附赠独家秘料
关注封底公众号获取

Python 文本数据分析与挖掘

利用 Python 来理解文本挖掘中会用到的自然语言处理技术、数学和统计学工具



[日] 山内长承 / 著

扫码即看视频
学习更高效更直观

◀Python 创始人
[荷兰] 范罗苏姆 (Gudio van Rossum)

中国青年出版社

需要全本请访问 www.ertongbook.com

Python 文本数据分析与挖掘

[日] 山内长承 著　张倩南 刘博 译

中国青年出版社

Original Japanese Language edition
PYTHON NI YORU TEXT MINING NYUMON
by Nagatsugu Yamanouchi
Copyright © Nagatsugu Yamanouchi 2017
Published by Ohmsha, Ltd.
Chinese translation rights in simplified characters by arrangement with Ohmsha, Ltd.
through Japan UNI Agency, Inc., Tokyo

律师声明

北京市京师律师事务所代表中国青年出版社郑重声明：本书由欧姆社授权中国青年出版社独家出版发行。未经版权所有人和中国青年出版社书面许可，任何组织机构、个人不得以任何形式擅自复制、改编或传播本书全部或部分内容。凡有侵权行为，必须承担法律责任。中国青年出版社将配合版权执法机关大力打击盗印、盗版等任何形式的侵权行为。敬请广大读者协助举报，对经查实的侵权案件给予举报人重奖。

主　　编　张　鹏
策划编辑　张　鹏
执行编辑　田　影
营销编辑　时宇飞
责任编辑　刘稚清
封面设计　乌　兰

侵权举报电话

全国“扫黄打非”工作小组办公室
010-65233456　65212870
http://www.shdf.gov.cn

中国青年出版社
010-59231565
E-mail: editor@cypmedia.com

版权登记号　01-2020-2532

图书在版编目（CIP）数据

Python文本数据分析与挖掘 /（日）山内长承著；张倩南，刘博译
. -- 北京：中国青年出版社，2021.3
ISBN 978-7-5153-6294-6

I. ①P... II. ①山... ②张... ③刘... III. ①软件工程-程序设计 IV. ①TP311.561

中国版本图书馆CIP数据核字（2021）第012707号

Python文本数据分析与挖掘

[日] 山内长承 / 著　张倩南 刘博 / 译

出版发行：中国青年出版社
地　　址：北京市东四十二条21号
邮政编码：100708
电　　话：（010）59231565
传　　真：（010）59231381
企　　划：北京中青雄狮数码传媒科技有限公司
印　　刷：天津旭非印刷有限公司
开　　本：880 x 1230　1/32
印　　张：5.5
版　　次：2021年6月北京第1版
印　　次：2021年6月第1次印刷
书　　号：ISBN 978-7-5153-6294-6
定　　价：79.80元（附赠独家秘料，关注封底公众号获取）

本书如有印装质量等问题，请与本社联系
电话：（010）59231565
读者来信：reader@cypmedia.com
投稿邮箱：author@cypmedia.com
如有其他问题请访问我们的网站：http://www.cypmedia.com

前　言

文本挖掘，顾名思义，就是从文本数据中挖掘信息，是一种从大量文本数据中把“有意义的信息”提取出来的处理技术。本书的目标是利用Python来理解文本挖掘中会用到的自然语言处理技术和数学、统计学工具。

在本书中，只是将自然语言处理技术和统计处理技术视为工具，不会涉及到其繁琐的原理、数学定理等。实际处理文本之后会知道些什么？能够做些什么分析？我们不妨带着这两个问题阅读本书。并且是在利用操作简便的包能处理的范围内，来探索文本挖掘可以帮我们做到的事情，而非用尖端的技术进行程序设计。

本书适用对象是多少有些编程经验，但没有语言处理经验的读者。虽然本书的编程语言选用的是Python，但读者并非必须有Python的经验，只要了解C、C++、Java等语言的基础编程概念即可，比如变量、代入、`if`语句、`for`语句之类的概念。对于完全没有接触过编程的读者，我建议先粗略地学习一下编程入门书籍，然后再尝试本书中的程序。当然，跳过程序部分来阅读本书也是可以的。

