Debortoli, S.、Junglas, I.、Müller, O. 和 vom Brocke, J. (2016)。信息系统研究人员的文本挖掘：带注释的主题建模教程。信息系统协会 (CAIS) 的通信。

通讯
信息协会
人工智能 系统

教程 国际标准期刊号:1529-3181

面向信息系统研究人员的文本挖掘：带注释的主题建模教程

斯蒂芬·德博托利
列支敦士登大学,信息系统研究所,瓦杜兹,列支敦士登,
stefan.debortoli@uni.li

奥利弗·穆勒
哥本哈根信息技术大学,丹麦哥本哈根

艾里斯·荣格拉斯
佛罗里达州立大学商学院,
美国佛罗里达州塔拉哈西

扬·冯·布洛克
列支敦士登大学信息研究所
系统,瓦杜兹,列支敦士登

抽象的：

据估计,当今超过 80% 的数据以非结构化形式存储（例如文本、音频、图像、视频）；
其中大部分是用丰富而模糊的自然语言表达的。传统上,自然语言分析
促使人们使用定性数据分析方法,例如手动编码。然而,文本数据集的大小
从互联网上获取的信息使得手动分析几乎不可能。在本教程中,我们将讨论挑战
在信息系统研究中应用自动文本挖掘技术时遇到的问题。特别是,我们
展示通过潜在狄利克雷分配（一种无监督文本挖掘)使用概率主题建模
技术,结合 LASSO 多项逻辑回归来解释用户对 IT 工件的满意度
通过自动分析超过 12,000 条在线客户评论。对于信息系统研究人员来说，
本教程为开展文本挖掘研究以及评估研究质量提供了一些指导。
其他的。

关键词:文本挖掘、主题建模、潜在狄利克雷分配、在线客户评论、用户满意度

1 简介

随着 Web 2.0 和社交媒体的出现,互联网上非结构化文本数据的数量急剧增长,尤其是在微观层面（Gopal、Marsden 和 Vanthienen，2011）。例如,在撰写本文时,仅 Amazon.com 就提供了超过 1.4 亿条客户评论,约超过 900 万种产品,由数百万亚马逊用户撰写,时间跨度近 20 年（McAuley、Pandey 和Leskovec,2015;McAuley、Targett、Shi 和 van den Hengel,2015）。超过 3 亿的活跃 Twitter 用户平均每天生成 5 亿条推文（Twitter,2015）。大量的公开数据为定性和定量信息系统 (IS) 研究人员创造了新的机会。

传统上,自然语言数据的分析促进了定性数据分析方法的使用,例如手动编码（Quinn、Monroe、Coaresi、Crespin 和 Radev,2010）。然而,互联网上可用的文本数据集的大小超出了单个研究人员甚至研究团队的信息处理能力。尽管有关于如何在使用多个编码器分析定性数据时提高有效性和可靠性的方法指南（Saldaña,2012）,但无法完全消除研究人员对数据的主观解释所产生的偏差（Indulska，Hovorka,&Recker,2012）。

文本挖掘技术能够以可扩展和可重复的方式从大量非结构化文本数据中自动提取隐含的、以前未知的和可能有用的知识（Fan、Wallace、Rich 和 Zhang,2006 年;Frawley、Piatetsky-Shapiro 和 Matheus,1992 年）。尽管文本的自动计算分析仅触及自然语言语义的表面，但事实证明,当输入足够大的数据集时,它是一种可靠的工具（Halevy、Norvig 和 Pereira,2009 年）。在此背景下,文本挖掘为 IS 研究提供了一种有趣且互补的探究策略,可以与其他数据分析方法（例如回归分析)很好地结合,或用于对从更传统的数据收集和分析方法中获得的研究结果进行三角测量。特别是,自动文本挖掘使 IS 研究人员能够 (1) 克服手动定性数据分析方法的局限性,以及 (2) 获得原本无法发现的见解。以下两个示例将作为说明。

在一项受到公众和学术界广泛关注的研究中,米歇尔等人。 (2011) 通过计算 Google 图书中单词出现的年度相对频率来调查文化趋势。这种简单的统计分析应用于超过 500 万本数字化图书,产生了一些有趣的见解。

例如,该研究发现,随着时间的推移,通过与某些技术（例如无线电、电话)相对应的词频来衡量,创新的扩散正在以越来越快的速度加速。19 世纪初,一项技术从发明到广泛采用平均需要 66 年,而到 1900 年左右,平均采用时间下降到 27 年。

另一个说明性的例子来自社会心理学领域。 Pennebaker 及其同事（Pennebaker,2011;Tausczik & Pennebaker,2010)开发了语言查询和字数统计 (LIWC) 工具,该工具允许通过计算不同功能词（例如代词、冠词、介词）。功能词在自然语言中很常见,但读者和程序员通常不会有意识地关注它们,而是关注实词（例如名词、动词）。然而,事实证明,功能词使用模式的细微差异是许多心理状态的重要预测因素。例如,研究人员使用 LIWC 来检测在线客户评论中的欺骗行为,因为与诚实的评论者（Ott、Choi、Cardie)相比,撒谎的评论者倾向于使用更多的人称代词和“我”词,以及不太具体的术语（例如数字）。 ,&汉考克,2011）。

在本教程中,我们讨论在信息系统研究中应用自动文本挖掘技术时遇到的挑战。应用文本挖掘需要多个领域的技能,包括计算机科学和语言学;并不是每个信息系统研究人员都熟悉这些领域的概念和方法。虽然存在大量关于特定文本挖掘算法的思想和方法的技术文献,例如主题建模（Blei,2012)或情感分析（Pang & Lee,2008）,但这些出版物很少涉及“操作方法”应用文本挖掘作为（信息系统)研究的探究策略的各个方面。我们特别关注概率主题建模,作为一种归纳发现大量文本（语料库)中运行的主题的技术,例如来自网络的用户生成的内容。除了概述主题建模的基础之外,我们

通过介绍可用的软件工具并借助在线客户评论领域的综合示例展示其应用来说明其具体用途。

本文的其余部分的结构如下。首先,我们概述了一般分析大型文本语料库的方法,然后特别深入研究概率主题建模。其次,我们讨论主题建模研究中遇到的典型挑战,并概述克服这些挑战的潜在方法。第三,我们介绍应用主题建模的工具,并借助集成示例说明其应用。最后,我们通过讨论所提出方法的局限性来得出结论。

2 背景

2.1 分析大型文本语料库

文本分析(手动和自动)中最基本的任务之一是文本分类,即将文本块(例如电子邮件、社交媒体评论、新闻)分配到一个或多个类别(例如,垃圾邮件或无垃圾邮件、正面或负面情绪、商业或政治或体育新闻)。

有不同的文本分类方法,每种方法都与一定的假设和成本相关(见表 1)。

社会科学研究中用于文本分类的传统方法是手动编码(Berg & Lune,2011)。编码的目的是将原始数据区分并组合成类别,以捕获数据块的本质含义(Miles & Huberman,1994)。存在各种编码技术(参见 Saldaña,2012);然而,在最基本的层面上,可以区分自下而上和自上而下的方法(Urquhart,2012)。作为自下而上编码的一部分,代码由数据(即单词和短语)建议 无论现有理论如何(Urquhart,2012)。编码员应以开放的心态处理分析任务,不要对数据强加先入为主的观念。相反,对于自上而下的编码,编码者使用从文献中得出的预定义编码模式,并将数据分配给这些代码(Urquhart,2012)。后一种编码方式有时也与代码实例计数结合使用,例如在进行系统内容分析时。

手工编码有许多优点,例如人类具有无与伦比的理解自然语言含义的能力,或者在文本特征和类别之间建立高度复杂且偶然的映射的可能性(Quinn 等人,2010 年)。然而,它也存在许多局限性。首先,它容易受到人类主观性的影响,因此不同的编码员可能会得到不同的结果(Urquhart,2001 年)。为了克服这些对有效性和可靠性的威胁,已经提出了各种实现主观可验证性的策略,包括使用代码本、拥有多个编码员或进行编码员间可靠性测试(Indulska 等人,2012 年)。然而,这些策略的适用性受到手工编码的第二个限制。手工编码在所需工时方面成本高昂,并且需要大量的领域知识(Quinn 等人,2010 年)。为了克服这些限制,研究人员通过应用基于词典或机器学习的算法开发了用于文本分析的计算机辅助方法。

基于字典的文本分类依赖于专家组装单词和短语列表,这些列表可能表明特定类别中文本块的成员资格(Quinn 等人,2010)。使用该词典,计算机可以自动解析大量文本并确定文本单元的分类。基于字典的分类仅适用于预定义类别并且文本特征(即单词和短语)与类别之间的映射是预先已知的并且可以被编码的情况。换句话说,基于字典的方法只能应用于自动化自上而下的手动编码。许多情感分析方法将文本分为积极或消极类别,例如流行的 SentiStrength (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010),使用基于字典的文本分类方法。

自动化自上而下的手动编码的第二种方法需要使用监督学习方法。与之前一样,类别是已知的且是预定义的,但是文本特征和类别之间的映射并不明确已知(Quinn 等人,2010)。使用一组手动分类的文档作为训练示例,然后可以应用监督机器学习算法来自动检测单词的使用与其类别分配之间的关系。然后可以使用学习到的模式对新的或未见过的文本进行分类。电子邮件垃圾邮件过滤是有效使用监督学习方法进行文本分类的典型示例,例如,通过挑选金融广告中的单词“$$$”、“credit”和“free e”。

最后,无监督的文本分类机器学习方法试图在不存在预定义分类的文本中找到隐藏的结构（Quinn 等人,2010 年）。无监督学习方法（例如,聚类、降维)利用文本的特征归纳发现潜在类别并将文本单元分配给这些类别。这种归纳方法与手动自下而上的编码或扎根理论方法中已知的开放式编码（Berente & Seidel,2014 年)相当。

无监督文本分类方法与手动编码相比具有一些明显的优势：(1)在预分析和分析阶段,它们只需要很少的人工干预和实质性知识；(2)它们产生可重复的结果,因为它们不受人类主观偏见的影响；(3)当今的算法和计算系统可以处理大量文本,即使大型编码团队也无法分析这些文本。缺点是,无监督方法需要大量的后分析阶段,这通常非常耗时,因为研究人员必须理解自动生成的归纳分类。

表 1. 不同文本分类方法的假设和成本（改编自 Quinn 等人,2010 年）

	手动编码（自下而上）	手动编码（自顶向下）	词典	监督机器学习	无监督机器学习
假设					
类别是预定义的	不	是的	是的	是的	不
相关文本特征已知	是的	是的	是的	是的	是的
文本特征和类别之间的映射是已知的	不	不	是的	不	不
成本					
预分析成本					
概念化所花费的人时	低的	高的	高的	高的	低的
实质性知识水平	低的	高的	高的	高的	低的
分析成本					
每篇文本花费的人时	高的	高的	低的	低的	低的
实质性知识水平	缓和	缓和	低的	低的	低的
后期分析成本					
口译人时	缓和	低的	低的	低的	缓和
实质性知识水平	高的	高的	高的	高的	高的

