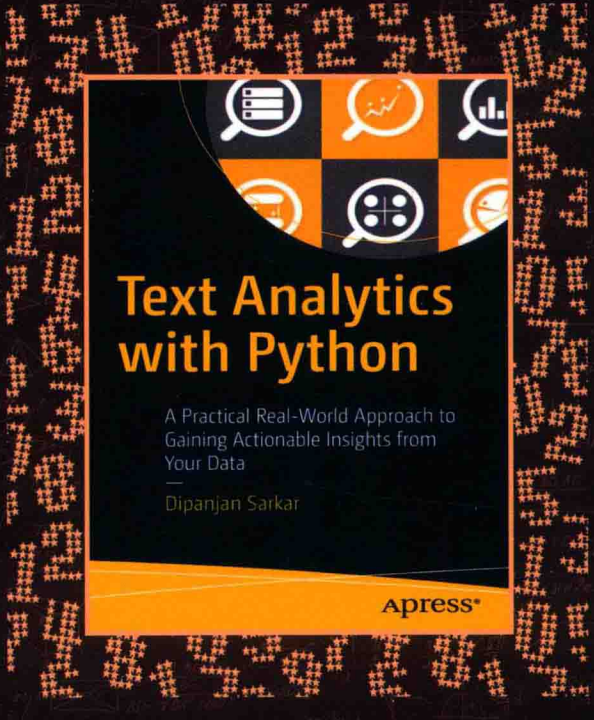
HZ BOOKS
华章IT
Apress®
数据科学与工程技术丛书

Python文本分析

[印度] 迪潘简·撒卡尔（Dipanjan Sarkar）著
闫龙川 高德荃 李君婷 译



TEXT ANALYTICS WITH PYTHON

A PRACTICAL REAL-WORLD APPROACH TO GAINING ACTIONABLE INSIGHTS FROM YOUR DATA

机械工业出版社
China Machine Press

数据科学与工程技术丛书

TEXT ANALYTICS WITH PYTHON

A PRACTICAL REAL-WORLD APPROACH TO GAINING ACTIONABLE INSIGHTS FROM YOUR DATA

Python文本分析

[印度] 迪潘简·撒卡尔（Dipanjan Sarkar）著

闫龙川 高德荃 李君婷 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 文本分析 /（印）迪潘简·撒卡尔（Dipanjan Sarkar）著；闫龙川，高德荃，李君婷译 . —北京：机械工业出版社，2018.3
（数据科学与工程技术丛书）
书名原文：Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data

ISBN 978-7-111-59324-9

I. P… II. ① 迪… ② 闫… ③ 高… ④ 李… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2018）第 043033 号

本书版权登记号：图字 01-2017-7335

Python 文本分析

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：唐晓琳　　责任校对：李秋荣
印　　刷：中国电影出版社印刷厂　　版　　次：2018 年 4 月第 1 版第 1 次印刷
开　　本：185mm×260mm 1/16　　印　　张：17.75
书　　号：ISBN 978-7-111-59324-9　　定　　价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066　　投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259　　读者信箱：hzit@hzbook.com

华章IT
HZBOOKS | Information Technology

译 者 序

自然语言处理和文本分析是当今人工智能研究和应用的重要方向，因其在人机交互方面的广泛应用和前景，吸引了学术界和产业界投入巨大的力量。目前，已经有一些产品陆续面世，在机器翻译、问答系统、语音助理、情感分析等方面取得了非常不错的进展，也给人们的生活带来了便利。

本书作者 Sarkar 是 Intel 公司的数据科学家，研究领域涉及数据科学与软件工程，有着丰富的文本分析研究和工程方面的经验，出版过多本 R 语言和机器学习方面的书籍。作者在 GitHub 上（https://github.com/dipanjanS/text-analytics-with-python）开源了本书相关的程序代码和数据集，感兴趣的读者可以下载研究。

本书首先介绍了与文本分析相关的自然语言基本概念以及 Python 语言的特点、特性和常用功能。然后，结合示例代码详细阐述了文本理解与处理、文本分类、文本摘要、文本相似性与聚类、语义与情感分析等内容，具有很强的实用性，内容覆盖了文本分析的重要方面，为相关应用的开发和研究提供了很好的参考借鉴。