本书的结构是，第1章介绍完文本挖掘的整体概念之后，在第2章中，概括介绍第3章以后使用Python时的必要知识。在第3章中了解文本处理相关的基本概念和观点之后，在第4章中利用Python学习作为基础处理的频率分析方法和其能得到的结果。最后，在第5章中介绍文本挖掘需要用到的各种具体方法以及在Python中的处理步骤。

本书中使用的文本处理和文本挖掘方法以原理上重要的、基本的东西为中心。虽然从实际需求来看有不足的地方，但是读者可以以本书为出发点扩充自己的知识结构。在实际应用中，可以把本书的知识看作基本的核心部分，或者看作大型项目中的一个模块并灵活运用起来。

在阅读本书的过程中，很多地方可以通过实际程序处理来加深理解。

虽然本书并非是必须通过实际操作来学习的教科书，但是书中有好多涉及到多方面的、可以实际尝试的程序和数据，所以请一定在实际的处理环境中重新写一下本书中出现的程序。这样做之后，一定能收获到很多关于文本挖掘处理的认知和可能性的知识。

本书是笔者所在的东邦大学研究室的学习会，通过整理近七年间出版的实际操作教程和素材编写而成。这次，我接受了日本欧姆社书籍出版社各位的建议，以基于Python编程为中心出版本书。最后，我要向帮我制作原始素材的研究室的各位学生和给我很多建议的长期访问的研究员笕义郎先生表示衷心的感谢。同时，我还要向为了出版本书而做出多方努力的欧姆社书籍出版社的各位表示深深的谢意。

山内长承

需要全本请访问 www.ertongbook.com

目 录

前言……3

第1章 文本挖掘的概要 7

1.1 什么是文本挖掘……8
1.2 应用实例……10

第2章 Python概要和实验准备 13

2.1 什么是Python……14
2.2 编写、运行程序的环境……15
2.3 Python的语法规则……18
2.4 可用于文本挖掘的程序包……35
2.5 数据的准备……48

第3章 文本分割和数据分析的方法 57

3.1 文本的构成元素……58
3.2 统计分析、数据挖掘的基本方法……63
3.3 文本挖掘特有的方法……97

第4章 频率统计的实际应用 103

4.1 文字单位的出现频率分析……104
4.2 单词的出现频率分析……114

第5章 文本挖掘的各种处理示例 125

5.1 连续・N-gram的分析和利用……126
5.2 词的重要性和TF-IDF分析……131
5.3 基于KWIC的检索……137
5.4 基于单词属性的积极消极分析……139

5.5 基于WordNet的同义词检索 ……151
5.6 句法分析和关联分析的实际操作 ……157
5.7 语义分析和Word2Vec ……160

附录 Python编程环境的简单安装 163

A.1 什么是开发环境 ……164
A.2 在Windows10上的安装 ……165
A.3 开始使用Jupyter Notebook ……167
A.4 作业结果的保存和Jupyter Notebook的结束 ……171

通过列表解析式提升处理速度 ……26
Python的编程环境 ……30
方差、标准差的分母 ……67
MeCab的使用方法 ……121
Python的脚本、模块、程序包以及_main_ ……127
N-gram程序中的处理说明 ……136
名义尺度、定序尺度、定距尺度、定比尺度 ……175
日语WordNet的进化和NLTK界面的增加 ……184
Stanford Parser的安装 ……185

本书使用的Python代码，是从欧姆社的网页（http://www.ohmsha.co.jp/）下载的。

注意

- 本文件仅供购买本书的人使用。本文件的著作权属于本书的作者山内长承。
- 使用本文件产生的直接或间接的损失，作者以及欧姆社不负任何责任。请在自己承担责任的前提下，进行使用。

Python 文本数据分析与挖掘

[日] 山内长承 著　张倩南 刘博 译

中国青年出版社

Original Japanese Language edition
PYTHON NI YORU TEXT MINING NYUMON
by Nagatsugu Yamanouchi
Copyright © Nagatsugu Yamanouchi 2017
Published by Ohmsha, Ltd.
Chinese translation rights in simplified characters by arrangement with Ohmsha, Ltd.
through Japan UNI Agency, Inc., Tokyo