2.2 概率主题建模

下面,我们将更详细地讨论概率主题建模,这是一种无监督的机器学习方法。无监督的机器学习方法仅依赖于底层文本数据的少数假设,并且需要的数据分析成本极低,这使得研究人员能够将其应用于各种来源和大量数据。

许多用于文本分类的无监督学习方法的基本思想植根于语言学的分布假设（Firth,1957;Harris,1954）,指的是“出现在相同上下文中的单词往往具有相似含义”的观察结果（Turney &潘特尔,2010,第 142 页）。例如,报纸文章中同时出现的“进球”、“球”、“前锋”和“犯规”等词可以被解释为共同类别（即“足球”)的标记,并用于相应地对文章进行分组。

在过去的几十年里,已经开发和扩展了几种用于无监督文本分类的分布式方法。 IS 研究中最常用的方法包括潜在语义分析 (Landauer, Foltz, & Laham, 1998)、潜在狄利克雷分配 (Blei, Ng, & Jordan, 2003) 和 Leximancer (Smith & Humphreys, 2006)。潜在语义分析 (LSA) 通过应用奇异值分解来降低术语文档矩阵的维数,从而提取分布式单词使用模式。由此产生的潜在语义因子与因子分析或主成分分析的输出有许多相似之处,通常被解释为主题（Landauer 等,1998）。

LSA 是计算语言学领域的一项突破性进展,但由于计算出的因子载荷通常没有明确的解释,因此存在可解释性问题。为了克服这些缺点,概率 LSA (pLSA) (Hofmann,1999) 和潜在狄利克雷分配 (LDA) (Blei 等人,2003;Blei,2012) 已作为经典 LSA 思想的扩展而开发。在这两种方法中,文档与主题之间以及主题与单词之间的关联都表示为概率分布,可用于进一步的统计分析。例如,估计的概率分布可以按文档元数据分组和聚合,或用作回归或分类模型中的预测因子。此外,还存在各种应用分布假设的商业工具。例如,Leximancer (http://www.leximancer.com) 将无监督的单词共现模式提取与概念映射和直观可视化相结合 (Smith & Humphreys, 2006)。然而，Leximancer 的算法和数据结构已申请专利,因此很少有记录。

下面,我们将详细描述LDA概率主题建模的应用。我们选择 LDA 的原因有以下三个:(1) LDA 是开创性的 LSA 思想的演变,两种方法都在学术研究中广泛使用[1]
, (2) 大多数统计编程语言（包括 R、Python、Java)都存在大量免费和开源 LDA 软件库,以及 (3) LDA 从文本中提取语义上有意义的主题并根据这些主题对文本进行分类的能力已在许多实证研究中得到验证（例如 Boyd-Graber、Mimno 和 Newman,2014 年;Chang、Boyd-Graber、Gerrish、Wang 和 Blei,2009 年;Lau、Newman 和 Baldwin，2014 年;Mimno、Wallach、Talley、Leenders 和 McCallum,2011 年）。

LDA 的核心思想由 Blei 等人 (2003) 首次提出,是一种虚构的生成过程,假设作者撰写D篇文档时,首先选择T 个主题的离散分布,然后从每个主题的典型离散单词分布中抽取W 个单词（见图 1）。换句话说,文档由一组固定主题的概率分布定义,而每个主题又由一组有限词汇的概率分布定义。虽然假设所有文档都是从同一组固定主题生成的,但每篇文档以不同的比例展示这些主题,可能从 0%（如果文档未能完全讨论某个主题)到 100%（如果文档只讨论某个主题)不等。LDA 算法的计算任务是根据观察到的每个文档单词出现次数估计隐藏的主题和单词分布。

这种估计可以通过采样方法（例如,吉布斯采样)或优化方法（例如,变分贝叶斯)来完成。

[1] 在撰写本文时,在 Google Scholar 上搜索“潜在语义分析”产生了超过 32,000 个结果,而搜索“潜在狄利克雷分配”产生了超过 19,000 个结果。

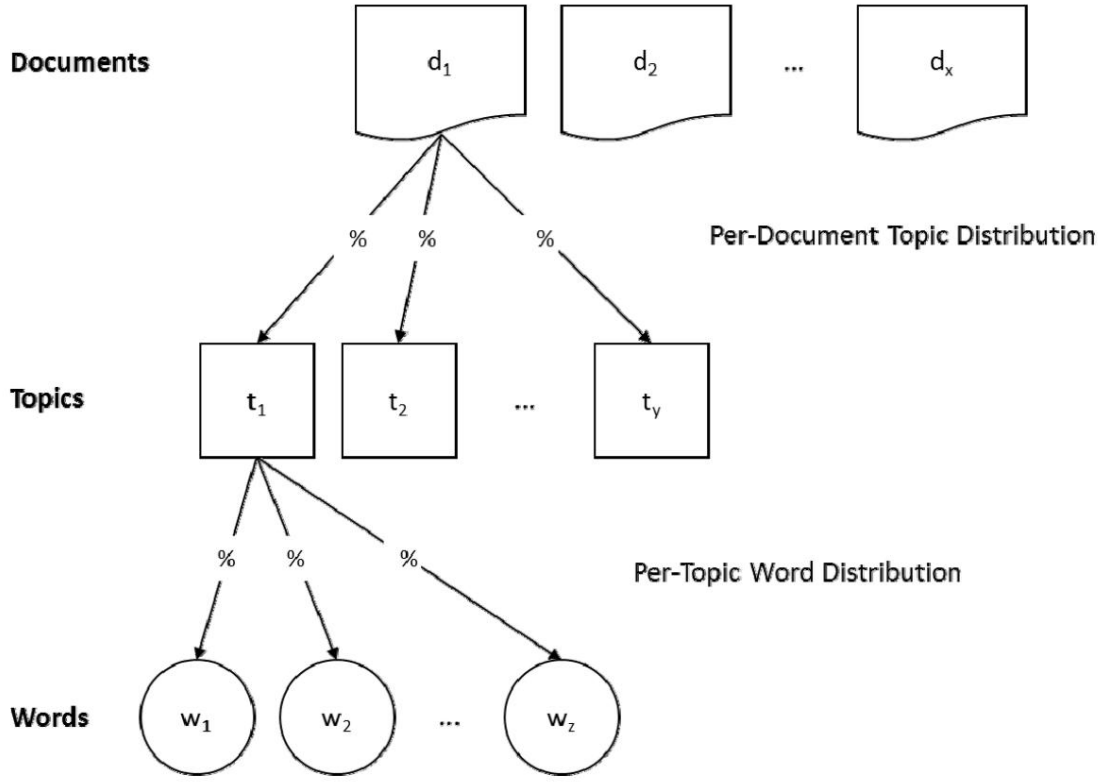

图 1. LDA 示意图

图 2 使用关于“Fitbit Flex”2设备的典型在线客户评论及其主题分布,以及六个主题及其词汇分布,说明了 LDA 背后的基本思想。典型评论涵盖三个不同程度的主题,即主题 3 (55%)、主题 2 (35%) 和主题 6 (10%);其他主题不存在 (0%)。每个主题依次由词汇分布表示。例如,主题 3 为“体重”(8%)、“损失”(5%) 和“磅”(4%) 等词分配了高可能性,表明该主题涵盖了使用 Fitbit 设备导致的减肥效果。另一方面,主题 2 具有高度可能的词,如“礼物”(10%)、“爱”(7%) 或“圣诞节”(7%),表示 Fitbit 设备已作为礼物赠送或收到。最后,主题 6 中最可能的词是“app”(12%)、“iphone”(8%)和“sync”(3%),指的是 Fitbit 和相应的 iPhone 应用程序之间的同步。

[2] Fitbit Flex 是一种可穿戴技术,可全天候跟踪和分析个人健康和健身数据。

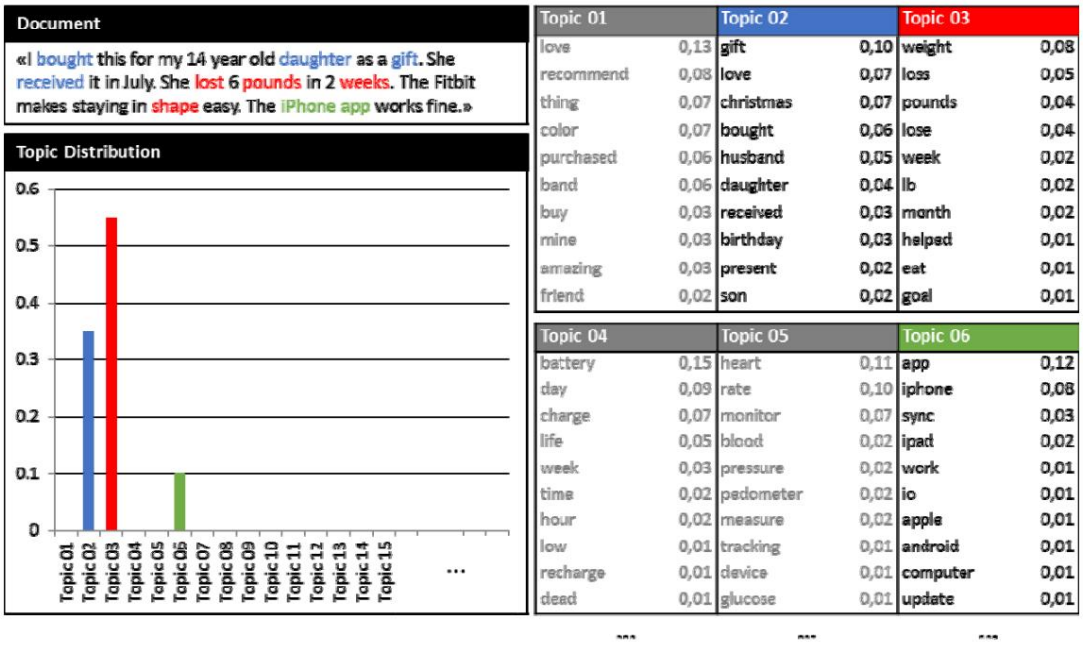

图 2. LDA 示例

应用主题建模的 3 个实际挑战

现在,我们转向应用主题建模作为自动分析大量定性数据的方法时所面临的实际挑战。这些挑战大致反映了典型研究过程的阶段。

3.1 挑战 1:从网络获取数据由于主题建模只有在输入足够大的数据集（n>1,000)时才能

产生有效的结果,因此它通常不用于分析研究人员自己收集的数据（例如,访谈记录、实地笔记）,而是用于分析由大量人群制作的文本（例如,来自社交媒体网站的用户生成内容、科学界撰写的研究文章)并作为互联网资源提供。广义上讲,有三种方法可以从互联网源中提取文本数据：(1)通过应用程序编程接口 (API)，(2)通过网络爬虫,或（3)通过文件下载。

API 是由数据提供商提供的编程数据访问点,用于以受控方式（例如,通过限制可访问的数据范围和数量)和结构化格式（例如,通过使用 XML 或 JSON 等标记语言)提供可重复使用的内容。虽然通过 API 提取数据通常可以确保高水平的数据质量,但提供商很少通过 API 公开其数据的全部广度和深度在最好的情况下,他们要求用户为此类数据请求付费。文本挖掘研究中常用的 API 是 Twitter API,因为它受到的限制很少。虽然 Facebook、LinkedIn 或 Google+ 等社交网络的 API 仅允许访问有关“好友”的数据,但 Twitter API 可以访问有关网络所有成员的数据。

