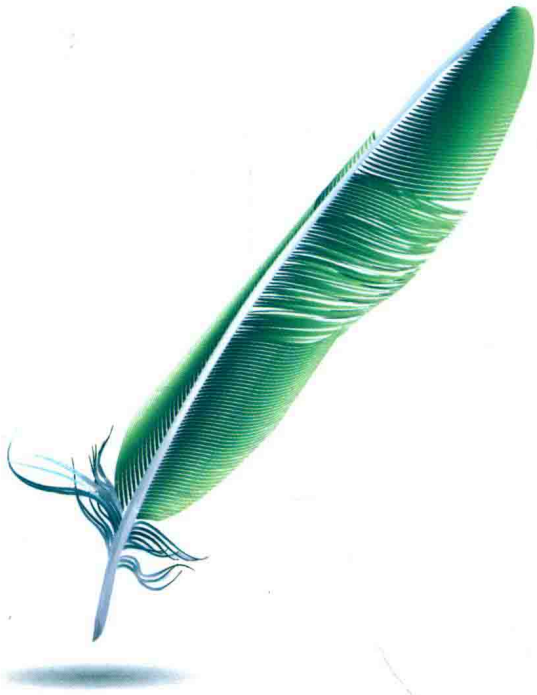
大数据与人工智能技术丛书



大数据专业英语教程

◎ 张强华 刘俊辉 郑聪玲 司爱侠 编著

清华大学出版社

大数据与人工智能技术丛书

大数据专业英语教程

◎ 张强华 刘俊辉 郑聪玲 司爱侠 编著

清華大學出版社
北京

内 容 简 介

本书是大数据专业英语教材，内容包括大数据基础、软件与开发技术、操作系统、Python与R编程语言、数据结构、数据库与数据仓库、云存储与数据备份、数据处理与数据清洗、数据挖掘、Hadoop与Spark、数据可视化、大数据安全等。

本书体例新颖，适合教学。每个单元均包含以下部分：课文——选材广泛、风格多样、切合实际的两篇专业文章；单词——给出课文中出现的新词，读者由此可以积累大数据专业的基本词汇；词组——给出课文中的常用词组；缩略语——给出课文中出现的、业内人士必须掌握的缩略语；难句讲解——讲解课文中出现的疑难句子，分析其语法结构，培养读者的阅读理解疑难句子的能力；习题——既有针对课文的练习，也有一些开放性的练习；短文翻译——培养读者的翻译能力；参考译文——让读者对照理解以提高翻译能力。

本书吸纳了作者近20年的IT行业英语翻译与图书编写经验，与课堂教学的各个环节紧密结合，支持备课、教学、复习及考试各个教学环节，有配套的PPT、参考答案等。

本书既可作为高等本科院校、高等专科院校大数据相关专业的专业英语教材，也可供从业人员自学；作为培训班教材，亦颇得当。

图书在版编目(CIP)数据

大数据专业英语教程/张强华等编著.—北京：清华大学出版社，2019
（大数据与人工智能技术丛书）
ISBN 978-7-302-52692-6

Ⅰ. ①大… Ⅱ. ①张… Ⅲ. ①数据处理–英语–教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2019）第057436号

策划编辑：魏江江
责任编辑：王冰飞　李　晔
封面设计：刘　键
责任校对：李建庄
责任印制：丛怀宇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954

印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张：**16　　**字　　数：**367千字
版　　次：2019年7月第1版　　**印　　次：**2019年7月第1次印刷
印　　数：1～1500
定　　价：39.80元

产品编号：079685-01

前言

我们正在从信息技术时代进入数据技术时代。我国的大数据产业已经进入高速发展期，许多高校都开设了大数据专业，培养急需的专业人员。由于大数据产业有极高的发展速度，从业人员必须掌握许多新技术、新方法，因此对专业英语要求较高。具备相关技能并精通专业外语的人员往往会赢得竞争，成为职场中不可或缺的核心人才与领军人物。

