入门新书，涉及 NLP 的语言理论、算法介绍和工程实践等。书中着重介绍了中文自然语言处理的传统统计方法，也涉及最新发展的深度学习方法；此外，还分享了很多一线的工业级开发经验、工程实现和技巧。

特别值得一提的是，何晗开发了中文分词库 HanLP。在 GitHub 上，HanLP 全球用户量在 2017 年 10 月就超过了斯坦福大学的 CoreNLP，以及老牌自然语言处理开发包 NLTK。目前，HanLP 的受欢迎程度持续增长，已经成为 GitHub Star 数最高的自然语言处理工具包。2019 年，在中国国际软件博览会上，HanLP 获得了优秀产品奖。

回到图书本身，可以说，这是第一本把读者阅读体验放在首位的中文 NLP 图书。著名物理学家霍金说，每增加一个公式，读者就少了一半。我猜何晗得到了霍金的“真传”。这本书的特点就是只允许必不可少的公式出现，采用从问题到算法再到工程实现的写作思路，通俗易懂、容易上手。何晗甚至设定了一个小目标：让大家在地铁上也能学会 NLP 开发。

最后，再次将这本优秀务实的中文 NLP 入门书分享给你。彻底搞懂本书后，你可以成长为自然语言处理类库的设计者。

夏志宏

首批长江学者，千人计划专家，

美国“青年科学家与工程师总统奖”得主，布拉门塞尔纯数学奖得主，

南方科技大学数学系创系主任，美国西北大学终身讲席教授，

大快搜索首席数学家

推荐语

最近几年 NLP 的研究进入高潮。很多人都想学习 NLP 但是不知道如何开始，目前国内 NLP 领域急需更多入门好书，HanLP 作者何晗即将出版的这本《自然语言处理入门》值得一读。这本书比较系统地介绍了 NLP 的基础技术，深入浅出、容易理解，对初学者很有帮助。

——周明，微软亚洲研究院副院长，国际计算语言学会会长

自然语言处理是人工智能最核心也最具挑战的领域，我衷心希望有更多的人能加入这个领域的技术研究、开发、应用之中。相信何晗的这本《自然语言处理入门》会对大家有很大的帮助。本书以 GitHub 开源项目 HanLP 的代码实现为基础，介绍了从分词到句法分析再到深度学习的自然语言处理最基本的技术。本书叙述简洁清晰，讲解透彻深入，非常适合初学者。强烈推荐！

——李航，字节跳动人工智能实验室总监，《统计学习方法》作者

作者从实践的角度用通俗易懂的语言解释自然语言处理的概念，用应用实例和类程序语言描述算法，有鲜明的特色和很强的实用性，我相信这本书会深受读者的欢迎。

——宗成庆，中国科学院自动化研究所研究员、博士生导师，《统计自然语言处理》作者

本书作者何晗原来也是一个自然语言的爱好者，现在已成为自然语言处理的专业人士，美国埃默里大学计算机科学专业的博士生。他自主开发了一套完全开放源代码的自然语言处理工具包 HanLP，受到使用者的好评。这本书依托于 HanLP 工具包，从基本的概念和原理出发，讲解了自然语言处理中一些常用的问题和算法。我相信这本书融入了作者对这个领域各项技术的深刻理解和切身体会，一定会是一本非常好的入门书。

——刘群，华为诺亚方舟实验室语音语义首席科学家

这本书不仅介绍了 NLP 的任务及算法，也提供了可以实际运行的生产级代码，非常适合 NLP 初学者入门并快速布置到生产环境。本书的文字十分流畅，连标点符号都鲜有错误，展示了作者严谨的写作态度和极强的文字能力。虽然本书深度学习相关的篇幅不多，但是了解传统的 NLP 方法能够大大提升对问题的理解能力，推荐阅读！

——王斌，小米人工智能实验室主任、NLP 首席科学家

近年来人工智能技术应用日益广泛深入，自然语言处理（NLP）也随之成为一门“显学”。作为教计算机学习理解和使用人类语言的学科，NLP 在搜索引擎、推荐系统、社会计算、智能音箱、机器翻译等几乎所有与“语言”有关的方向发挥着重要作用。由于人类语言的复杂特点，NLP 所涉及的基础知识和技术非常多，虽然国内外有一些经典的教材，但与实际应用密切结合深入浅出讲授的著作凤毛麟角。本书作者是著名的中文 NLP 工具包 HanLP 的开发者，本书结合 HanLP 细致讲解 NLP 的关键技术，是上手 NLP 的优秀读物。我非常高兴将这本书推荐给对 NLP 感兴趣的朋友们。