网络爬虫提供了另一种从网络中提取数据的方法。这些自动化程序遍历网络的拓扑结构并下载相关页面和超链接（Liu,2011）。它们不是由数据提供者操作的,而是由数据消费者或中介机构操作的。网络爬虫使用简单的自然语言处理启发式方法（例如正则表达式)解析或“抓取”页面内容。由于网页通常包含大量噪音（例如 HTML 标签、广告横幅）,许多网络爬虫试图过滤掉不相关的元素;或多或少成功。除了提取内容外,网络爬虫还能够通过构建相互关联的参与者（例如网页、用户)的图表来捕获网络的底层链接和社会结构。总体而言,网络爬虫为研究人员提供了很大的灵活性。例如,研究人员可以通过使用一组

搜索词或种子 URL。然而,这种灵活性是有代价的。爬虫通常需要深入编程,而且收集的数据质量可能达不到分析所需的水平。

最后,研究人员还可以使用可从网络下载的开放数据存储库。开放数据包括任何人都可以自由使用、重复使用和重新分发的数据,只需遵守归属要求（OKF,2012）。在过去的几年里,政府（例如,http://www.data.gov/）、非营利组织（例如,http://en.wikipedia.org/wiki/Wikipedia:Database_download）、研究机构 https ://snap.stanford.edu/data/）（例如,http://www.yelp.com/dataset_challenge)已经建立了开放数据存储库,其中包含 IS 研究人员可能感兴趣的大量文本数据集合。大多数这些数据集都是集成和整理的,这简化了访问并确保了数据质量。

（例如， 和 私人的 组织

在选择数据收集方法时,必须考虑可以捕获的时间范围。数据快照是最容易实现的,大多数 API 和网络爬虫都支持它。此外,许多开放数据集是横断面调查的结果,因此代表快照。收集纵向数据更成问题。虽然一些 API 和开放数据文件提供历史数据,但网络爬虫是定期批量运行的,无法捕获网页的全部波动性。最后,最难捕获的时间范围是实时数据。只有所谓的流式 API（例如 Twitter 的 Firehose API)能够提供此功能,它允许实时访问完整的推文流（目前约为每秒 4,000 条推文）。但是,Firehose 访问权限仅限于选定的合作伙伴组织。

3.2 挑战 2:准备好要分析的数据

自然语言数据的特点是缺乏明确的结构和高比例的噪音。因此,在几乎所有情况下,数据都需要经过广泛的预处理阶段,然后才能通过主题模型进行统计分析。尽管在研究结果的展示中很少强调数据准备步骤,但它们通常需要整体工作量的 45% 到 60%（Kurgan & Musilek,2006）。

作为第一步,应执行高级探索性数据分析 (EDA),以便对数据集有一个初步了解并识别潜在的数据质量问题。除了计算汇总统计数据（例如,数据集中的文档数量、每个文档的平均单词数)之外,研究人员还应使用可视化。例如,词频图提供了有关所需数据清理和自然语言处理步骤的宝贵信息。同样,在时间线上绘制文档的时间戳可以快速发现缺失数据,可能指向数据收集过程中的错误,或时间趋势和季节性模式。如果获得的文本文档包含将在以后的分析中使用的数字信息,例如作为回归分析中的独立或因变量,则还应绘制它们以可视化其分布并识别潜在异常。

在探索了整个数据集之后,需要在文档级别检查和预处理所获得的文本。典型的准备步骤包括:数据清理、数据构建、数据格式化和自然语言处理。

数据清理是将自然语言数据准备好进行分析的基本步骤之一,通过删除重复项和噪音来做到这一点。由于文本挖掘研究中使用的许多数据集都是二手数据,因此它们包含“不干净”数据的可能性相当高。例如,Twitter 等在线社交网络上的帖子可能包含重复记录（转发、垃圾邮件）,而网络爬虫收集的数据可能充满 HTML 标签形式的噪音。如果不加以处理,重复项和噪音不仅可能导致偏差,还可能导致不正确的结果。

数据构建需要派生新属性和/或记录。派生属性的示例包括涉及多个属性的计算（例如,通过从当前日期中减去创建日期来计算在线评论的持续时间)或单个属性转换（例如,用地理位置标记评论）。创建新数据属性的必要性在很大程度上取决于后续的数据分析程序。为了确保透明度,应提供派生新属性的精确公式。

在这些初始步骤之后,大多数文档需要（重新)格式化以允许使用特定的分析工具或方法进行处理。重新格式化的范围可以从单个值的简单更改（例如,删除非法字符或更改字符编码)到复杂的数据模型转换。为了

例如,通过 API、Web 爬虫或下载提取的数据大多以平面文件（例如 CSV)或分层数据模型（例如 XML、JSON)表示;对于分析和存储,将此类数据转换为关系（例如 SQL 数据库)或键值数据模型（例如 NoSQL 数据库)可能会很有用。理想情况下,原始数据模型及其各种来源以及用于分析目的的最终数据模型应详细说明并充分记录。

在文档级预处理之后,单个文档集将经历许多低级自然语言处理 (NLP) 步骤,例如标记化（即将文档拆分成句子,将句子拆分成单词）、n-gram 创建（即创建 n 个连续的单词:1-gram 为“fast”、“food”或“chain”;2-gram 是两个 1-gram 的连接,例如“fast food”;3-gram 由三个 1-gram 组成,例如“fast food chain”）、停止（即删除常见或不具信息量的单词）、词性过滤（即根据词性识别和过滤单词）、词形还原（即将单词还原为其词典形式,例如将名词的复数形式还原为单数形式,将动词还原为一般现在时）、词干提取（即将单词还原为其词干)以及创建文档集合的结构化数字表示（例如创建向量或矩阵表示）(Miner et al. al.,2012）。这些转换的共同目标是消除噪音,并逐渐将定性文本数据转换为适合后期统计分析的数值表示。不幸的是,没有简单的方法来选择合适的自然语言预处理步骤组合。

这在很大程度上取决于研究目标及其基础数据集。不过,可以应用一些策略,如删除停用词、文本规范化和搭配发现（Boyd-Graber 等,2014）,在一定程度上缓解这一困境。

为了识别停用词,生成词频列表（即文本语料库中每个单词出现的次数)是一种有用的方法。例如,在研究有关 Apple iPhone 的在线客户评论时,“Apple”和“iPhone”这两个词的频率计数很高,但对分析没有特别的价值,因此可以删除。其他方法,例如词数的 tf-idf 加权,也可以用于自动过滤无信息的术语（Salton & McGill,1983）。

文本规范化通常包括将所有字符转换为小写以及对每个单词进行词形还原。例如,“dog”、“Dog”、“dogs”和“Dogs”这些词都将转换为“dog”,从而只产生一个单词,而不是四个不同的标记。通过应用词干提取,文本规范化的概念甚至可以得到进一步的推进。例如,在词干提取中,“analyze”和“analysis”这两个词将缩减为“analy”。然而,这种词数的减少可能会导致另一个问题（Evangelopoulos、Zhang 和 Prybutok,2012 年） 这将使我们无法区分“analy”在给定上下文中指的是名词还是动词。

最后,发现单词的搭配或多单词表达（即 n-gram | n > 1)有助于找到单词的正确含义。例如,“房子”一词在给定上下文中表示一件事,但“白宫”一词在大多数情况下具有完全不同的含义。因此,建议执行 n > 1 的 n 元语法分析,特别是在稍后由人类解释结果的情况下。

3.3 挑战3:拟合和验证主题模型

将主题模型拟合到文档集合可能具有挑战性。 LDA 算法对其参数的变化和输入数据的变化很敏感,例如通过不同的数据准备过程引入的变化。

最关键的 LDA 参数是要提取的主题数量（Blei 等人,2003 年;Boyd-Graber 等人,2014 年）。当选择太多主题时,算法可能会发现过多的主题,这些主题之间的区别很小（例如,主题在写作风格上不同,但在内容上没有差异）,而选择太少的主题可能会不必要地限制主题建模的探索潜力。因此,最佳实践是改变主题数量,并根据研究目标评估生成模型的质量。如果研究目标是创建一个可由人类通过生成大量文本的定量表示来解释的主题模型,那么要选择的主题数量通常较低,可能在 10 到 50 之间。如果,相反,主题模型旨在充当另一个统计模型（例如回归、分类、聚类)的输入,并且人类可理解性不是重要因素,要选择的主题数量由模型拟合决定,而不是由模型拟合决定它的可解释性;这里,主题的数量可能在 30 到 100 之间,甚至更多。

必须选择作为 LDA 设置一部分的另一组参数是超参数 α
和 β,分别控制每个文档主题分布和每个主题单词分布的形状。大的α导致广泛的主题分布（即,文档包含许多主题）,而大的β导致广泛的单词分布（即,主题包含许多单词）。相反,较小的α和β值会导致更稀疏的分布,即假设文档仅包含很少的主题,并且假设主题仅包含很少的单词。尽管大多数主题建模工具允许用户明确定义 α 和 β,但常见的做法是使用既定的标准值（例如,一除以主题数量）,或依赖优化技术,如 Wallach、Mimno 和 McCallum 所描述的(2009),自动确定适当的值。

一旦计算出主题模型,研究人员就必须解释结果。出于演示目的,LDA 结果通常以列表的形式显示,显示每个主题前 n 个最有可能的单词（Ramage、Rosen、Chuang、Manning 和 McFarland,2009 年）。虽然这是一种直观的呈现方式,但它可能会使调查人员产生偏见,因为每个主题实际上是语料库中整个词汇的分布。因此,在解释主题的含义时,建议研究人员检查实际的单词概率（而不仅仅是它们的排名）,以及与每个主题密切相关的文档（可以通过每个文档的主题分布获得）。

通常,研究人员会为主题分配描述性标签,以帮助读者解释主题。与文本的手动编码一样,建议至少由两名独立研究人员对主题进行解释和标记。

验证主题模型可能很困难。由于其无监督性质,目前还没有关于如何评估主题建模结果的基本规则或黄金标准。在计算机科学界,主题模型通常通过测量其在后续任务（例如信息检索、回归、分类)中的表现来评估,或者通过测量在给定语料库上训练的模型与未见过的或保留的文本的匹配程度来评估（有关概述,请参阅 Wallach、Murray、Salakhutdinov 和 Mimno,2009 年）。这两种方法都假设主题模型由另一种算法使用。然而,实验表明,具有高预测准确性的主题模型不一定具有良好的人类可解释性（Chang 等人,2009 年）。

对于旨在由人类解释的主题模型,Boyd-Graber 等人 (2014) 提出了两个指导性问题来评估其语义质量：

1. 各个主题是否有意义、可解释、连贯且有用？
2. 为文档分配主题是否有意义、适当且有用？

对单个主题（问题 1)的可解释性的常见威胁是多方面的（Boyd-Graber 等,2014）。太多的常用词　或者太多的特定词（例如,名称、数字）　会导致主题过于宽泛或过于具体,可能会阻止研究人员对语料库获得更深入的理解。调整停用词列表并重新运行分析可能有助于解决这些问题。