本书是关于自然语言处理的实践教程，通过学习本书，读者可以全面地掌握文本分析的基础技术和机器学习的一些经典方法，包括 SVM、贝叶斯分类器、k-means 聚类、层次聚类等，为进一步的学习和研究奠定基础。感兴趣的读者可以继续研究和探索深度学习技术在文本分析中的应用，这是人工智能应用中发展非常迅速的领域，相信阅读本书打下的基础会对你大有帮助。

最后，感谢本书的作者和机械工业出版社华章公司的编辑，是他们的鼓励和支持使得本书能与读者见面。感谢我们家人的理解。尽管我们努力准确地表达作者的思想和方法，但仍难免有不当之处。译文中的错误，敬请指出，我们将非常感激，请将相关意见发往 yanlongchuan@iie.ac.cn。

闫龙川　高德荃　李君婷

2017 年 12 月

前　言

从高中开始接触数学和统计学以来，我就一直对数字着迷。分析学（analytics）、数据科学以及最近的文本分析技术均出现较晚，大概是在几年前，当时关于大数据（big data）和数据分析的炒作越来越猛烈，甚至有些疯狂。就个人而言，我认为其中很多都是过度炒作，但是也有一些令人兴奋的东西，因为这些技术在新工作、新发现以及解决人们先前认为不可能解决的问题方面展现了巨大的可能性。

自然语言处理（Natural Language Processing，NLP）一直深深吸引着我，因为人脑科学和人类认知能力确实令人着迷。如果尝试在机器中重塑这种传递信息、复杂思维和情绪的能力，那一定是令人惊喜的。当然，尽管我们在认知计算（cognitive computing）和人工智能（Artificial Intelligence，AI）方面的发展突飞猛进，但现在尚且无法实现这一点。仅通过图灵测试可能是不够的，机器真正能复制人的方方面面吗？

当今，对于NLP和文本分析应用，迫切需求从非结构化、原始文本数据中提取有用信息和可行见解的能力。到目前为止，我一直在努力解决各种问题，面临诸多挑战，并随着时间的推移吸取了各种各样的经验教训。本书涵盖了我在文本分析领域学到的大部分知识，仅仅从一堆文本文档中建立一个花哨的词云是不够的。

在学习文本分析方面，最大的问题或许不是信息缺乏，而是信息过多，通常这称为信息过载（information overload）。海量的资源、文档、论文、书籍和期刊包含了大量的理论资料、概念、技术和算法，它们常常使该领域的新手不知所措。解决问题的正确技术是什么？文本摘要如何真正有效？哪些才是解决多类文本分类的最佳框架？通过将数学和理论概念与现实用例的Python实现相结合，本书尝试解决这个问题，并帮助读者避免迄今为止我所遇到的一些急迫问题。

本书采用了全面的和结构化的介绍方法。首先，它在前几章中介绍了自然语言理解和Python结构的基础知识。熟悉了基础知识之后，其余章节将解决文本分析中的一些有趣问题，包括文本分类、聚类、相似性分析、文本摘要和主题模型。本书还将分析文本的结构、语义、情感和观点。对于每个主题，将介绍基本概念，并使用一些现实世界中的场景和数据来实现涵盖每个概念的技术。本书的构想是呈现一幅文本分析和NLP的蓝海，并提供必要的工具、技术和知识以处理和解决工作中遇到的问题。我希望你能觉得本书很有帮助，并祝你在文本分析的世界中旅途愉快！

目　录

第1章
自然语言基础

我们已迎来了大数据时代，组织和企业越来越难以管理由各种系统、过程和事务生成的所有数据。虽然，*大数据*的“3 V（Volume，Variety，Velocity）”（高容量，多样性，高速度）特征被广泛认可，但是其定义却比较模糊，导致了大数据（Big Data）这个术语常常被误用。这是因为人们有时很难准确地量化什么数据属于“大”数据。一些人可能把数据库中的10亿条记录看作大数据，但是与各种传感器甚至社交媒体生成的PB级数据相比，实际上该数据就显得量级相对较小。在所有组织，无论其从属何种行业领域，现今都有体量巨大的非结构化文本数据。仅举一些例子，我们有巨量的各种形式的数据，如推特、状态更新、评论、井号标签、文章、博客、维基信息和其他更多的社会媒体信息。即使在零售业和电子商务商店，也会基于新产品信息、客户评价和反馈元数据生成大量的文本数据。