本书的特点与优势如下：

（1）选材全面，包括大数据基础、软件与开发技术、操作系统、Python 与 R 编程语言、数据结构、数据库与数据仓库、云存储与数据备份、数据处理与数据清洗、数据挖掘、Hadoop 与 Spark、数据可视化、大数据安全等。书中许多内容非常实用，具有广泛的覆盖面。

（2）体例新颖，非常适合教学，与课堂教学的各个环节紧密结合，支持备课、教学、复习及考试各个教学环节。每个单元均包含以下部分：课文——选材广泛、风格多样、切合实际的两篇专业文章；单词——给出课文中出现的新词，读者由此可以积累大数据专业的基本词汇；词组——给出课文中的常用词组；缩略语——给出课文中出现的、业内人士必须掌握的缩略语；难句讲解——讲解课文中出现的疑难句子，分析其语法结构，培养读者的阅读理解疑难句子的能力；习题——既有针对课文的练习，也有一些开放性的练习；短文翻译——培养读者的翻译能力；参考译文——让读者对照理解以提高翻译能力。

（3）习题量适当，题型丰富，难易搭配，便于教师组织教学。

（4）教学支持完善，有配套的 PPT、参考答案等。

（5）作者有近 20 年 IT 行业英语图书的编写经验。在作者编写的英语书籍中，有三部国家级“十一五”规划教材，一部全国畅销书，一部获华东地区教材二等奖图书。基于这些图书的编写经验有助于本书内容的完善与提升。

在使用本书的过程中，有任何问题都可以通过电子邮件与我们交流，我们一定会给予答复。邮件标题请注明姓名及“索取大数据英语参考资料”字样。我们的 E-mail 地址为 zqh3882355@sina.com 和 zqh3882355@163.com。

如本书有任何不妥之处，望大家不吝赐教，让我们共同努力，使本书成为一部“符合学生实际、切合行业实况、知识实用丰富、严谨开放创新”的优秀教材。

作　者

2019 年 4 月

目 录

Unit 1

Text A

Big Data

Big data is changing the way people work together within organizations. It is creating a culture in which business and IT leaders must join forces to realize the value from all data. Insights from big data can enable all employees to make better decisions — deepening customer engagement, optimizing operations, preventing threats and fraud, and capitalizing on new sources of revenue.

1. The Big Vs

1.1 Value

This is indeed the holy grail of big data and what we are all looking for. One has to demonstrate value that can be extracted from big or small data in order to justify the investments, whether on big data or on traditional analytics, data warehouse or business intelligence tools, whatever may be the buzzing nomenclature. There seems to be an increasing interest related to the value of big data, as indicated by the number of Google searches looking for similar terms over the last two years.

1.2 Volume

There is no doubt that the information explosion has redefined the connotation of volumes. There are several such staggering statistics going around and it has become

increasingly difficult to keep track of the number and magnitude of the prefixes attached to "bytes" while measuring the volume. Since there is a "helluva lot of data", the term "Hellabyte" has been coined beyond Petabytes, Exabytes, Zettabytes and Yottabytes. However, since these measures will be superseded by the likes of Brontobytes, Geopbytes and more, lets move on!

1.3 Velocity

Similarly, velocity refers to the speed at which the data is generated. Some of the factors that exacerbate this trend are the proliferation of social media and the explosion of IoT (Internet of Things). In the context of business operations that have not yet been touched by social media or IoT, the velocity arises from sophisticated enterprise applications that capture each and every minute detail involved in the completion of a particular business process. Enterprise applications have traditionally captured such information but the world has woken up to the power of such information largely in the big data era.

1.4 Variety

The last of the original attributes of big data is variety. Since we are living in an increasingly digital world where technology has invaded into our glasses and watches, the variety of data that is generated is mind-boggling. The computing power available is able to process unstructured text, images, audio, video and data from sensors in the IoT (Internet of Things) world that capture (almost) everything around us. This attribute of big data is more relevant today than it ever was.