低质量主题的另一个原因是所谓的混合主题。虽然单词放在一起时没有意义,但它们包含单词的子集,这些子集放在一起时完全有意义。

换句话说,混合主题包含多个主题,应该进行拆分。对于相同的主题,情况正好相反,其中算法提出两个语义上等效的主题。通过增加或减少要提取的主题数量,可以避免混合主题和相同主题。

最后,总是有可能遇到无意义的主题。例如,如果文档表现出特定的结构模式和/或具有共同的写作风格和词汇,则可能会出现这样的主题。例如,一组经常包含单词“图”和“表”的研究论文可能会导致算法根据这些词生成主题。虽然将这些词添加到算法的停用词列表中似乎是一个可行的选择,但它很可能会影响其他主题的质量。

因此,从进一步的分析中排除这些主题通常是最好的解决方案。

直到最近,研究人员才开始开发一些定量标准,通过将算法的单词分配与人类用户的单词分配进行比较（Ramage 等,2009)或通过测量主题的统计属性（Boyd-Graber)来评估各个主题的语义质量。等人,2014）。

例如,Chang 等人（2009)提出的单词入侵任务旨在量化主题的语义连贯性。在这个任务中,向人类评估者展示六个随机排序的单词。五个单词是从给定主题中最可能的单词中抽取的,一个单词

入侵者 从语料库的词汇表中随机选择。其理念是,对于语义连贯的主题,人类判断者应该能够轻松发现入侵者。例如,大多数人会将单词“苹果”识别为由单词{狗、猫、马、苹果、猪、牛}定义的主题中的入侵者（Boyd-Graber 等人,2014 年）。相比之下,在语义不连贯的单词列表{桌子、天空、苹果、黄色、城市、咖啡}定义的下一个主题中,单词“咖啡”很难被识别为入侵者。

除了人工测量外,主题连贯性的测量也可以采用自动化方式进行（Lau 等人,2014 年;Mimno 等人,2011 年;Newman、Lau、Grieser 和 Baldwin,2010 年）。大多数自动化方法将主题中最常用的单词与已知具有高语义连贯性的文本（如维基百科或报纸文章)进行比较。其理念是,高度连贯的主题的单词（例如,{狗、猫、马、苹果、猪、牛})应该经常在参考文本（例如,关于动物的维基百科文章)中同时出现;如果没有,则表明语义连贯性低。

可以应用类似的逻辑来验证主题与文档的分配（问题 2）。使用主题入侵任务,向人类评估者呈现一份随机文档,提供四个主题选择,每个选择由其前 n 个最有可能的单词表示,可以评估主题与文档分配的有效性。其中三个主题是与所讨论的文档具有很高可能性的主题,一个主题是随机选择的（Chang 等人,2009 年）。测量人类编码员识别入侵者主题的能力可以指示 LDA 算法所做的文档主题分配的质量。

3.4 挑战#4:超越描述

默认情况下,主题模型本质上是描述性的,即它们代表大型文档集合的定量摘要。特别是对于探索性研究,描述性模型通常是研究的主要目标（例如,Debortoli 等人（2014)从招聘广告中得出的能力分类法）。

由于主题模型中的所有关联都以概率表示,研究人员不仅可以展示相关主题以及选定的单词和文档分布,还可以按不同的文档元数据（例如按作者、地理位置、时间)对主题概率进行分组和汇总。例如,这允许按流行程度对主题进行排序,比较其在特定子组中的流行程度,或跟踪主题随时间的演变（有关示例,请参阅 Grimmer & Stewart (2013)）。

除了描述目的之外,主题模型还可用于解释或预测目的（Blei 等人,2003 年）。为此,估计的每个文档主题概率被用作回归或分类模型中的独立变量或预测因子。例如,Müller 等人（2016 年)使用超过 100 万条关于视频游戏的在线客户评论的概率主题分配来构建一个统计模型,该模型能够预测新评论或未见过的评论的有用性。

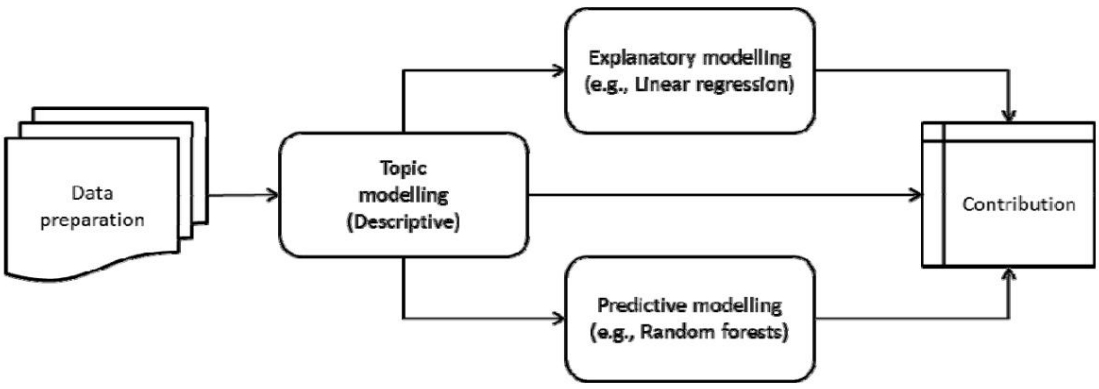

图 3. 主题建模与解释和预测建模之间的关系

如前所述,研究的目的（即描述、解释、预测)对主题模型拟合具有重要意义。旨在描述或打算将主题模型输入到后续解释模型中进行假设检验的研究人员倾向于应用更细粒度的主题模型（例如,10-50 个主题）,以便能够完整且易于理解地呈现其结果。在

相反,当研究的目的是预测时 过程和结果的可理解性不那么重要 文档的高维表示（例如,100 多个主题)与非线性回归或分类技术（例如,高度准确但“黑箱”随机森林模型)相结合,已证明可以产生最准确的结果。然而,这些模型和技术很难（甚至不可能)被人类理解,因此对于描述和解释目的来说用处不大（有关更详细的讨论,请参阅 Martens 和 Provost (2014)）。

最后,为了超越给定语料库的纯定量描述,利用现有理论和文献至关重要。实现此目的的一种方法是尝试将自动识别的主题与已知的理论构造进行映射,以便将它们置于其法理网络中。与主题标记方法类似,多个研究人员应该参与此主题构造映射任务。为此,为了得出有效的结论,深入了解感兴趣的领域及其理论基础至关重要。向所有参与编码人员提供可能发现的理论构造的定义列表以建立共同理解也可能会有所帮助。

如果某个主题与现有构造不对应,研究人员可能想要对其本体进行理论化。

4 主题建模示例研究

在本节中,我们使用在线客户评论作为示例数据源,说明主题建模与解释性回归分析相结合的实际应用。说明性示例的呈现是根据 CRISP-DM（数据挖掘的跨行业流程)框架松散地构建的,该框架包括业务理解（我们将其重命名为研究问题）、数据理解、数据准备、建模、评估阶段,以及部署（我们将其重命名为解释）（Shearer,2000）。

4.1 研究问题

我们进行说明性文本挖掘研究的目的是通过挖掘评论的文本和非结构化部分来解释用户对消费电子产品的满意度（由其星级定义）。我们的方法是基于这样的直觉:在线客户评论中某些主题的出现会对相应的星级产生重大影响。我们选择了“Fitbit Flex 无线活动和睡眠腕带”（https://www.fitbit.com/flex)作为示例产品,这是早期的可穿戴技术之一,用于全天候跟踪和分析个人健康和健身数据。

4.2 数据理解在线客户评论被定义为“同

行在公司或第三方网站上发布的产品评价”（Mudambi & Schuff,2010）。除了自由形式的文本评论外,评论通常还包含数字产品评级（通常从 1 到 5 星)以及其他元数据（例如,评论者姓名、评论日期、有用性投票）。亚马逊是世界上最大的互联网零售商,也是最大的在线客户评论来源之一（Business Wire,2010）。对于“Fitbit Flex”设备,亚马逊上有超过 12,900 条客户评论（截至 2015 年 5 月）,涵盖了两年多来客户对该产品的反馈。

由于大多数电子商务平台不提供访问客户评论的 API,因此通过网络爬虫收集评论通常是首选方法。为了完成本教程,我们开发了一个网络爬虫,用于捕获亚马逊上“Fitbit Flex”的所有历史产品评论。我们使用了 Python 包“Beautiful Soup”（http://www.crummy.com/software/BeautifulSoup/）,该包旨在从 HTML 文件中提取数据。从亚马逊下载评论后,我们将其格式化为 JSON（JavaScript 对象表示法)对象列表,以与我们用于主题建模的文本挖掘工具兼容。图 4 显示了 JSON 格式的示例性客户评论。除了文本评论外,它还包含其他元数据,例如星级（1 到 5 之间）、作者（匿名)和评论日期。总的来说,我们抓取了 12,910 条评论,时间跨度超过三年,从 2012 年 3 月到 2015 年 5 月。

```
{
        "rating": 2.0,
        "author": "Anonymous",
        "text": "Disappointed. It came with only one band and it said that it would come with two, both the large and small.",
        "date": "2015-04-25"
}
```


图 4. JSON 格式的在线客户评论示例

下一步,我们进行了探索性数据分析。计算和绘制描述性统计数据,例如评论数量 (12,910)、单词数量 (457,239)、唯一单词数量 (4,556) 和总体单词频率,提供了数据集的初步概述。例如,初始词频图显示（图 5）,冠词和代词等功能词在语料库中占主导地位。由于这些词意义不大,我们决定在后续的数据准备阶段将其删除。

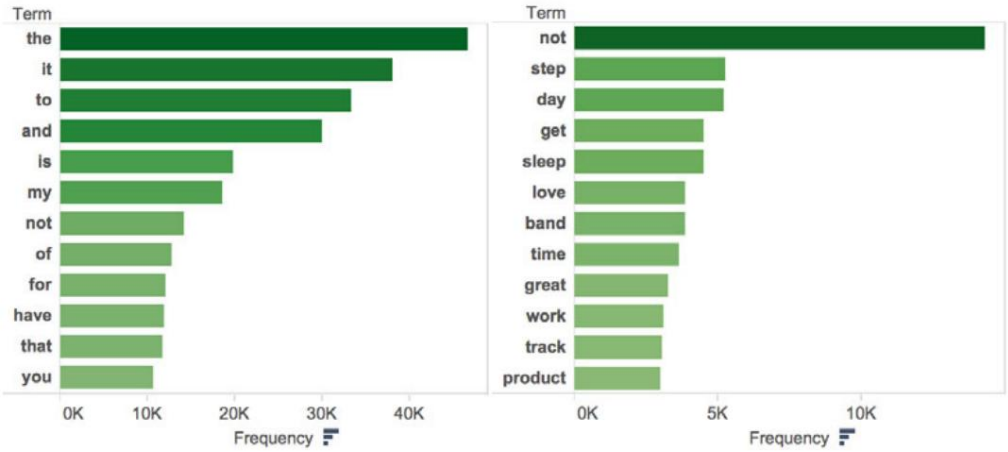

图 5. 数据准备之前（左)和之后（右)的词频图

每个客户评论的元数据的存在使我们能够绘制评论沿时间维度的分布（图 6）。一个有趣的观察是,评论数量在 12 月最后一周和 2015 年 2 月中旬激增,这可能表明“Fitbit Flex”设备是受欢迎的圣诞节和情人节礼物。