与文本数据相关的挑战主要有两个方面。第一方面的挑战涉及数据的有效存储和数据管理。通常，文本数据是非结构化的，与关系数据库不同，其不遵循任何特定的预定义数据模型或模式。然而，基于数据语义，可以将数据要么存储在基于SQL的数据库管理系统（DBMS）中，如SQL Server，要么存储在基于NoSQL的系统中，如MongoDB。拥有海量文本数据集的组织通常使用基于文件的系统，例如Hadoop系统，其中所有数据以Hadoop分布式文件系统（Hadoop Distributed File System，HDFS）格式存储，并按需进行访问，这是*数据湖*（data lake）的主要原则之一。

第二方面的挑战是数据分析，以及尝试提取对于组织有价值、有意义的模式和有用的洞见。虽然有大量能任意使用的机器学习和数据分析技术，但其中大多数技术适用于数值型数据，所以必须借助于*自然语言处理*（Natural Language Processing，NLP）领域和专门的技术、变换和算法来分析文本数据，或者更具体地称为*自然语言*，它与机器容易理解的编程语言有着显著不同。请记住，高度非结构化的文本数据并不遵循结构化或规范化的语法和模式，因此不能直接使用数学模型或统计模型来分析它。

在深入文本数据分析的具体技术和算法之前，本章将讨论与文本数据特性相关的一些主要概念和理论基础。本章的主要目的是让你熟悉与*自然语言理解、处理和文本分析*相关的概念和领域。在本书中将使用Python编程语言，其主要用于访问和分析文本数据。本章中的示例将非常简单，而且易于理解。不过，如果想在阅读本章之前了解Python，你也可以快速浏览第2章的内容。本书中所有的例子可在原书作者的GitHub库（https://github.com/dipanjanS/text-analytics-withpython）中下载，其中包括程序、代码片段和数据集。本章介绍与自然语言、语言学、文本数据格式、句法、语义和语法相关的概念，然后再介绍高级的主

题，如文本语料库（text corpora）、NLP 和文本分析。

1.1 自然语言

虽然文本数据是非结构化数据，但它通常属于特定语言，遵循特定的语法和语义。任何文本数据片段（简单的单词、句子或文档）大多数情况下可追溯到一些自然语言。本节将讨论自然语言的定义、语言哲学、语言采集和语言的使用。

1.1.1 什么是自然语言

要理解文本分析和自然语言处理，我们需要了解到底是什么造就了语言的“自然性”。简单来说，自然语言是人类基于自然使用和交流而发展演化而来的语言，而不是像计算机编程语言那样由人工构造和创建的语言。

如英语、日语和梵语等人类语言都是自然语言。自然语言可以以不同的形式进行沟通和传递，包括言语、文字甚至符号。已经有大量的学术研究工作致力于理解语言的起源、本质和哲学。这将在下一节中简要讨论。

1.1.2 语言哲学

我们现在知道了自然语言是什么意思。但想想以下的问题：一门语言的起源是什么？

英国人的语言是因何形成“英语（English）”的？“fruit”这个词的意义是如何产生的？人们之间如何使用语言进行交流沟通？无疑，这些都是一些非常重大的哲学问题。

语言哲学主要解决以下四个问题，并探寻答案来解决它们：

- 语言中意义的本质。
- 语言用法。
- 语言认知。
- 语言与现实之间的关系。
- 语言中意义的本质涉及语言的语义和语意本身的特性。这里，语言哲学家或语言学试图找出语言对实际事物对象的意义——也就是说，任何词或句子的意义如何起源和产生，以及语言中不同的词语如何成为彼此的同义词并形成联系。另一个重要的研究内容是语言的结构和句法如何为语义奠定好基础，或者更具体地说，如何结构化地组织具有自身意义的词语以形成有意义的句子。语言学是对语言的科学研究，它是一个处理这些问题的专业领域，后面将对此进行更详细地讨论。句法、语义、语法和分析树是解决这些问题的一些方法。意义的语言学本质可以在两个人之间交流沟通时得以展现，这里的两个人分别记为发送者和接收者。对于发送者，意义是在向接收者发送消息时尝试表达或交流的内容，而对于接收者，意义是从接收消息的上下文中所理解或推断的内容。另外，从非语言的角度来看，诸如身体语言、既有经验和心理作用等都是语意的影响因素，考虑到这些因素，每个人都以自己的方式感知或推断出语言的意义。
- 语言用法更关心语言在各种场景和人类之间的交流中是如何使用的。这包括言语分析和说话时的语言用法，包括说话者的意图、语气、内容和表达消息时涉及的相关