律师声明

北京市京师律师事务所代表中国青年出版社郑重声明：本书由欧姆社授权中国青年出版社独家出版发行。未经版权所有人和中国青年出版社书面许可，任何组织机构、个人不得以任何形式擅自复制、改编或传播本书全部或部分内容。凡有侵权行为，必须承担法律责任。中国青年出版社将配合版权执法机关大力打击盗印、盗版等任何形式的侵权行为。敬请广大读者协助举报，对经查实的侵权案件给予举报人重奖。

主　　编　张　鹏
策划编辑　张　鹏
执行编辑　田　影
营销编辑　时宇飞
责任编辑　刘稚清
封面设计　乌　兰

侵权举报电话

全国“扫黄打非”工作小组办公室
010-65233456　65212870
http://www.shdf.gov.cn

中国青年出版社
010-59231565
E-mail: editor@cypmedia.com

版权登记号　01-2020-2532

图书在版编目（CIP）数据

Python文本数据分析与挖掘 /（日）山内长承著；张倩南，刘博译
. -- 北京：中国青年出版社，2021.3
ISBN 978-7-5153-6294-6

Ⅰ. ①P… Ⅱ. ①山… ②张… ③刘… Ⅲ. ①软件工程-程序设计 Ⅳ. ①TP311.561

中国版本图书馆CIP数据核字（2021）第012707号

Python文本数据分析与挖掘

[日] 山内长承 / 著　张倩南 刘博 / 译

出版发行：中国青年出版社
地　　址：北京市东四十二条21号
邮政编码：100708
电　　话：（010）59231565
传　　真：（010）59231381
企　　划：北京中青雄狮数码传媒科技有限公司
印　　刷：天津旭非印刷有限公司
开　　本：880 x 1230　1/32
印　　张：5.5
版　　次：2021年6月北京第1版
印　　次：2021年6月第1次印刷
书　　号：ISBN 978-7-5153-6294-6
定　　价：79.80元（附赠独家秘料，关注封底公众号获取）

本书如有印装质量等问题，请与本社联系
电话：（010）59231565
读者来信：reader@cypmedia.com
投稿邮箱：author@cypmedia.com
如有其他问题请访问我们的网站：http://www.cypmedia.com

前　言

文本挖掘，顾名思义，就是从文本数据中挖掘信息，是一种从大量文本数据中把“有意义的信息”提取出来的处理技术。本书的目标是利用Python来理解文本挖掘中会用到的自然语言处理技术和数学、统计学工具。

在本书中，只是将自然语言处理技术和统计处理技术视为工具，不会涉及到其繁琐的原理、数学定理等。实际处理文本之后会知道些什么？能够做些什么分析？我们不妨带着这两个问题阅读本书。并且是在利用操作简便的包能处理的范围内，来探索文本挖掘可以帮我们做到的事情，而非用尖端的技术进行程序设计。

本书适用对象是多少有些编程经验，但没有语言处理经验的读者。虽然本书的编程语言选用的是Python，但读者并非必须有Python的经验，只要了解C、C++、Java等语言的基础编程概念即可，比如变量、代入、`if`语句、`for`语句之类的概念。对于完全没有接触过编程的读者，我建议先粗略地学习一下编程入门书籍，然后再尝试本书中的程序。当然，跳过程序部分来阅读本书也是可以的。

本书的结构是，第1章介绍完文本挖掘的整体概念之后，在第2章中，概括介绍第3章以后使用Python时的必要知识。在第3章中了解文本处理相关的基本概念和观点之后，在第4章中利用Python学习作为基础处理的频率分析方法和其能得到的结果。最后，在第5章中介绍文本挖掘需要用到的各种具体方法以及在Python中的处理步骤。

本书中使用的文本处理和文本挖掘方法以原理上重要的、基本的东西为中心。虽然从实际需求来看有不足的地方，但是读者可以以本书为出发点扩充自己的知识结构。在实际应用中，可以把本书的知识看作基本的核心部分，或者看作大型项目中的一个模块并灵活运用起来。

在阅读本书的过程中，很多地方可以通过实际程序处理来加深理解。

虽然本书并非是必须通过实际操作来学习的教科书，但是书中有好多涉及到多方面的、可以实际尝试的程序和数据，所以请一定在实际的处理环境中重新写一下本书中出现的程序。这样做之后，一定能收获到很多关于文本挖掘处理的认知和可能性的知识。