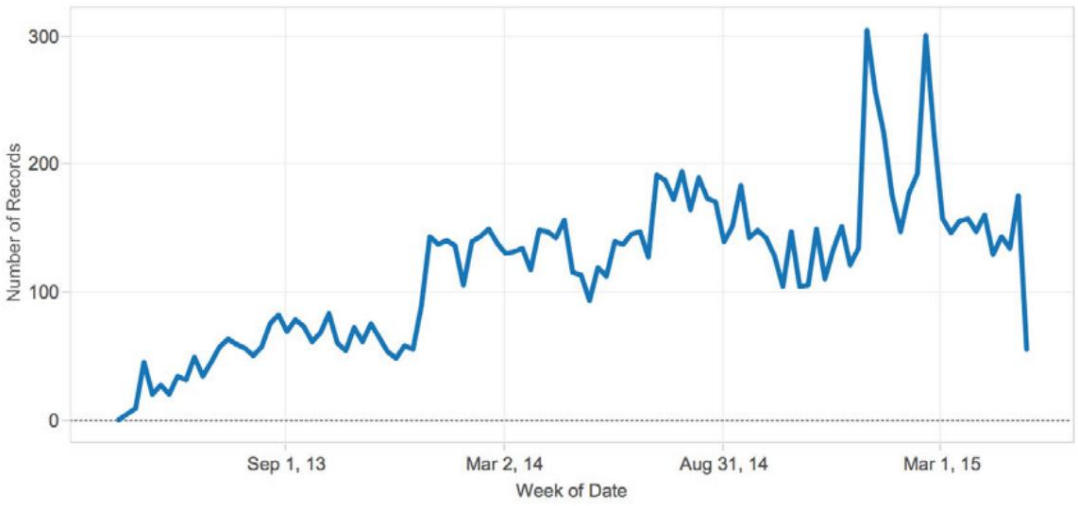

图 6. 随时间变化的评论数量

绘制随时间变化的平均星级评级图支持了用户对设备持续满意的假设（图 7），平均为 3.64（满分 5 颗星）。直方图显示了星级评分的 J 形分布（图 8），这是在线客户评论的常见现象，由购买和代表性不足偏差引起（Hu、Zhang 和 Pavlou，2009 年）。

图 7. 随时间变化的平均星级评分

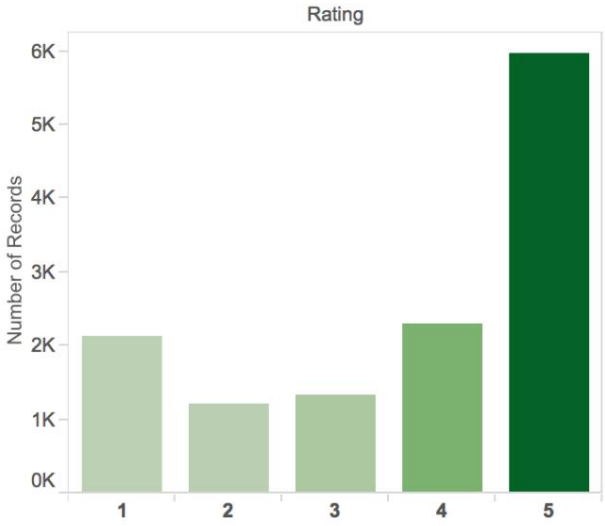

图 8. 星级的 J 形分布

4.3 数据准备、建模和评估

如前所述，数据准备过程会对主题建模结果的质量产生重大影响，在文本挖掘研究中，数据准备、建模和评估之间的反复进行非常常见。因此，以下将结合这三个步骤进行报告。所有自然语言处理和主题建模步骤均使用基于云的工具 MineMyText.com 执行，结果可在https://app.minemytext.com/fitbit 上公开访问。

我们首先进行了一系列准备-建模-评估循环，以确定从文档集合中提取的适当主题数量。我们测试了不同的替代方案，包括

20 到 100 个主题（步长为 10），并定性评估了所得主题的凝聚性。我们确定 50 个主题是最佳解决方案，因为更细粒度的主题模型（50 到 100 个主题）产生的近似重复主题数量越来越多，而更粗粒度的模型（20 到 50 个主题）无法明确区分主题。

将主题数量设置为 50 后，我们尽可能地清除了评论中的噪音。其中包括：

1. n-gram 标记化，即将文档拆分为单个单词（即 1-gram：例如“产品”、“爱”）、两个连续单词的组（即 2-gram：例如“强烈推荐”、“应用商店”）或三个连续单词的组（即 3-gram：例如“心率监测器”）。

2. 删除无信息但频繁出现的停用词（例如“the”、“and”）， 3. 词性 (POS) 过滤（即根据词性删除单词，例如名词、动词、形容词或副词），

4. 词形还原（即，将单词简化为其字典形式，例如，“评论”和“评论”为“审查”），

5. 删除数字（例如“2014”），以及

6. 删除可能源自网络抓取活动的 HTML 标签和其他技术符号。

表 2 显示了初始 LDA 分析结果的摘录。它显示了八个选定主题的最可能的单词，揭示了许多损害主题正确解释的数据质量问题。例如：

1. 主题 7 中最可能的术语是“device”和“devices”，主题 15 中最可能的术语是“band”和“bands”。为了协调这些术语，我们在预处理管道中添加了词形还原步骤。

2. 许多主题在前 10 个最可能的单词中包含单词“fitbit”、“flex”、“fit”和“bit”（例如，参见主题 29）。这并不奇怪，因为所有评论都是关于 Fitbit 设备的。
从文本挖掘的角度来看，这些词语不会给评论添加新的信息；相反，它们甚至可能会影响统计分析或妨碍结果的解释。

因此，我们将这些术语添加到自定义停用词列表中，从而消除了它们。

3. 参与感兴趣领域的词汇（例如，正在审查的产品的功能）对于解释和理解任何文本挖掘研究的结果至关重要。例如，我们发现主题28涉及设备的“静音闹钟”功能。

不幸的是，LDA 算法将“alarm”和“silent”这两个词视为独立词。通过修改我们的模型以包含 n-gram，我们迫使算法创建一个新的复合词“silent_alarm”，这有助于更好地理解该主题。在主题 41（“heart”、“rate”和“monitor” “heart_rate_monitor”；“blood”和“pressure” “blood_pressure”）中也观察到了同样的问题。

4. 有些客户在评论中提供了大量详细信息，例如购买的月份和年份。由于这些信息已被元数据（日期字段）捕获，因此我们选择删除数字词。

表 2. 初始主题模型的示例主题（数据准备前）

主题ID	最可能的词
T1	活动生活设备活动时间人长制作生活方式思想
T2	应用程序设备网站 iPhone Android 数据良好的应用程序界面设备设备跟踪
T7	日数据类似监控帐户人点带带子佩戴轻松爱彩色手腕小舒适大
T15	
T28	闹钟静音唤醒睡眠设置功能闹钟振动时钟早晨比特适合爱身体bodymedia购买伟
T29	大的媒体功能思想2014购买购买充电2013七月收到六月月份周
T31	

表 2. 初始主题模型的示例主题（数据准备前）

T41	心率监测器血压跟踪测量活动事物睡眠

为了微调主题模型，我们尝试了不同的数据准备选项，重新运行了 LDA 算法，并使用自动化方法评估生成的描述性主题模型的质量（见表 3 中列出的变体）。我们应用了 Lau 等人（2014 年）的方法，通过计算主题前 n 个词中的术语对在参考语料库上滑动的窄窗口（例如 10 个词）内同时出现的频率来自动评估主题模型的语义一致性（有关该技术的详细信息，请参阅 Lau 等人（2014 年）和 Newman 等人（2010 年））。在实验中，生成的规范化逐点互信息 (NPMI) 度量范围在 -1（最差）和 +1（最好）之间，与人类对语义一致性的判断具有高度相关性（皮尔逊相关性在 0.84 和 0.98 之间）（Lau 等人，2014 年）。我们使用原始评论语料库作为参考语料库，计算了不同预处理选项集的 NPMI 分数。结果表明，表 3 中的配置 #5（即 3-gram 标记、删除标准停用词、删除数字、词形还原、POS 标记（名词、动词、形容词）和一小串自定义停用词（fitbit、flex、fit、bit））产生了具有最佳可解释性的主题模型。

表 3. 不同的数据准备选项及其对语义一致性的影响

#	标记化	标准停用词		删除号码	词形还原	词性过滤	自定义停用词	语义一致性	（国家PMI）
1	1 克	1 克							0.1281
2	1 克		是的						0.1615
3			是的					Fitbit、Flex	0.1872
4	3 克	3 克	是的	是的	是的			Fitbit,弯曲,适合,位	0.2390
5	3 克		是的	是的	是的	N、V、ADJ		Fitbit,弯曲,适合,位	0.2826
6			是的	是的	是的	N、V、ADJ、ADV	fitbit、flex、fit、bit		0.2760

表 4 通过展示 50 个主题中每个主题最可能出现的前 10 个单词来总结最终的主题模型，图 9 则可视化了语料库中主题的总体分布（即，某个主题的概率越高，谈论该主题的评论就越多）。

表 4. 最终主题模型的主题

主题	最可能的单词
T1	天步数周工作步行时间步数_天步行夫妇结束
T2	穿戴式淋浴出水时间带爱充电舒适穿戴_淋浴游泳
T3	体重减轻 减重 一周减重 磅 一个月减重
T4	分钟活跃活动步行步数英里跑步活跃分钟跑步轨迹
T5	腕带佩戴时间拉链计步器放东西丢失夹
T6	心率 心率监测器 心率监测器 心率监测器 血压 血压 计步器
T7	睡眠跟踪时间夜间步骤睡眠跟踪白天模式特征小时
T8	指令工作集网站网站查找网页用户时间图
T9	礼物爱圣诞节买的丈夫女儿收到生日送的礼物
T10	积极进取爱日行走迈出步伐移动激励保持积极进取
T11	产品很好 推荐 很棒_产品爱推荐_产品好 优秀 不_推荐 好产品