动作。在语言学中这通常称为言语行为。语言创作的起源和人类认知活动——例如负责学习和使用语言的语言习得等更高级的概念也受到了重点关注。

- 语言认知侧重于研究人脑的认知功能如何实现理解和解释语言。考虑典型的发送者和接收者例子，涉及从消息沟通到解释的许多行为。语言认知试图发现在连接和关联特定词语成为句子以及构成一个有意义的消息时思维如何作用，当发送者和接收者使用语言来沟通消息时，语言与他们思维过程的关系是什么。
- 语言与现实之间的关系探讨语言表达的真实程度。通常，语言哲学家试图度量这些表达的事实符合度，以及它们如何与我们现实世界中的特定事件相关。这种关系可以用几种方式表示，我们将探讨其中的一些方式。

语义三角形模型是最流行的语义模型之一，其用于解释词语如何表达接收者心中的意义和想法，以及解释该意义如何与现实世界中的实体对象或事实联系起来。语义三角形模型由查尔斯·奥格登（Charles Ogden）和艾佛·理查德（Ivor Richards）在他们的著作《意义之意义》中提出，该书于1923年第一次出版，模型如图1-1所示。

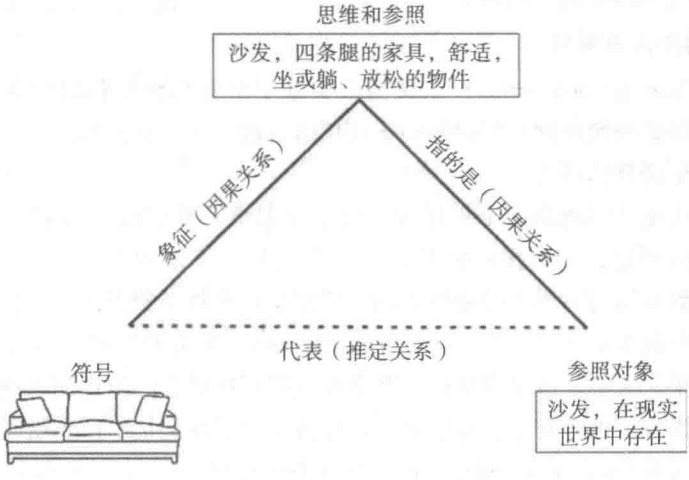


图1-1 语义三角形模型

语义三角形模型也称为意义之意义模型，图1-1中描述了一个真实例子，一个沙发置于某人面前，则可以被其感知。符号（symbol）表示语言符号，就像能唤起人脑中思想的词或物体。在这个例子中，符号是沙发，这唤起了什么是沙发的思想，一件家具，可以用于坐着或躺下和放松，它是让我们感到舒适的物体。这些思想称为参照，通过这个参照，人们能够将它与存在于现实世界中的称为参照对象的物体相关联。在本例中，参照对象是现实世界中的沙发。

找到语言和现实之间关系的第二种方式称为方向匹配（direction of fit），我们将在这里讨论两个主要方向。“词－世界”（word-to-world）的匹配方向谈论了语言使用能够反映现实的案例情况，即，使用词语来匹配或关联现实世界中正在发生的或已经发生的事情。举一个例子，句子“The Eiffel Tower is really big”，它强调了现实世界中的一个事实。另一个匹配方向，称为“世界－词”（world-to-word），它谈论语言使用可以改变现实的情况。举一个例子，句子“I am going to take a swim”，在这里人物“I”通过将要游泳正在改变现实，这通过正在交流的语句表达了相同的意义。图1-2显示了两种匹配方向之间的关系。

根据前面的描述，可以很清楚地看到，基于从现实世界感知的参照对象，一个人可以形成符号或词形式的表达，并且可以将其完全相同地传递给另一个人，这个人则基于所接收的符号建立对真实世界的画像表示，这样就构成一个循环。