本书是笔者所在的东邦大学研究室的学习会，通过整理近七年间出版的实际操作教程和素材编写而成。这次，我接受了日本欧姆社书籍出版社各位的建议，以基于Python编程为中心出版本书。最后，我要向帮我制作原始素材的研究室的各位学生和给我很多建议的长期访问的研究员笕义郎先生表示衷心的感谢。同时，我还要向为了出版本书而做出多方努力的欧姆社书籍出版社的各位表示深深的谢意。

山内长承

需要全本请访问 www.ertongbook.com

目 录

前言……3

第1章 文本挖掘的概要 7

1.1 什么是文本挖掘……8
1.2 应用实例……10

第2章 Python概要和实验准备 13

2.1 什么是Python……14
2.2 编写、运行程序的环境……15
2.3 Python的语法规则……18
2.4 可用于文本挖掘的程序包……35
2.5 数据的准备……48

第3章 文本分割和数据分析的方法 57

3.1 文本的构成元素……58
3.2 统计分析、数据挖掘的基本方法……63
3.3 文本挖掘特有的方法……97

第4章 频率统计的实际应用 103

4.1 文字单位的出现频率分析……104
4.2 单词的出现频率分析……114

第5章 文本挖掘的各种处理示例 125

5.1 连续・N-gram的分析和利用……126
5.2 词的重要性和TF-IDF分析……131
5.3 基于KWIC的检索……137
5.4 基于单词属性的积极消极分析……139

5.5 基于WordNet的同义词检索 ……151
5.6 句法分析和关联分析的实际操作 ……157
5.7 语义分析和Word2Vec ……160

附录 Python编程环境的简单安装 163

A.1 什么是开发环境 ……164
A.2 在Windows10上的安装 ……165
A.3 开始使用Jupyter Notebook ……167
A.4 作业结果的保存和Jupyter Notebook的结束 ……171

通过列表解析式提升处理速度 ……26
Python的编程环境 ……30
方差、标准差的分母 ……67
MeCab的使用方法 ……121
Python的脚本、模块、程序包以及_main_ ……127
N-gram程序中的处理说明 ……136
名义尺度、定序尺度、定距尺度、定比尺度 ……175
日语WordNet的进化和NLTK界面的增加 ……184
Stanford Parser的安装 ……185

本书使用的Python代码，是从欧姆社的网页（http://www.ohmsha.co.jp/）下载的。

注意

- 本文件仅供购买本书的人使用。本文件的著作权属于本书的作者山内长承。
- 使用本文件产生的直接或间接的损失，作者以及欧姆社不负任何责任。请在自己承担责任的前提下，进行使用。