表 4. 最终主题模型的主题

T12充电充电器充电保持月单位电池问题hold_charge问题
T13 睡眠模式 sleep_mode put time tap 轻敲 put_sleep 忘记转
T14 追踪睡眠步数 keep_track track_steps 爱情活动 track_sleep 很棒 keep_track
T15 楼梯超计数轨道爬升步骤大飞行交易地板
T16 让时间成为生命中的长事 爱情 make_sure 意识到改变
T17带腕带 硬扣扣 硬腕带 快照时间手环
T18 退货 亚马逊 一天 产品 物品 一周 收费 工作 购买 开心
T19 工作很棒 作品_很棒 not_work 项目广告精美 爱情想法 not_great
T20 工作 停止月份 停止工作周 工作 充电 购买 退出 停止充电
T21精准步手计步器佩戴步幅设置距离优势手腕
T22 丢失带扣手腕坠落时间设计腕带坠落安全
T23 卡路里消耗 卡路里消耗 追踪 燃烧 步数 一天 多少卡路里 吃 体重
T24 个月前 我的坏了 le 买的表带问题 保修期已到
T25 应用程序 iphone 同步 ipad iphone_app 工作 io 苹果 安卓 电脑
T26光力显示时间带步进进度手表点显示
T27计步手臂运动精准手部计数行走count_steps移动
T28 活动睡眠监测级别 每日 Activity_level 活动_睡眠日感知 daily_activity
T29 手机同步电脑应用程序设备 android 同步蓝牙 not_sync 工作
T30 妻子买了爱情秤 aria 妻子_爱买了_妻子礼物 aria_scale
T31 jawbone nike band fuel app fuelband nike_fuel 准确 fuel_band 通缉
T32目标日步每日达成爱设定遇见命中进度
T33 工具 健身 伟大的健康目标 great_tool 程序 追踪功能 活动
T34 好东西工作坏主意设备制造给予价格跟踪
T35 朋友 乐趣 爱 伟大的家庭 挑战 竞争 步骤 竞争 很多
T36 追踪器 睡眠追踪器 睡眠活动健身 很棒 活动追踪器 步骤健身追踪器 准确
T37品评明星买给人好东西读问题
T38 电池 一天的充电寿命 电池寿命 一周 充电时间 小时 低
T39 金钱价值浪费时间浪费金钱不值得一块计步器产品购买
T40 步 英里 一天行走了 Many_steps 行走次数 距离 see_many steps_take
T41 Great Stay Motivator Moving Motivator Great_Motivator Love Active Keep_Motivator 轨道
T42带腕带手腕月破大周小更换腕带
T43 客户服务 customer_service 支持 电子邮件 当天更换 联系问题产品
T44 健身伙伴 Fitness_pal myfitnesspal 应用程序同步 同步爱情应用程序 mfp
T45习惯积极睡眠健康帮助改变模式中等睡眠护理
T46爱推荐东西颜色推荐_任何人一天购买乐队爱_爱
T47 食物日志摄入量卡路里活动水分睡眠跟踪锻炼应用程序
T48闹钟静音唤醒silent_alarm设置功能睡眠振动时钟早上
T49 轻松 easy_use 套装 喜欢 穿着舒适 很棒 easy_set 应用程序 超级
T50设备数据活动信息跟踪软件制作给出时间界面

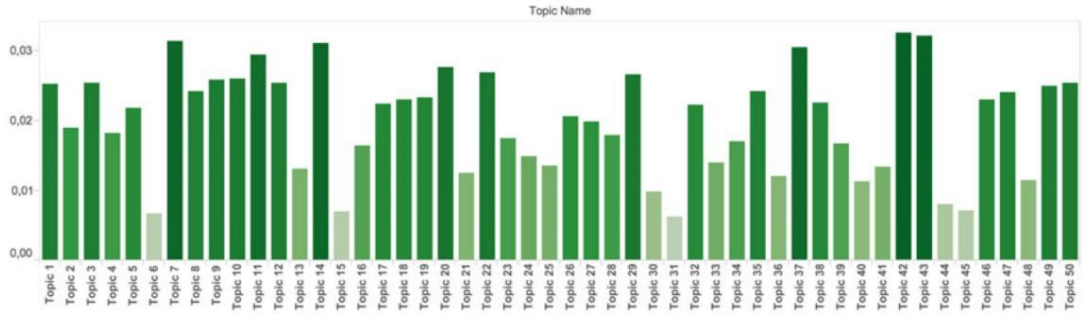

图 9. 总体主题分布

建模阶段的最后一步是量化已识别主题（独立变量)对用户满意度（因变量)的影响。为此,可以应用不同的回归分析技术。最常见的选择是使用线性最小二乘 (OLS) 回归;但是,星级评定是按序数而不是连续尺度来衡量的。因此,有序逻辑回归将是更好的选择。然而,根据我们的数据集测试有序逻辑回归的比例几率假设表明,主题对星级评定的影响在星级评定级别之间有所不同 - 这是用户满意度 J 形分布的结果。因此,我们决定使用多项逻辑回归,它将因变量的不同级别（即 1、2、3、4、5 星)视为无序类别。因此,它会为因变量的每个级别生成单独的系数;在我们的示例中,50 个主题中的每一个都有 5 个系数（即 250 个系数）。为了管理结果模型的复杂性,并提高其可解释性,我们选择了 LASSO（最小绝对收缩和选择算子)来将模型与数据拟合。LASSO 是一种线性回归方法,它通过将不具影响力的独立变量的系数收缩到零来进行变量选择,从而生成一个仅包含最重要的独立变量来解释因变量的模型（Hastie、Tibshirani 和 Friedman,2013 年）。

图 10 可视化了 LASSO 回归模型的系数。3例如,分析显示与 5 星评级相关的前 5 个主题包括:主题 46（“推荐给他人”）、主题 10（“运动动机”）、主题 3（“减肥”）、主题 35（“与朋友竞争”)和主题 49（“易于使用”）。相比之下,与 1 星评级相关的前 5 个主题包括:主题 39（“负成本/收益比”）、主题 18（“亚马逊的产品退货政策”）、主题 20（“故障”）、主题 43（“客户服务”)和主题 8（“操作说明”）。估计模型的拟合优度（以模型解释的偏差分数衡量)为 0.26,分类准确度为 0.57。

[3] 为了为套索惩罚选择最合适的 lambda 参数,我们使用 10 倍交叉验证进行了网格搜索,结果为 =0.00021。

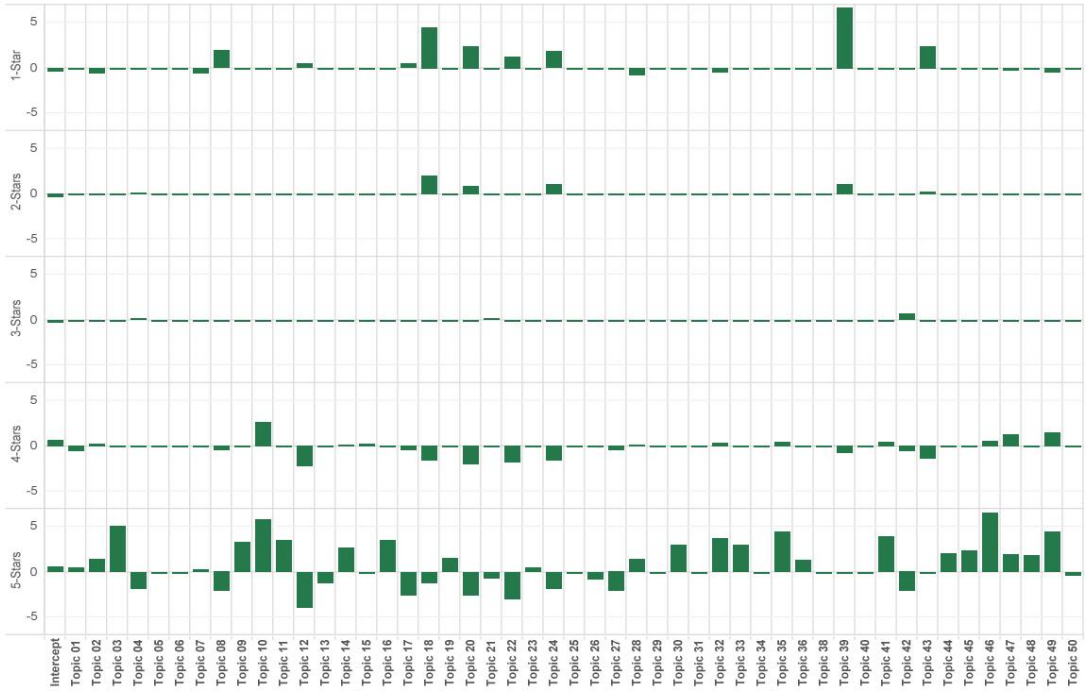

图 10. LASSO 多项逻辑回归的系数

4.4 解释

最后一步是理解和理解发现的主题及其对用户满意度的影响。通过分析主题中最可能的术语以及相关的最可能文档,可以揭示主题的含义。图 11 显示了主题 3 的单词分布的气泡图。气泡的大小和颜色都代表给定术语在给定主题中的概率。此主题的首次标记可能是“减肥”。但是,为了验证基于单词概率的初始解释是否有意义,建议彻底调查相关文档。表 5 证实,客户很高兴地报告了他们借助 Fitbit 设备减肥的成功故事。总体而言,两位研究人员独立解释和标记了所有 50 个主题,除了细微的措辞差异外,编码员之间的一致性达到了 84%。

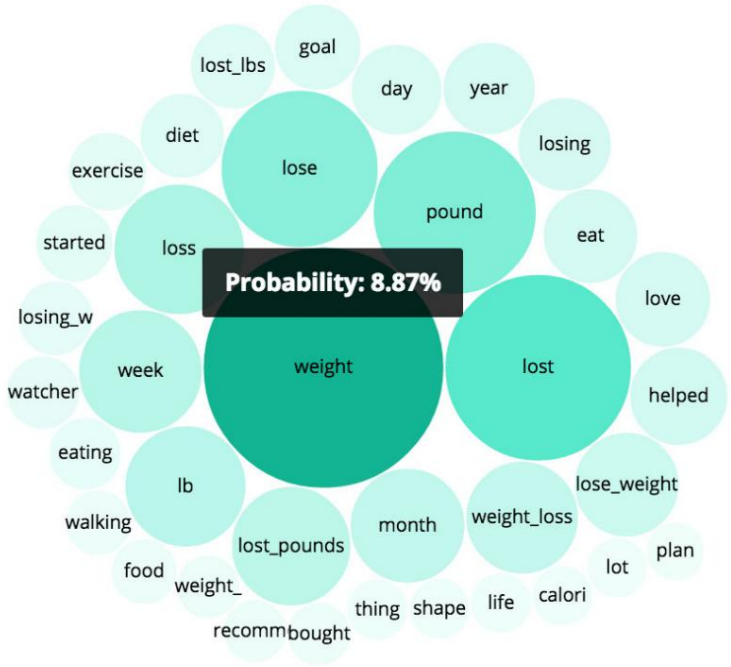

图 11. 主题 3 最可能出现的单词（“减肥”）

表 5. 与主题 3（“减肥”）最相关的评论

概率文本帮助我达到了减		星级评定
0.81	肥目标,并且已经维持了 6 周的体重。已经成为我 24/7 的一部分。	5
0.78	FitBit 帮助我彻底改变了生活方式。到目前为止,我已经减掉了 27 磅,而且还在继续减。	5
0.78	今年夏天我正在尝试减肥 3 项活动。 Fitbit 极大地激发了我的动力。我的体重并没有减轻很多。但是当有人看到我并且他们已经有一段时间没有看到我时,他们会说“你变瘦了”。我会向任何想要获得运动和健身动力的人推荐这款产品。	5
0.76	我很喜欢它。它帮我减掉了 20 磅。	5
0.72	我在 2014 年 7 月底买了这个。我想记录我吃了什么,减掉了大约 15 磅,并提高了我的整体健康水平。 5月份我就60岁了。我改变了我的生活方式,我吃更少的食物和更健康的东西。由于我坚持不懈,我已经减掉了大约 33 磅。我感觉比20年来更好了,自从我注意以来,我的整体健康状况现在非常出色。 Fitbit Flex 一直是我的计划中不可或缺的一部分。完成这一切对我来说是非常有价值的。唯一的问题是,我需要很多新衣服,在三个多月的时间里,我的腰围减少了 4 英寸。为了锻炼身体,我每天快步走两次。如果您学习如何使用 Fitbit,那真的非常简单而且非常值得。	5