——刘知远，清华大学副教授，MIT“35 岁以下科技创新 35 人”中国区榜单获得者

大数据与人工智能已经成为当今世界各国的战略必争之地，自然语言处理是人工智能科学皇冠上的明珠，大数据为自然语言处理的跨越式发展提供了算源与算力基础。HanLP 吸收了我所开源的汉语分词系统 ICTCLAS 的精髓，何晗跟我深入讨论过我发表的论文，其学习能力与勤奋严谨给我留下了深刻印象。何晗结合 HanLP 宝贵的开发经验与 NLP 领域最新研究成果所写的这本书，是一部难得的 NLP 启蒙之作，推荐阅读。

——张华平，北京理工大学副教授、NLPIR-ICTCLAS 创始人，
钱伟长中文信息处理科学技术奖一等奖获得者

几年前，第一次得知 HanLP 的作者何晗是上外一名非科班同学时，我很吃惊。要知道，即使科班出身，要开发一个如此完备的 NLP 工具都相当有挑战，更不用说 HanLP 在中文 NLP 开源领域还相当成功了。而今，何晗在美国就读 CS 领域的博士，他在课余时间坚持写作，结合自己的学习历程和 HanLP 的开发经验给大家呈现了一本不太一样的 NLP 入门书。我很乐意把这本书推荐给大家。

——杨攀，我爱自然语言处理（52nlp）博主，公众号 AINLP 主理人

前　言

为什么要写这本书

自然语言处理是一门交叉学科，属于人工智能的一个分支，涉及计算机科学、语言学、数学等多个领域的专业知识。外行人很难入门这个小众的圈子，非科班出身的我对此深有体会。经典教材虽然高屋建瓴，但自学的话很难读懂，缺乏代码也无法落地；工程类书籍则往往侧重对开源项目的接口介绍，缺乏深度与宏观系统性。我曾经跟天书般的术语与公式顽强斗争，也在迷宫般的教学代码中苦苦挣扎。现在回顾自学历程，我才认识到：当时缺少一本面向普通人的入门书，走了许多弯路。

在我的开源自然语言处理项目 HanLP 流行起来后，我接触了大量 NLP 初学者，我看到不少人碰到了我当初苦苦思索的问题。许多用户不理解“统计自然语言处理”的设计理念，对“语料”“训练”“模型”等概念十分陌生。同时，如果你缺乏自然语言处理基础的话，也无法掌握 HanLP 中的高级功能。还有部分学习热情高涨的用户尝试阅读 HanLP 的代码，却反应即便代码有注释，也看不懂为什么要这么写……用户的问题和困惑越积越多，有些朋友建议我写一本 HanLP 的书。然而我认为一本书不应当局限于代码，而应当让读者知其所以然，而彼时我觉得自己才疏学浅，写不出满意之作。后来经过几年的完善，HanLP 成为 GitHub 上最受欢迎的自然语言处理项目，我对自然语言处理的理解也系统化了一些。正巧图灵的王军花老师跟我约稿，我想是时候将这些年的收获总结一下了。

这本书跟其他图书有什么不同

避免大而全式地泛泛而谈，又不拘泥于工程实践，这是我写作这本书秉持的原则。我希望这本务实的入门书，能够帮助零起点的你上手这门新学科，并且真正将自然语言处理应用在生产环境中。

书中不是枯燥无味的公式罗列，而是用白话阐述的通俗易懂的算法模型；书中不是对他人开源代码的堆砌，而是工业级开发经验的分享。

我以 HanLP 作者的身份，从基本概念出发，逐步介绍中文分词、词性标注、命名实体识别、信息抽取、文本聚类、文本分类、句法分析这几个热门问题的算法原理与工程实现。通过