1.1.3 语言习得和用法

到目前为止，我们已经掌握了自然语言所表示的意思和语言背后的概念、它的本质、意义和使用。本节将更深入地讨论如何使用人类的认知能力来感知、理解和学习语言，最后将讨论语言用法的主要形式，作为言语行为进行简要讨论。重要的是，不但要理解自然语言的含义，而且要理解人类如何解释、学习和使用相同的语言，以便我们尝试从文本数据中提取洞见时，能够在算法和技术中编程模拟其中的一些概念。

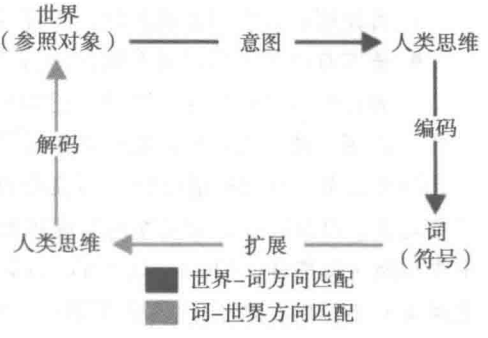


图 1-2 方向匹配表示图

1. 语言习得和认知学习

语言习得（language acquisition）定义为人类基于听觉和感知利用认知能力、知识和经验来理解语言并开始按照单词、短语和句子使用语言进行相互交流的过程。简单来说，语言习得是获取和产生语言的能力。

语言习得的历史可追溯到几个世纪前。哲学家和学者曾试图推理和理解语言习得的起源，并提出了几种理论，例如把语言看作一种代代相传下来的天赋的能力。柏拉图指出，词 - 义之间的映射在语言习得中起关键作用。许多学者和哲学家提出了现代理论，在一些流行理论中，最为著名的是 B. S. 斯金纳（B. S. Skinner）指出的知识、学习和语言的使用更多是一种行为结果。人类，或更具体地，儿童在使用任何语言的特定词语或符号时，会体验基于某些刺激的语言，由于对它们反复使用，导致了它们在记忆中得到增强。这个理论基于操作性条件反射（operant conditioning）或工具性条件反射（instrumentation conditioning），这是一种条件学习，其中基于其后果进行特定行为或行动强度的调整，例如奖励或惩罚，并且这些随后的刺激有助于增强或控制行为和学习。举一个例子，孩子们会学到由反复使用某个单词组成的一个特定的声音组合，这可能是通过他们的父母，或者通过他们正确地说这个字时得到的奖赏回报，或者说错时而被纠正。未来，这种反复的条件反射最终将强化孩子记忆中对该词的实际意义和理解。总而言之，孩子们基本上通过行为模仿和听成年人讲话来学习和使用语言。

然而，这种行为理论受到著名语言学家诺姆·乔姆斯基（Noam Chomsky）的挑战，他认为，孩子们不可能只通过模仿成人的一切来学习语言。这个假设在下面的例子中是成立的。虽然像 go 和 give 这样的词是有效的，但是孩子们常常最后会使用这个单词的无效形式，例如使用 goed 或 gived 来代替过去时态的 went 或 gave。可以确保他们的父母没有在孩子面前说出这些话，所以根据先前的 Skinner 理论，孩子们习得这些是不可能的。所以，乔姆斯基提出儿童不仅要模仿他们所听到的言语，还从相同的语言结构中提取模式、语法和规则，这与仅仅运用基于行为的通用认知能力是不同的。

按照乔姆斯基的观点，认知能力以及与特定语言相关的知识和能力，如句法、语义、言

语概念和语法，一起形成了他所称的语言习得机制，从而使人类能够具备语言习得的能力。除了认知能力之外，在语言学习中独特和重要的是语言本身的语法，这在他的名句“Colorless green ideas sleep furiously”中进行了强调。如果你观察这个句子并重复多次，会发现它是没有意义的。Colorless 与 green 矛盾，ideas 不能与 green 关联，它们也不能与 sleep furiously 搭配。然而，从语法上来讲，这个句子符合句法。这正是乔姆斯基试图解释的——句法和语法描述的信息独立于词的意义和语义。因此，他提出与其他认知能力相比，语言句法的学习和识别是人类的一种独特能力。这个假设也称为句法自主性（autonomy of syntax）。虽然这些理论仍然受到学者和语言学家的广泛争论，但是它对于探索人类心智如何习得和学习语言都非常有益。现在，我们将看看通常情况下语言使用的典型模式。