为了根据现有理论背景理解发现的主题,我们尝试将解释的主题映射到技术接受领域的理论构造,特别是技术接受模型 (TAM)（Venkatesh & Davis,2000;Venkatesh、Morris、Davis 和 Davis,2003)和 IS 成功模型（例如,DeLone 和 McLean,2003）。与通过标签解释主题类似,两位研究人员根据一系列理论定义（表 7)独立地在主题和构造之间进行了映射。编码员之间的一致性达到了 86%。

表 6. 与用户满意度相关的结构定义

构造子构造定义			来源
系统 质量		信息系统的理想特性。	Petter、DeLone 和 麦克莱恩 2013
	可靠性	系统运行的可靠性 系统适应用户需求变化的方式 Wixom	威克瑟姆和托德 2005
	灵活性	and Todd 2005	
	一体化	系统允许集成来自不同来源的数据的方式	威克瑟姆和托德 2005
	无障碍	访问或从系统中提取信息的难易程度;系统对信息或行动请求做出及时响应的程度;系统输出（内容、报告、仪表板)的理想特征。	
	及时性		威克瑟姆和托德 2005
信息 质量			Petter、DeLone 和 麦克莱恩 2013
	完备性:与	表示在存储的 信息	Nelson、Todd 和用户群体 威克瑟姆 2005
	准确性	信息的正确性、明确性、意义性、可信度和一致性	纳尔逊、托德和 威克瑟姆 2005
	格式	信息以用户可理解和可解释的方式呈现的程度,从而有助于完成任务	纳尔逊、托德和 威克瑟姆 2005
	货币	信息的最新程度,或者信息准确反映其所代表的世界当前状况的程度 系统用户从 IS 部门和 IT 支持获得的支持质量 信息系统的程度为个人、团体、组织、行业和国家的成功做出贡献。	纳尔逊、托德和 威克瑟姆 2005
服务 质量			彼得、德隆和 麦克莱恩 2013
净收益			Petter、DeLone 和 麦克莱恩 2013
用处		个人认为使用特定系统可以提高其工作绩效的程度 个人使用特定系统可以轻松完成工作的程度 用户对信息系统的满意度	戴维斯、巴戈齐和 华沙 1989
使用方便			戴维斯、巴戈齐和 华沙 1989
用户 满意			Petter、DeLone 和 麦克莱恩 2013

表 6 显示了映射过程的示例结果。大多数主题都明确地映射了现有的结构。例如，“减肥”（主题 3)被映射为净收益的结构,因为减肥似乎是使用 Fitbit Flex 设备通过增加体力活动的间接结果。客户对“活动跟踪的准确性”（主题 4)的评论与准确性的构造（IS 成功模型中信息质量构造的子构造)以及“亚马逊的产品退货政策”（主题 18)与服务相对应质量。关于 TAM,我们能够将主题 49“易于使用”映射为易于使用的结构,并将主题 1“步数跟踪”映射为有用,因为它代表了设备的核心功能之一。

总体而言,在确定的 50 个主题中,我们用系统质量映射了 14 个,用实用性映射了 12 个,用净收益映射了 6 个,用信息质量映射了 3 个,用服务质量映射了 2 个,用易用性映射了 2 个 所有这些都是用户满意度的先决条件。其余一些主题被归类为用户满意度的指标,而不是先决条件(主题 11、33 和 46)。

[4] 此外,我们还发现了八个既不符合 IS 成功模型也不符合 TAM 结构的主题。例如,主题 39“负成本/效益”或主题 31“与竞争对手产品的比较”在这两个模型中都没有理论上的等价物。这可能会扩展现有理论或开发全新的理论 这两个目标超出了本教程的目的。

表 7. 与用户满意度相关的结构定义

主题 最可能的术语	高度相关的评论句子示例	标签	映射到现有构造
T1 天步数 周工作步行时间 steps_day 步行情侣结束	我平均每天走 12,350 步。我期待有一天我可以用它来追踪我早起和深夜的跑步。 刚开始使用时,我很幸运每天能走 4000 步,因为我大部分时间都在办公桌前工作。我必须非常努力才能达到 10000 步的目标。现在我已经使用几个月了,并将我的目标提高到每天 12000 步。		用处(文卡特什& Davis,2000)
T3 减重 磅 减重 周 磅 减重 月 减重 减重 帮助 爱吃 减肥 FitBit 帮助我实现了全年目标 减重 磅 生活方式的改变。到目前为止,我已经减掉了 27 磅,并且还在继续减。	帮助我实现了减肥目标,现在已经减肥并保持了 6 周的体重。体重已经成为我24/7的一部分。		净效益(德隆 & 麦克莱恩,2003)
T4 分钟 活跃活动 步行 步数 英里 跑步 活跃分钟数 跑步 轨迹 锻炼 精确记录 自行车 注册 体重测量 跑步机 小时	我喜欢 fitbit,喜欢看到我的步数不断增加。但是,我经常使用椭圆机或自行车,它不会记录这些活动。只适用于步行。这让我很失望。 就在今天,佩戴 Fit Bit 行走在 GPS 测量的 3 英里步道上。我花了48分钟。 Fit Bit 记录了 2.5 英里和 23 分钟的活动。所以如果你能接受 50% 的准确度就可以了。	活动追踪的准确性	信息质量-> 准确性(德隆和麦克莱恩,2003)
T9 礼物爱圣诞节买丈夫女儿收到生日送礼物儿子日圣诞节 _礼物购买年买丈夫姐姐妈妈爱	我买了它作为圣诞礼物送给我姐夫,他很喜欢。 我把这个物品作为礼物送人。我想她和我一样喜欢我收到的礼物。	Fitbit 作为礼物	暂无对应 已识别 IS 构造

[4] 从解释性回归模型中删除这三个主题仅略微降低了其拟合优度和预测性准确度(解释的偏差分数:0.2440,分类准确度:0.5596)。

表 7. 与用户满意度相关的结构定义

T10 积极行动 爱日散步 迈出步伐 行动 激励 keep_motivated		这让我更加意识到我白天需要多运动。它帮助我变得更加健康。 爱它!真正激励您站起来并开始行动!期待充分利用它！	搬家的动机	净收益 （德隆 & 麦克莱恩,2003）
T12充电充电器充电保持月单位电池问题	hold_charge问题not_charge not_hold充电时间usb工作 not_hold_charge灯触点	在电池拒绝充电之前工作了大约六个月。经过多次电子邮件后,制造商确实发送了一封新邮件。现在,六个月后,同样的事情 - 电池无法充电。 喜欢我的 Fitbit。然而,2.5 个月后,我遇到了严重的电池充电问题。希望尽快得到解决。	电池充电 问题	系统质量->可靠性 （德隆 & 麦克莱恩,2003）
T18 亚马逊退货日 fitbit 无法充电,亚马逊不接受 30 天后的退货。这是第一个 fitbit 购买后收到的第二个 fitbit 的充电情况,很高兴退款也没有收费。退货失望政策更换购买购买购买保修 T31 jawbone nike band fuel app fuelband nike_fuel 准确 fuel_band 想要		该商品仅使用了不到 90 天,想要退回亚马逊更换,但不允许退货。	亚马逊的产品 退货 政策	服务质量 （德隆和 麦克莱恩,2003）
		同时购买了这个和 Jawbone。 Jawbone 要好得多。它的应用程序更好,蓝色竞争对手的牙齿连接也很有帮助。产品 我拥有一款 Nike+ 表带,佩戴了将近一年。这款表带的弹性更小,表带可更换（如果你能找到的话）,而且比 Nike 表带多了很多功能。	比较 没有相应的 这在性能上比 IS 构造确定的	
一分钱可以扔掉。产品买值得的东西太贵了,因为它能做什么,能用,但很贵,因为它不准确。我可以得到不浪费的,花同样的东西,免费的,或者很便宜的,失望		我再说一遍,不要浪费你的钱：（不要浪费你的钱。除非你有	负成本/效益 比	没有对应的 已识别 IS 构造
T43 客户服务 customer_service	支持 电子邮件 当天更换 联系问题产品	客户服务太糟糕了。产品有缺陷,Fit Bit 公司让你费尽周折来修理或更换价值 100 美元的产品,希望你放弃。我对产品缺陷的投诉仍然没有得到解决。糟糕的客户服务和体验。 希望你能获得退款或更换。	客户服务	服务质量 （德隆和 麦克莱恩,2003）

表 7. 与用户满意度相关的结构定义

T49 easy	easy_use 设置爱戴舒适 很棒的 easy_set 应用程序 超准确 简单设置 easy_wear 让仪表板工作舒适 _wear 带	非常适合问责制。易于设置和使用。 我买了它给我最好的朋友,她很喜欢它! 她说,这款眼镜安装起来非常简单,而且几乎可以与任何服饰搭配。	易于使用 易于使用	（Venkatesh & 戴维斯,2000）

4.5 总结

我们的说明性主题建模研究展示了如何使用开放和自然发生的文本数据以完全数据驱动、归纳和高度自动化的方式解释给定产品的客户满意度。我们从亚马逊收集了超过 12,900 条关于“Fitbit Flex”可穿戴技术的在线客户评论,并应用 LDA 主题建模算法提取独立变量,以构建用户满意度的解释性统计模型。我们能够将大量归纳识别的主题映射到现有的理论构造中,并将它们放在一个法理网络中,然后我们使用 LASSO 多项逻辑回归对其进行分析。结果表明,净收益和感知有用性方面对积极的用户满意度（4 星和 5 星)的影响最大,而糟糕的系统和服务质量对消极的用户满意度（1 星评级)的影响最大。此外,我们还确定了解释因素,例如“负成本/收益比”,这些因素不属于现有的技术接受度 IS 理论。

5 结论

在本教程中,我们讨论了文本挖掘的挑战,特别是主题建模,并通过一个示例展示了其应用。研究人员可以将本教程用作自己主题建模研究的蓝图和示例,或判断他人研究的质量。

文本挖掘方法提供了广泛的工具,可以在合理的假设和成本下分析大量不同的文本,从而使信息系统研究人员能够利用以前基本上无法访问的新数据源。然而,尽管过去十年在自然语言处理和机器学习方面取得了所有进步,但这些统计技术利用简化模型来处理自然语言的复杂性,并且远未复制人类分配的过程对语言的意义。例如,大多数主题模型将文本视为无序的单词集,完全忽略词序或句子结构。此外,仅仅因为已经证明主题建模在某些数据集上提供了高质量的结果,并不意味着它在每个数据集上都表现良好。例如,如果总体文本集合很小（例如,调查中的开放式问题）、范围非常广泛（例如,电子邮件）、嘈杂（例如,从网站上删除的文本）,或者单个文档是很短（例如推文）,主题建模可能无法产生有洞察力的结果。因此,通过实验和三角测量来评估主题建模结果的有效性至关重要。毕竟,主题建模等文本挖掘方法无法取代人类分析,而只能增强人类分析。