对多种算法的讲解和实现，比较各自的优缺点和适用场景。这些实现并非教学专用，而是生产级别的成熟代码，可以直接用于实际项目。

理解这些热门问题的算法之后，本书会引导你根据自己的项目需求拓展新功能，最终达到理论和实践上的同步入门。

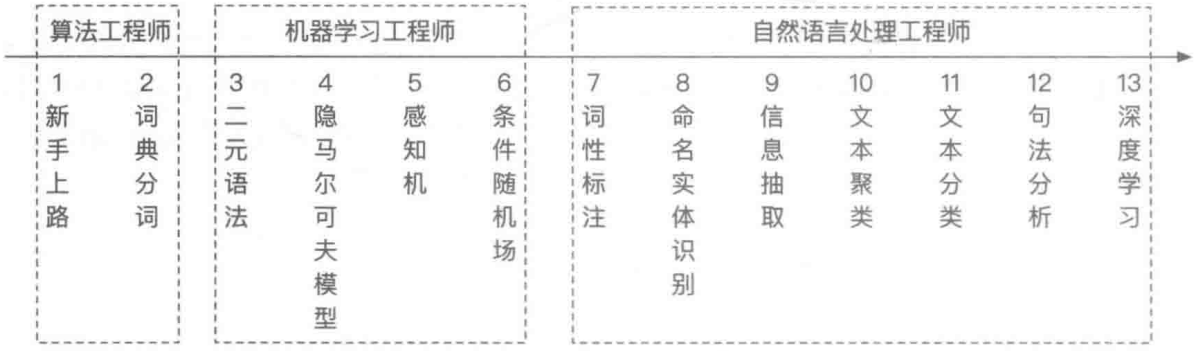
书中还会穿插一些你在网络资料中难得一见的实现技巧，巧妙运用的话会成为你高效开发的秘诀。读完本书后，你不光会理解理论、掌握接口，还能成长为自然语言处理类库的设计者。

无论是书还是代码，我都坚持“递归深入”“延迟加载”（lazy loading）的思想，即只在使用的时候才去加载必要的资料。也就是说，全书是自顶而下循序渐进的：

- 你首先看到的是一个摸得着的实际问题，为了解决该问题才去接触一个具体方案；
- 为了理解这个方案，才会介绍必要的背景知识；
- 为了实现这个方案，才会介绍相关细节；
- 为了克服这个方案的问题，才会过渡到新的方案。

主要内容

本书是自包含的，编排上尊重一般人的认知规律而不是学术上的纲目顺序。本书面向普通程序员，将内容粗略划分为下图所示的三大部分。



第一部分介绍一些字符串算法，让普通程序员从算法的角度思考中文信息处理。

第二部分由易到难地讲解一些常用的机器学习模型，让算法工程师晋级为机器学习工程师。这部分并非空谈理论，而是由中文分词这一应用问题贯穿始终，构成一种探索式的递进学习。这些模型也并非局限于中文分词，会在第三部分应用到更多的自然语言处理问题上去。

第三部分新增了许多与文本处理紧密相关的算法，让机器学习工程师进化到自然语言处理

工程师。特别地，最后一章介绍了当前流行的深度学习方法，起到扩展视野、承上启下的作用。你也可根据自身情况，灵活跳过部分章节。

翻阅本书，你会得到观影一般的流式体验。我曾经也是一无所知的外行，自学时走过不少弯路，深知数学语言的艰深晦涩，并且痛恨罗列公式故作高深的文章，也不喜欢大而全的综述书籍。因此，我写了这本不太一样的入门书，将阅读体验排在学术规范之前。尽量用自然的语调讲几个具体算法，把每个算法讲清楚，力争做到让你在地铁上也能把书读完读懂。

图片、公式与配套代码

本书为双色印刷，我们根据图书的特点对大部分图片进行了双色处理，其中有些跟作者使用代码输出的原始图片样式略有区别，但是在表达上的效果一致。

本书只保留了必不可少的公式和推导，以确保你充分理解为选择标准。书中的数学符号约定在“目录”之前的“主要数学符号表”中单独给出了，建议开始阅读之前了解一下。书中公式与代码相互印证，配套代码由 Java 和 Python 双语言写成，与 GitHub 上的最新代码同步更新，分别位于 https://github.com/hankcs/HanLP/tree/v1.7.5/src/test/java/com/hankcs/book 和 https://github.com/hankcs/pyhanlp/tree/master/tests/book[①]。为保证兼容，读者也可以使用命令 `git checkout v1.7.5` 切换到本书写作时的版本。对于 Java 代码，使用“类路径 # 方法名”来索引，比如 com.hankcs.book.ch01.HelloWord#main 表示源代码 HanLP/src/test/java/com/hankcs/book/ch01/HelloWord.java 中的 `main` 函数。对于 Python 代码，使用“模块路径 . 方法名”来索引，比如 tests.book.ch01.hello_word.main 表示源代码 pyhanlp/tests/book/ch01/hello_word.py 中的 `main` 函数。引用整个源码文件时，则直接使用文件的相对路径。另外，区别于正文，配套代码在书中印刷的背景色为淡蓝色。