第1章

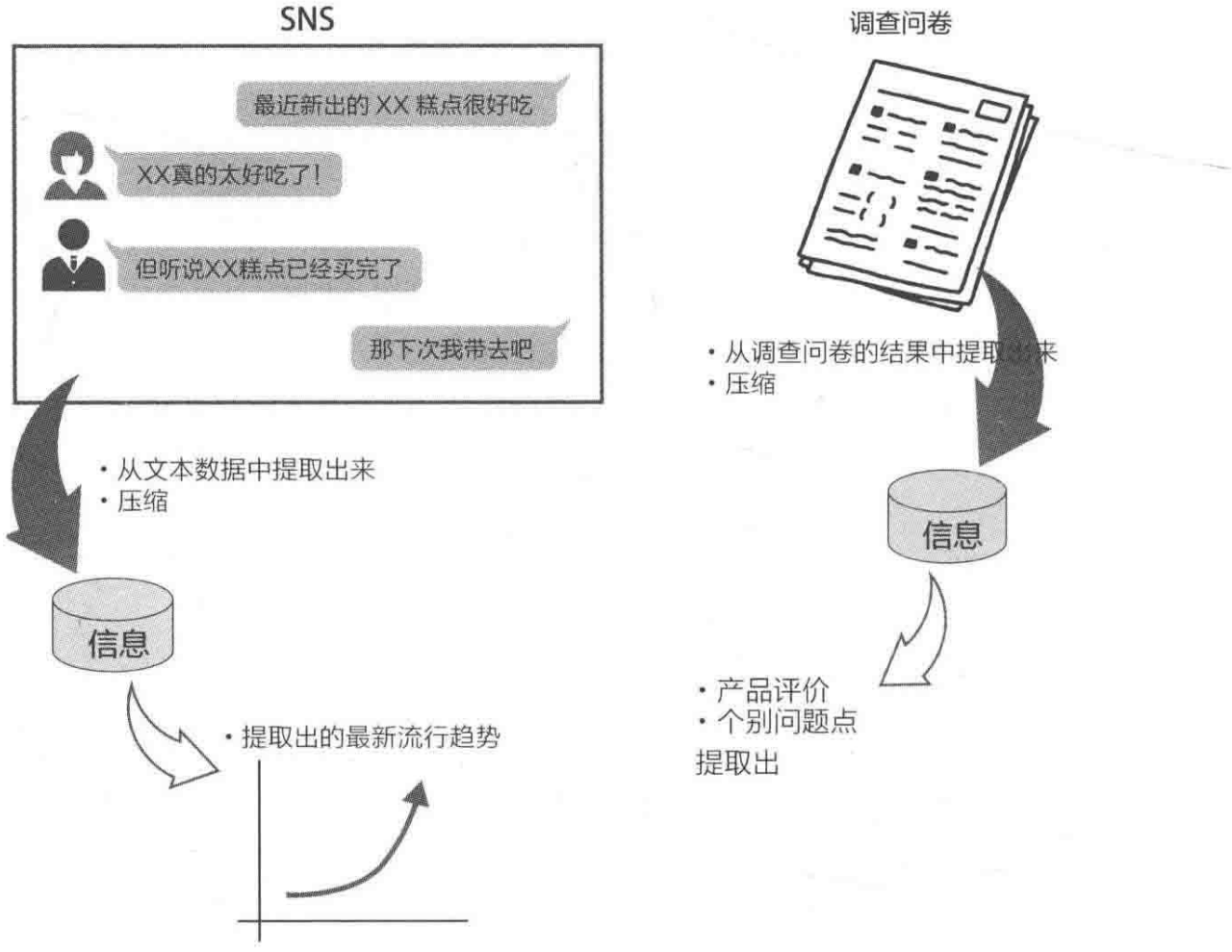
文本挖掘的概要

首先，我们要弄清楚三个问题：文本挖掘是干什么的？其内容是什么？在什么场合会用到文本挖掘？文本挖掘就是通过分析文本，把其中的信息压缩后提取出来的一种技术。需要准备的工具有两个：能把语言特征提取出来的自然语言处理技术和能从特征中把信息提取出来的统计处理技术。文本挖掘技术在问卷调查的意见分析、评价分析、话题的关联性分析和文书检索、分类等领域都得到了广泛的应用。

1.1 什么是文本挖掘

文本挖掘是指从文本数据中把信息挖掘出来，如同从大量沙土中找出被掩埋的宝贵钻石，文本挖掘是从大量文本数据中找出被掩埋的“有意义的信息”。

通过使用自然语言处理技术和统计学工具，从大量的文本数据中提取出压缩后有意义的信息。在这里，我们分别使用“文本数据”和“信息”这两个词。先说文本数据，比如各种各样的文件、在社交平台上发布的信息、关于产品或服务的问卷调查结果，这些数据的产生都有原本的目的。也就是说，我们是为了写文件、为了在社交平台上发布信息、为了评价产品或服务等目的，才去写作各种文本，而这些文本就是我们输入的“数据”。文本挖掘，就是从这些数据入手，比如从社交平台中提取出最近的流行趋势，从问卷调查中提取出对某种商品或服务的整体评价和出现的问题（图1-1）。而上面说的从社交平台上提取的流行趋势、通过问卷得到的评价和问题，和原来的文本数据相比，被大幅度地压缩了。我们把压缩后得到的东西称为“信息”。