2. 语言用法

上一节讨论了言语行为以及如何使用方向匹配模型将词和符号与现实相关联。本节将介绍与言语行为相关的一些概念，重点讲述在交流中使用语言的不同方式。

言语行为主要分为三类：言内行为（locutionary）、言外行为（illocutionary）和言后行为（perlocutionary）。言内行为主要涉及当句子通过说话从一个人表达给另一个人时的实际传递行为。言外行为更关注于所表达句子的实际语义和意图。言后行为从心理或行为角度更注重言语表达对其接收者的实际影响。

举一个简单的例子，父亲对他孩子说的短语“Get me the book from the table”。这位父亲说出短语时则形成了言内行为。这句话的意图是一条指令，命令孩子替他去从桌上取书，它形成一种言外行为。孩子听到后所采取的行动，也就是说，如果他把书从桌子取给他的父亲，则形成了言后行为。

在上面这个例子中，言外行为是一个指令。根据哲学家约翰·塞尔（John Searle）的理论，总共有如下五种不同类型的言外行为。

- 断言（assertive）是说明事物已经存在于世界的言语行为。当发言者试图断言在现实世界中命题可能是真或假时，发言者说出的就是断言。这些断言可以是声明或陈述。举一个简单的例子，“The Earth revolves round the Sun”。这些信息表示前面讨论的“词－世界”的匹配方向。
- 指令（directive）是说话者向听话者表达请求或指挥他们做某事的言语行为。这表示听话者在接收到来自说话者的指令之后可能采取的自愿行动。既然指令是自愿性的，就可以遵从或不遵从它。这些指令可以是简单的请求，甚至命令或指令。“Get me the book from the table”是一个指令示例，我们先前谈论言语行为类型时讨论过。
- 承诺（commisive）是说话者或发言者按其所说，承诺一些未来自愿行为或行动的言语行为。诸如承诺、誓言、保证和宣誓等言语行为即为承诺，其匹配方向可以有两种。“I promise to be there tomorrow for the ceremony”是一个承诺类的例句。
- 表达（expressive）表明发言者或说话者对通过消息所传递的特定主张的意向和观点。它可以包含各种表达形式或情感类型，例如祝贺、讽刺等。“Congratulations on graduating top of the class”是一个表达类的例句。
- 宣告（declaration）是强大的言语行为，具有基于说话者/发送者表达消息中所宣称的主张来改变现实的能力。它的匹配方向通常是“世界－词”模式，但也可以是相反方向。“I hereby declare him to be guilty of all charges”是一个宣告类的例子。

这些言语行为是人类之间使用和传递语言的主要方式，我们往往并不会意识到，在一天中，我们可能会使用上述言语行为上百次。下面我们来看看语言学以及一些与它相关的主要研究领域。

1.2 语言学

我们已经介绍了是什么自然语言，如何学习和使用语言以及语言习得的起源。在语言学中，这些是语言学家等研究人员和学者们所研究的内容。严格来讲，*语言学*定义为对语言的科学研究，包括语言的形式和句法、意义和由语言用法和使用语境描述的语义。语言学的起源可追溯到公元前4世纪，当时的印度学者和语言学家帕尼尼（Panini）对梵语语言进行了形式化描述。1847年首先提出专业术语*语言学*（linguistics），用来表明对语言的科学研究，在此之前用于表达相同意思的专业术语是*文献学*（philology）。虽然文本分析不需要语言学的详细知识，但是了解语言学的不同领域还是十分有用的，因为其中一些领域在自然语言处理和文本分析算法中广泛使用。在语言学中，主要的专业研究领域如下。