1.5 Veracity or Validity

Veracity or validity of data is extremely important and fundamental to the extraction of value from the underlying data. Veracity implies that the data is verifiable and truthful. If this condition is violated, the results can be catastrophic. More importantly, there are several cases in which the data is accurate but may not be valid in the particular context. For instance, if we are trying to ascertain the volume of searches on Google related to big data, we will also obtain results pertaining to the hit single "dangerous" from "big data".

1.6 Visible

Information silos have always existed within enterprises and have been one of the major roadblocks in the attempt to extract value from data. Relevant information should not only exist, but also be visible to the right person at the right time. Actionable data needs to be visible transcending the boundaries of functions, departments and even organizations for value unlocking. Individuals might have believed that information in their hands is power but

in the age of big data, collective information available to the world at large is truly omnipotent!

1.7 Visual

We live in an increasingly visual world and the statistics of increase in the number of images and videos shared on the Internet is staggering. According to official statistics, 300 hours of video are uploaded every minute on YouTube. In a business context, appropriate visualization of data is critical for the management to be able to extract value from their limited time, resources and even more limited attention span!

2. More Contenders

In addition to the 7 V's described above, there are several other V's that may be considered:

2.1 Volatility

With more applications such as SnapChat and IoT sensors, we may have data in and out in a snap. Volatility of the underlying data sources may become one of the defining attributes in the future.

2.2 Variability

One of the cornerstones of traditional statistics is standard deviation and variability. Whether or not it makes to an extended list of V's relating to big data, it can never be ignored.

2.3 Viability

Embedded in the concept of value is the need to check the viability of any project. Big data projects can scale up to gigantic proportions and guzzle a lot of resources very quickly. Those who do not learn this fast and get fascinated with fads will funnel funds towards futility resulting in failure. In a nutshell, viability of any project needs to be established and big data projects do not have the liberty of exemption, whether or not it remains a trending buzzword.

2.4 Vitality

Vitality or criticality of the data is another concept that is crucial and is embedded in the concept of Value. Information that is more meaningful or critical to the underlying business objective needs to be prioritized. Analysis paralysis needs to be replaced with a more pragmatic approach. Technology allows marketers to create segments of one, but is such extreme segmentation vital or even aligned to the organizational strategy?

2.5 Vincularity

Derived from Latin, it implies connectivity or linkage. This concept is very relevant in today's connected world. There is significant value arbitrage potential by connecting diverse information sets. For instance, the government has forever been trying to connect the details of major expenditure heads and correlating the same with the income declared in tax returns to identify concealment of income. The same purpose may now be achieved by drawing information from social media posts.

3. An Example of Big Data

An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people — all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible.

New Words

realize	['riəlaiz]	*vt.*认识到，了解，实现，实行
engagement	[in'geidʒmənt]	*n.*参与度，敬业度
fraud	[frɔ:d]	*n.*欺骗，欺诈行为
indeed	[in'di:d]	*adv.*真正地，确实；当然
demonstrate	['demənstreit]	*vt.*示范，证明，论证
nomenclature	[nə'mənclet∫ə]	*n.*系统命名法；命名；术语；专门名称
analytics	[ˌænə'litiks]	*n.*分析学，解析学，分析论
redefine	[ˌri:di'fain]	*v.*重新定义
connotation	[ˌkɔnəu'tei∫ən]	*n.*内涵
staggering	['stægəriŋ]	*adj.*令人惊愕的，难以置信的
helluva	['heləvə]	*adj.*很大的
Hellabyte	['heləbait]	*n.*数据单位，$=10^{27}$ Byte
Exabyte	['eksəbait]	*n.*数据单位，缩写为 EB
Zettabyte	['zetəbait]	*n.*数据单位，缩写为 ZB
Yottabyte	['jɔtəbait]	*n.*数据单位，缩写为 YB
Brontobyte	['brɔntəbait]	*n.*数据单位，缩写为 BB
Geopbyte	['dʒiəpbait]	*n.*数据单位，缩写为 GB
velocity	[vi'lɔsiti]	*n.*高速性；速度，速率
exacerbate	[eks'æsəbeit]	*vt.*使恶化，使加剧
trend	[trend]	*n.*倾向，趋势