■图1-1　文本挖掘

简而言之，文本挖掘可以说是一种从大量文本数据中获得压缩信息的过程。

文本挖掘近年来被广泛使用，结合其背景来看，可能有以下几个原因。首先，能够获得机器可读的文本数据是一个重要原因。比如，社交平台本来就是一种线上服务，文本会转换成计算机可读的形式。问卷调查在很多情况下也是通过网页在线上收集的*1。其次，值得特别指出的一个变化是，用来分析文本数据的程序和框架大部分被公开了，大家可以很容易地获取并使用这些程序。以前公司开发的软件和字典数据主要是限制性开放，并且是收费的，最近大学或个人开发的软件都是无偿公开给大家使用的，而且，不仅可以使用，使用者还可以加以改善，做出更加优秀的程序。而这种现象正在成为趋势。本书中介绍的各种软件也是这样积累而来的。

再者，最近有一个非常显著的变化，那就是有些文本分析框架结合了机器学习的方法。在文本分析程序中，除了以语言学、数学理论为基础一步步地分析处理这种方法以外，还有一种机器学习框架，即把大量常见的文本收集起来，把其统计学上的特性输入到程序中，使程序具备文本分析的功能。通过机器学习来构成复杂理论的这一方法以前也被人提出过，近几年取得了令人瞩目的进展，以图像处理为中心，被应用到很多场合，也有很多关于文本分析的研究正在进行。大量地收集机器学习需要的相关文本是非常有必要的，由于互联网的普及，我们可以轻易地在各种联网软件、网页等地方收集到很多文章。通过深度学习等最新的机器学习，来实现文本处理的研究成果还很少，所以在本书中不会涉及到这些内容，我们只是粗略地了解一下简单的机器学习。

通过Python的程序来理解文本挖掘的各种分析处理是本书的写作目的。但在实际应用中，文本分析以外的东西也是必要的，比如用Web来收集整理线上问卷的系统。如果是纸质问卷，还需要有人专门负责将其输入到计算机中，这个叫做“预处理”的部分，因为实际情况会涉及到多个方面，所以本书中只好将这部分舍弃。还有如何解释、说明分析的结果这一问题，因为这个问题会随着目的、数据的性质而变化，所以本书不会涉及。本书的主题是，在可以把文本作为分析对象的场合，思考其分析方法和应该使用的程序。

*1 目前需要手写的纸质问卷也不少，在这种情况下，就需要把问卷收集后，再输入到计算机中。

1.2 应用实例

文本挖掘技术有多种多样的实际应用，下面通过举例来大致了解都有哪些应用，并希望能以此来了解文本挖掘功能的整体面貌。

1.2.1 关于调查问卷的开放性问题或客服中心收到的提问和意见的分析

通过纸质媒体或互联网做问卷调查来探究大众对产品或服务的印象、评价的方法被大家广泛使用。调查问卷的项目中不是只有选择题，还有需要答卷人自己写答案的开放性问题。通过分析其中的文本内容，我们可以获知顾客对作为调查对象的产品或服务的反响。

对调查问卷的开放性问题这一栏目的分析，与其他文本分析相比，有一个不同之处，那就是其文本内容可以结合别的选择题一起分析。多个项目对比讨论的分析方法叫做交叉列联表，这和统计分析里的“多变量分析”是相对应的，但交叉列联表能够包含作为项目之一的开放性问题栏目的特征量。因此，一方面，我们可以做出有深度的分析，另一方面，我们可以得到问卷制作人没有预想到的内容，而这是选择题无法涵盖的。即使只是注意开放性问题栏目中出现的单词，也有机会抓住答卷人的动向。

在客服中心接待客人电话的情况下，分析对象通常是客服人员记录的内容，所以肯定不能涵盖用户所说的全部内容，但我们仍可以从中提取出提问内容的倾向、投诉的特征等。在客服接电话时，通常都会记录事先准备好的检查项目，但如果我们能分析客服听写记录下的内容，就有可能找出准备的检查项目中没有包含的要素。

1.2.2 社交平台上对特定商品和服务的评价分析

在互联网上发表的个人想法和心情，也可以通过文本挖掘进行分析。推特、脸书、博客等SNS（网络社交平台），用户只要登陆就可以在上面发表自己的想法和心情，由于发表门槛低，因此可以在这些平台上收集到很多人的意见。利用这个特点，我们可以提取出大众对于某个特定服务或商品的意见倾向，或者不针对特定对象，而是提取出社会整体的氛围和感情。

需要全本请访问 www.ertongbook.com