- *语音学*（phonetics）：这是对在讲话过程中由人类声道产生的声音的声学性质的研究。它包括研究声音的特性，以及人类如何创造声音。在具体语言中，人类语音的最小个体单位称为*音位*（phoneme）。对于这个语音单位，一个更通用的跨语言术语是*音素*（phone）。
- *音韵学*（phonology）：这是对人类思维如何解释声音模式的研究，用于区分不同的音位，以找出哪些音位是重要的。通常，通过逐一考虑特定语言来详细研究音位的结构、组合和解释。英语由大约45个音位组成。音韵学通常不仅研究音位，还包括口音、语气和音节结构等内容。
- *句法*（syntax）：这通常是对句子、短语、词语及其结构的研究。它包括研究词语如何在语法上组合在一起以构成短语和句子。在短语或句子中，使用词语的句法顺序很重要，因为顺序会完全改变其含义。
- *语义*（semantics）：这涉及语言中意义的研究，并且可以进一步细分为词汇语义（lexical semantics）和成分语义（compositional semantics）。
  - *词汇语义*：使用形态和句法来研究词和符号的意义。
  - *成分语义*：研究词语和词语组合之间的关系，理解短语和句子的意义以及它们之间的相关性。
- *形态学*（morphology）：*语素*（morpheme）是具有区别性意义的最小语言单位。这包括诸如词语、前缀、后缀等具有各自不同含义的内容。形态学是对一门语言中这些不同单元或语素的结构和意义的研究。特定的规则和句法通常是控制语素组合在一起的方式。
- *词汇*（lexicon）：这是对语言中使用的单词和短语属性以及它们如何构建语言词汇的研究。这些包括何种声音与词语的含义相关联、词语所属的词性，以及它们的形态构成。
- *语用学*（pragmatics）：它研究语言和非语言因素（如上下文和场景）如何影响一条消息或一个话语所表达的意义。这包括设法推断在沟通交流中是否存在隐含或间接的意义。

- 话语分析（discourse analysis）：它通过人们的对话来分析语言和以句子形式的信息交换。这些对话可以通过口头、书面，甚至手势进行表达。
- 文体学（stylistics）：这是侧重于对语言写作风格的研究，包括语气、口音、对话、语法和语态类型。
- 符号学（semiotics）：这是对符号、标记和符号过程以及它们如何传达意义的研究。这个领域涵盖类比、隐喻和象征主义等内容。

上述内容是语言学研究探讨的主要领域，但语言学属于一个庞大的领域，其范围远远大于这里所提及的内容。然而，诸如语言句法和语义之类的内容是一些最重要的概念，这些概念基本上奠定了自然语言处理的基础。以下章节将更深入地介绍它们。

1.3 语言句法和结构

我们已经知道语言、句法和结构所表示的内涵。句法和结构通常是密切相关的，一组特定的规则、惯例和原则通常规定了单词组成短语、短语组成从句、从句组成句子的形式。本节专门讨论英语的句法和结构，因为在本书中我们将处理英文类型的文本数据。但是，这些概念也可以扩展到其他语言类型。关于语言结构和句法的知识对许多领域有所帮助，例如文本处理、文本标注，以及用于进一步的文本分析，如文本分类或摘要。

在英语中，单词常常组合在一起形成其他语言成分。这些成分要素包括单词、短语、从句和句子。所有这些成分在任何消息中一起存在，并且在层次结构中彼此相关。此外，句子是一种表达一组词汇的结构化格式，只要它们遵循某些句法规则，如语法。请看图 1-3 中所示的单词。

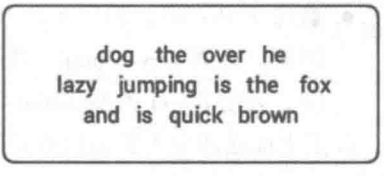


图 1-3 没有任何关系或结构的单词集合

在图 1-3 的单词集合中，很难确定它可能正在尝试传递什么或意味着什么。事实上，语言不仅仅是非结构化词汇的堆叠。正确句法不仅有助于我们获取正确的句子结构和相关词，还可以帮助句子根据单词的顺序或位置来表达意义。就我们先前的“句子→从句→短语→单词”的层次结构而论，我们能使用浅层分析（shallow parsing）构建图 1-4 中的分层句子树，浅层分析是一种用于查找句子成分的技术。