proliferation	[prəuˌlifəˈreiʃən]	*n.*增殖；扩散
explosion	[iksˈpləuʒən]	*n.*爆发，爆炸
era	[ˈiərə]	*n.*时代，纪元，时期
variety	[vəˈraiəti]	*n.*多样性；品种，种类
attribute	[əˈtribju(:)t]	*n.*属性，品质，特征
mind-boggling	[maind-ˈbɔgliŋ]	*adj.*令人难以置信的
unstructured	[ʌnˈstrʌktʃəd]	*adj.*非结构化的，未组织的
veracity	[vəˈræsiti]	*n.*真实性
validity	[vəˈliditi]	*n.*有效性；合法性，正确性
extremely	[iksˈtri:mli]	*adv.*极端地，非常地
fundamental	[ˌfʌndəˈmentəl]	*adj.*基础的，基本的 *n.*基本原则，基本原理
verifiable	[ˈverifaiəbl]	*adj.*能证实的
truthful	[ˈtru:θful]	*adj.*诚实的，说实话的
violate	[ˈvaiəleit]	*vt.*违犯，冒犯，干扰；违反
catastrophic	[ˌkætəˈstrɔfik]	*adj.*悲惨的，灾难的
visible	[ˈvizəbl]	*adj.*看得见的，明显的，显著的 *n.*可见物
transcend	[trænˈsend]	*vt.*超越，胜过
boundary	[ˈbaundəri]	*n.*边界，分界线
omnipotent	[ɔmˈnipətənt]	*adj.*全能的，无所不能的
visualization	[ˌvizjuəlaiˈzeiʃən]	*n.*可视化
span	[spæn]	*n.*跨度，跨距，范围
contender	[kənˈtendə]	*n.*竞争者
volatility	[ˌvɔləˈtiliti]	*n.*波动率；波动性；波动
variability	[ˌvəriəˈbiliti]	*n.*变异性；可变性
cornerstone	[ˈkɔ:nəstəun]	*n.*奠基石，基础，最重要的部分
viability	[ˌvaiəˈbiliti]	*n.*可行性，切实可行，能办到；生存能力
gigantic	[dʒaiˈgæntik]	*adj.*巨人般的，巨大的
proportion	[prəˈpɔ:ʃən]	*n.*比例；均衡；部分 *vt.*使成比例；使均衡，分摊
guzzle	[ˈgʌzl]	*vt.*狂饮，暴食；消耗
fascinate	[ˈfæsineit]	*vt.*使……着迷，使……神魂颠倒 *vi.*入迷，极度迷人的
fad	[fæd]	*n.*时尚，一时流行的狂热，一时的爱好
funnel	[ˈfʌnəl]	*vt.& vi.*把……灌进漏斗；使成漏斗状；成漏斗形；使汇集 *n.*漏斗；漏斗状物

futility	[fju:'tiləti]	*n.*无益，无用
nutshell	['nʌtʃel]	*n.*简言之，一言以蔽之
exemption	[ig'zempʃən]	*n.*解除，免除
vitality	[vai'tæliti]	*n.*时效性；动态性，灵活
criticality	[kriti'kæliti]	*n.*临界点；临界状态；紧急程度，危险程度
prioritize	[prai'ɔritaiz]	*vt.*把……区分优先次序
pragmatic	[præg'mætik]	*adj.*实际的，注重实效的
arbitrage	['ɑ:bitridʒ]	*n.*套汇，套利交易
correlate	['kɔrileit]	*vt.*使相互关联 *vi.*和……相关
incomplete	[ˌinkəm'pli:t]	*adj.*不完全的，不完善的