思维导图

为了让读者纵览 NLP 领域的宏观图景，也为了帮助读者梳理 NLP 知识点，我精心打磨了一份 NLP + ML “双生树”思维导图。这可能是目前你所见到的最为详尽的思维导图，印刷尺寸大约宽 60 厘米，高 74 厘米，它是随书附赠给你参考学习的。

你不仅可以从中了解 NLP 领域的详尽知识脉络，还可以彻底弄清楚 NLP 与 ML 知识点之间的关联。这些关联知识点不仅涵盖本书中的核心知识，甚至涉及许多前沿研究与应用。不论你处于入门还是进阶阶段，这份思维导图都可以帮你厘清学科脉络。放在手边，时常拿出来参考一下，会相当便利。

① 你也可以到图灵社区（iturling.cn）本书主页“随书下载”下载源代码文件。

关于封面

本书封面上的图案是一只蝴蝶形状的词云，由全书 60 个关键术语构成。蝴蝶同时也是码农场与 HanLP 的标志，为业内人士熟知。蝴蝶象征着蝴蝶效应、非线性与混沌理论——虽然微小，但足以改变世界。本书虽属入门读物，但希望能成为读者漫漫修行路上那只扇动翅膀的蝴蝶。

致谢

本书的撰写得到了许多亲友和老师的帮助。

感谢我的父母，为我创造舒适的写作环境。

感谢我的导师 Choi 教授，你严谨的教研态度深深地感染着我。

感谢图灵编辑王军花和英子，在书稿的审核过程中提出了许多细致入微的高标准建议。

感谢夏志宏教授、周明副院长、李航博士、宗成庆教授、刘群教授、王斌教授、刘知远副教授、张华平副教授、@52nlp 为后辈小生作推荐。

感谢大快搜索创始人孙燕群、首席科学家汤连杰，为 HanLP 的研发提供诸多资源。

感谢开源社区的每一位用户与参与开发的黑客，是你们推动了中文信息处理在工业界的落地。

感谢每一位为本书做出贡献的朋友，我谨以此书作为回礼。

互动与勘误

虽然水平有限，但我对改进内容的热情是无限的。本书配套代码承诺与 HanLP 同步更新维护，欢迎大家积极参与开源项目。此外，也欢迎读者朋友们将对本书的评价和问题发在留言板 https://forum.hankcs.com/ 或者图灵社区本书主页上，大家一起探讨，谢谢。

何晗

2019 年 7 月

主要数学符号表

a	标量
$\boldsymbol{a}$	向量，$\boldsymbol{x}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}\in\mathbb{R}^{n\times 1}$
$\boldsymbol{A}$	矩阵或张量
a_i	向量$\boldsymbol{a}$的第i个元素，索引从1开始
$\boldsymbol{A}_{i,j}$	位于矩阵$\boldsymbol{A}$的第i行j列的元素
$\boldsymbol{A}_{i,:}$	矩阵$\boldsymbol{A}$的第i行
$\frac{\mathrm{d}y}{\mathrm{d}x}$	y关于x的导数
$p(a)$	随机变量a的概率分布
$\hat{p}(a)$	估计随机变量a的概率分布
$a\sim p$	a服从分布p
$p(a\mid b)$	随机变量a与b的条件分布
$p(a,b)$	随机变量a与b的联合分布
$\{0,1\}$	包含0和1的集合
$\neg\{0,1\}$	包含0和1的集合的补集
$\cup$	并集运算
$\cap$	交集运算
$\lvert A\rvert$	集合A的大小
$P(n,k)$	排列数，从n个元素中取k个排序的种数
$C(n,k)$	组合数，从n个元素中取k个组合的种数
$\boldsymbol{x}^{(i)}$	数据集的第i个样本的特征向量
$y^{(i)}$	数据集的第i个样本的标准答案（非结构化预测）
$\boldsymbol{y}^{(i)}$	数据集的第i个样本的标准答案（结构化预测）
$\hat{y}^{(i)}$	数据集的第i个样本的预测输出（非结构化预测）
$\hat{\boldsymbol{y}}^{(i)}$	数据集的第i个样本的预测输出（结构化预测）
$\boldsymbol{y}^{*}$	问题的最优解
e	自然常数，$\mathrm{e}=\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n}\approx 2.71828$
$\log x$	x以e为底的对数

目　录