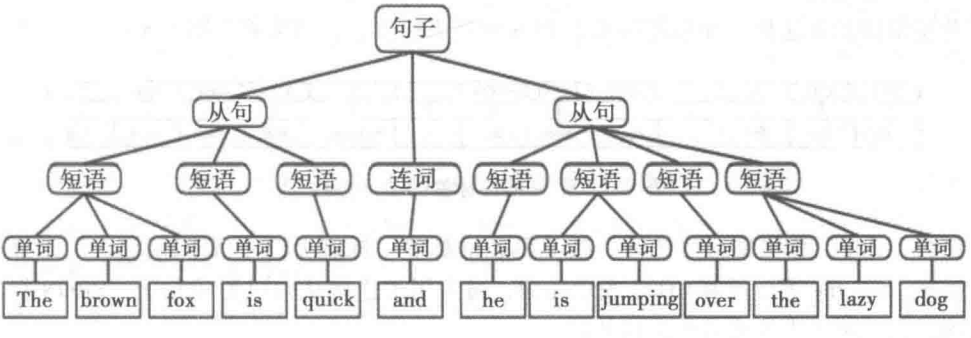


图 1-4 遵循分层句法的结构化语句

根据图 1-4 中的分层树，我们得到了句子“The brown fox is quick and he is jumping over the lazy dog”。可以看到，分层树的叶节点由单词组成，这些单词是树的最小单元，单词之间组合成短语，短语再依次形成从句。从句通过各种填充词或词语（如连词）连接在一起，形成最终的句子。下一节将进一步详细介绍这里提及的每种组成部分，了解如何分析它们，并找出主要的句法类别。

1.3.1 词

词（word）是一门独立语言中的最小单位，具有其特有的含义。虽然语素是有独特意义的最小语言单位，但语素与词不同，它不是独立的，而一个单词可以由几个语素组成。标注和标记单词并分析其词性（Parts Of Speech，POS）对于查看主要句法类别是有用的。这里将介绍各种 POS 标签的主要类别和意义。第 3 章将进一步详细研究它们，并学习编程生成 POS 标签的方法。

通常，单词可以归为以下主要的词类之一。

- 名词（noun）：通常表示描述一些可能有生命或无生命的物体或对象的单词。例如，fox、dog、book 等。名词的 POS 标签是 N。
- 动词（verb）：动词是用来描述某些动作、状态或事件的词。动词有各种各样的子类组成，如助（auxiliary）动词、反身（reflexive）动词和及物（transitive）动词等。动词的一些典型例子如 running、jumping、read 和 write。动词的 POS 标签是 V。
- 形容词（adjective）：形容词是用于描述或限定其他词——通常是名词和名词短语的词。短语 beautiful flower 有一个名词（N）flower，用形容词（ADJ）beautiful 来描述或限定。形容词的 POS 标签是 ADJ。
- 副词（adverb）：副词通常用作其他单词的修饰词，包括名词、形容词、动词或其他副词。短语 very beautiful flower 有副词（ADV）very，它修饰形容词（ADJ）beautiful，表明 flower 的 beautiful 程度。副词的 POS 标签是 ADV。

除了上面这四个主要类别的词类外，还有其他词类在英语中经常出现。这些词类包括代词（pronoun）、介词（preposition）、感叹词（interjection）、连词（conjunction）和限定词（determiner）等。而且每个 POS 标签均可进一步细分，如名词（N）可以进一步细分为单数名词（NN）、单数专有名词（NNP）和复数名词（NNS）。第 3 章将更详细地介绍 POS 标签，到时我们将处理和解析文本数据并实现使用 POS 标签来标注文本。

考虑先前的例句（The brown fox is quick and he is jumping over the lazy dog），我们已经构建了分层结构的句法树，如果使用基本 POS 标签来标注它，结果将如图 1-5 所示。

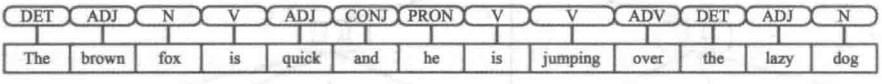


图 1-5 带有词性标签的单词标注

在图 1-5 中，你可能会注意到一些不熟悉的标签。标签 DET 代表限定词，用于描述数量，如 a、an、the 等。标签 CONJ 表示连词，通常用于连接从句以形成句子。PRON 标签代表代词，表示用于表达或替代名词的词。

标签 N、V、ADJ 和 ADV 是典型的开放性词类，代表属于开放性词汇表的词。开放性词