Phrases

big data	大数据
capitalize on	充分利用；资本化
holy grail	圣杯；无处寻觅的稀世珍宝，努力却无法得到的东西
extracted ... from	从……中抽取，从……中提取
data warehouse	数据仓库
business intelligence tool	商业智能工具
information explosion	信息爆炸，知识爆炸
be superseded by …	被……取代
wake up	活跃起来；引起注意；（使）认识到
invade into	侵入
unstructured text	非结构化文本
underlying data	源数据；基础数据；基本数据
pertain to	属于，关于，附属
in the attempt to	试图，企图
at large	普遍的；一般的；整体的
according to	依照
in a snap	立刻，马上
standard deviation	标准差，标准偏差
scale up	按比例增加，按比例提高
get fascinated with	迷上，沉溺于
in a nutshell	简言之，一言以蔽之
analysis paralysis	过度分析
be replaced with	由……代替
be aligned to	与……一致

be derived from	来自，源于
draw from…	从……抽取
consist of	构成，组成
customer contact center	客户联络中心，客户服务中心

Abbreviations

IT (Information Technology)	信息技术
IoT (Internet of Things)	物联网

Notes

[1] One has to demonstrate value that can be extracted from big or small data in order to justify the investments, whether on big data or on traditional analytics, data warehouse or business intelligence tools, whatever may be the buzzing nomenclature.

本句中，that can be extracted from big or small data 是一个定语从句，修饰和限定 value。in order to justify the investments, whether on big data or on traditional analytics, data warehouse or business intelligence tools 是一个目的状语从句，修饰主句的谓语 demonstrate。whatever may be the buzzing nomenclature 是一个让步状语从句。

[2] In the context of business operations that have not yet been touched by social media or IoT, the velocity arises from sophisticated enterprise applications that capture each and every minute detail involved in the completion of a particular business process.

本句中，that have not yet been touched by social media or IoT 是一个定语从句，修饰和限定 business operations。that capture each and every minute detail involved in the completion of a particular business process 也是一个定语从句，修饰和限定 enterprise applications。在该从句中，involved in the completion of a particular business process 是一个过去分词短语，做后置定语，修饰和限定 each and every minute detail。

[3] Since we are living in an increasingly digital world where technology has invaded into our glasses and watches, the variety of data that is generated is mind-boggling.

本句中，Since we are living in an increasingly digital world where technology has invaded into our glasses and watches 是一个原因状语从句，修饰和限定主句的谓语 is mind-boggling。在该从句中，where technology has invaded into our glasses and watches 也是一个定语从句，修饰和限定 digital world。that is generated 是一个定语从句，修饰和限定 the variety of data。

[4] Embedded in the concept of value is the need to check the viability of any project.

本句是一个表语前置的倒装句。the need to check the viability of any project 是主语，Embedded in the concept of value 是表语。正常语序应为：The need to check the viability

of any project is embedded in the concept of value.

Exercises

【Ex. 1】 根据课文内容回答问题。

1. What can insights from big data do?
2. What does velocity refer to? What are some of the factors that exacerbate this trend?
3. Why is the variety of data that is generated is mind-boggling?
4. What does veracity imply?
5. What have always existed within enterprises and have been one of the major roadblocks in the attempt to extract value from data?
6. What should relevant information be?
7. How many hours of video are uploaded every minute on YouTube according to official statistics?
8. What is one of the cornerstones of traditional statistics?
9. What kind of information needs to be prioritized?
10. Where is the word vincularity derived from? What does it imply?

【Ex. 2】 把下列句子翻译为中文。

1. I hope that this talk has given you some insight into the kind of the work that we've been doing.
2. The new systems have been optimized for running Microsoft Windows.
3. These designs demonstrate her unerring eye for colour and detail.
4. Let me make this clear: A bar chart is not analytics.
5. A good dictionary will give us the connotation of a word as well as its denotation.
6. The latest lifestyle trend is downshifting.
7. The end of an era presupposes the start of another.
8. You cannot combine structured and unstructured exception handling in the same function.
9. Finally, the practical application shows the feasibility and veracity of this approach.
10. The viability of multilayer switches depends on the protocol supported.

【Ex. 3】 短文翻译。

Cloud computing is a general term for anything that involves delivering hosted services over the Internet. These services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). The name cloud computing was inspired by the cloud symbol that's often used to represent the Internet in flowcharts and diagrams.