在本教程中,我们仅介绍了一种文本挖掘技术:主题建模。根据研究目标,应用其他技术可能更合适。由于其无监督性质,主题建模特别适合在大型文本集合中归纳发现模式。特别是对于缺乏构造和理论的领域的探索性研究,或者对于扩展现有理论,这种方法可能很有用。相反,如果研究对象是确认性的,那么基于字典的方法可能更合适。使用基于字典的方法,研究人员可以仔细生成字典和规则,以便将模型与一组预定义的可测试假设相匹配;然而,它们的探索潜力非常有限。

最后,在本教程中,我们尝试以一种易于理解的方式向广大受众介绍两种复杂的统计方法（即 LDA 和 LASSO）。建议对 LDA 或 LASSO 应用感兴趣的研究人员在解释其输出之前,先彻底研究原始文献,以更深入地了解这些方法。

参考

贝伦特,N. 和塞德尔,S. (2014)。大数据和归纳理论发展:走向计算基础理论?第 20 届美洲信息系统会议论文集（第 20 页）
1–11）。 萨凡纳。

Berg, BL 和 Lune, H. (2011)。社会科学研究的定性研究方法。波士顿:皮尔逊。

布莱,D.（2012）。概率主题模型。 ACM 通讯,55(4), 77–84。

Blei, D.、Ng, A. 和 Jordan, M. (2003)。潜在狄利克雷分配。机器学习研究杂志，
3（1）,993–1022。

Boyd-Graber, J.、Mimno, D. 和 Newman, D. (2014)。主题模型的维护和养成:问题、诊断和改进。收录于 EM Airoldi、D. Blei、EA Erosheva 和 SE Fienberg (Eds.) 的《混合成员模型及其应用手册》（第 3-34 页）。博卡拉顿:CRC Press。

美国商业资讯。 （2010）。 2010 年社交购物研究揭示了消费者在线购物习惯使用客户评论的变化。来自http://www.businesswire.com/news/home/20100503005110/en/2010-Social-Shopping-Study-Reveals-Consumers 和 的 Online 检索

Chang, J.、Boyd-Graber, J.、Gerrish, S.、Wang, C. 和 Blei, D. (2009)。阅读茶叶:人类如何解释主题模型。神经信息处理系统进展会议论文集（第 1-9 页）。温哥华。

德隆,WH 和麦克莱恩,ER (2003)。信息系统的德隆和麦克莱恩模型
成功:十年更新。管理信息系统杂志,19(4), 9–30。

Evangelopoulos, N.、Zhang, X. 和 Prybutok, VR (2012)。潜在语义分析:五项方法建议。欧洲信息系统杂志,21（1）,70-86。

范 W.、华莱士 L.、里奇 S. 和张 Z. (2006)。发挥文本挖掘的力量。通讯
ACM,49（9）,76-82。

JR 弗斯 (1957)。 1930-1955 年语言理论概要。语言分析研究（第 1-32 页）。
牛津:语言学会。

Frawley, W.、Piatetsky-Shapiro, G. 和 Matheus, C. (1992)。数据库中的知识发现:
概述。人工智能杂志,13(3),57–70。

Gopal, R.、Marsden, JR 和 Vanthienen, J. (2011)。信息挖掘 - 对数据、文本和媒体挖掘的最新进展和未来发展的反思。决策支持系统,51(4), 727–731。

格里默,J. 和斯图尔特,B. (2013)。文本即数据:自动内容分析的前景和陷阱
政治文本的方法。政治分析,21(3),1-31。

Halevy, A., Norvig, P., & Pereira, F. (2009). 数据的不合理有效性。IEEE 智能
系统,24(2),8-12。

哈里斯,Z.（1954）。分配结构。字,10（23）,146-162。

Hastie, T.、Tibshirani, R. 和 Friedman, J. (2013)。《统计学习的要素》。纽约:Springer。

霍夫曼,T.（1999）。概率潜在语义索引。第 22 届国际 ACM SIGIR 信息检索研究与开发会议论文集（第 50-57 页）。伯克利。

胡 N.、张 J. 和 Pavlou, P. a. （2009）。克服产品评论的 J 形分布。
ACM 通讯,52(10), 144-147。

Indulska, M.、Hovorka, DS 和 Recker, J. (2012)。内容分析的定量方法:识别出版渠道之间的概念漂移。欧洲信息系统杂志,21(1), 49–69。

Kurgan, LA 和 Musilek, P. (2006)。知识发现和数据挖掘过程模型的调查。
知识工程评论,21（1）,1-24。

Landauer, TK、Foltz, PW 和 Laham, D. (1998)。潜在语义分析简介。话语
过程,25(2-3), 259–284。

Lau, JH、Newman, D. 和 Baldwin, T. (2014)。机器阅读茶叶:自动评估主题连贯性和主题模型质量。《计算语言学协会欧洲分会第 14 届会议论文集》,530–539 页。

刘,B.（2011）。网络数据挖掘。柏林:施普林格。

Martens, D. 和 Provost, F. (2014)。解释数据驱动的文档分类。 《管理信息系统季刊》,38(1),
73–99。

McAuley, J.、Pandey, R. 和 Leskovec, J. (2015)。推断可替代和互补产品的网络。第 21 届 ACM SIGKDD 国际知识发现和数据挖掘会议论文集。悉尼。

McAuley, J.、Targett, C.、Shi, Q. 和 van den Hengel, A. (2015)。基于图像的样式和替代推荐。第 38 届国际 ACM SIGIR 信息检索研究与开发会议论文集。圣地亚哥。

米歇尔,J.-B.,沉,YK,艾登,AP,韦雷斯,A.,格雷,MK,皮克特,JP，……艾登,EL（2011）。
使用数百万本数字化书籍对文化进行定量分析。科学,331（6014）,176-182。

Miles, M. 和 Huberman, A. (1994)。定性数据分析:扩展资料手册。千橡市:
Sage 出版公司

Mimno, D.、Wallach, H.、Talley, E.、Leenders, M. 和 McCallum, A. (2011)。优化主题模型中的语义一致性。自然语言处理经验方法会议论文集。斯特劳兹堡。

Miner, G.、Elder, J.、Hill, T.、Nisbet, R.、Delen, D. 和 Fast, A. (2012)。非结构化文本数据应用的实用文本挖掘和统计分析。沃尔瑟姆:学术出版社。

Mudambi, SM 和 Schuff, D. (2010)。什么是有用的在线评论?顾客研究
亚马逊网站上的评论。 《管理信息系统季刊》,34(1),185–200。

Müller, O.、Junglas, I.、vom Brocke, J. 和 Debortoli, S. (2016)。利用大数据分析进行信息系统研究:挑战、承诺和指南。欧洲信息系统杂志,即将出版。

Newman, D.、Lau, J.、Grieser, K. 和 Baldwin, T. (2010)。主题连贯性的自动评估。载于计算语言学协会北美分会会议论文集。洛杉矶。

好的。 (2012)。 打开 数据 手册文档。 http://opendatahandbook.org/pdf/ 检索 从
OpenDataHandbook.pdf

奥特,M.,崔,Y.,卡迪,C.,&汉考克,JT（2011）。尽其所能地发现欺骗性意见垃圾邮件。计算语言学协会第 49 届年会论文集:人类语言技术。波特兰。

庞 B. 和李 L. (2008)。意见挖掘和情感分析。信息基础和趋势
检索,2(1-2),1–135。

彭尼贝克,JW (2011)。代词的秘密生活:我们的言语如何讲述我们。纽约:布卢姆斯伯里出版社。

Quinn, KM、Monroe, BL、Colaresi, M.、Crespin, MH 和 Radev, DR (2010)。如何以最小的假设和成本分析政治关注。《美国政治科学杂志》,
54(1),209–
228.

Ramage, D.、Rosen, E.、Chang, J.、Manning, CD 和 McFarland, DA (2009)。社会科学主题建模。主题模型应用研讨会论文集。惠斯勒。

Saldaña, J. (2012).《定性研究人员编码手册》。伦敦:Sage Publications, Inc.

Salton, G., & McGill, M. (1983). 现代信息检索简介。纽约:McGraw-Hill。

希勒,C.（2000）。 CRISP-DM 模型:数据挖掘的新蓝图。数据仓库杂志,5(4), 13–22。

史密斯,AE 和汉弗莱斯,MS (2006)。使用 Leximancer 概念映射评估自然语言的无监督语义映射。行为研究方法,38(2), 262–279。

Tausczik, YR 和 Pennebaker, JW (2010)。词语的心理意义:LIWC 和计算机文本分析方法。语言与社会心理学杂志,29（1）,24-54。

Thelwall, M.、Buckley, K.、Paltoglou, G.、Cai, D. 和 Kappas, A. (2010)。简短强度检测非正式文本中的情感。美国信息科学与技术学会杂志,61（12）,2544-2558。

特尼,P.,&潘特尔,P.（2010）。从频率到意义:语义的向量空间模型。杂志
人工智能研究,37(1), 141–188。

推特。 （2015）。 Twitter 使用情况和公司情况。 2015 年 10 月 30 日检索自
https://about.twitter.com/company

Urquhart, C. (2001). 与扎根理论的邂逅:解决实际和哲学问题。收录于 EM Trauth 主编的《信息系统中的定性研究:问题与趋势》（第 34 页)中。
104–140）。赫尔希:Idea Group Publishing。

Urquhart, C. (2012)。定性研究的扎根理论:实用指南。Sage Publications, Inc.

Venkatesh, V. 和 Davis, FD (2000)。技术接受模型的理论扩展:四
纵向实地研究。管理科学,46（2）,186–204。

Venkatesh, V.、Morris, M.、Davis, G. 和 Davis, F. (2003)。用户对信息技术的接受程度:
迈向统一观点。MIS季刊,27(3),425–478。

Wallach, H.、Mimno, D. 和 McCallum, A. (2009)。重新思考 LDA:为什么先验很重要。在诉讼程序中
神经信息处理系统会议。温哥华。

Wallach, H.、Murray, I.、Salakhutdinov, R. 和 Mimno, D. (2009)。主题模型的评估方法。国际机器学习会议论文集。蒙特利尔。

关于作者

Stefan Debortoli是一名研究助理和博士。列支敦士登大学信息系统研究所的候选人。他的博士研究重点是应用大数据分析作为信息系统研究的新探究策略。在大数据分析领域,他专注于文本挖掘技术的研究应用。他的研究成果发表在《欧洲信息系统杂志》、《信息系统协会通信》以及《商业与信息系统工程》上。 Stefan 在各行业的软件工程和 IT 项目管理领域拥有多年的工作经验。

Oliver Müller是哥本哈根信息技术大学的副教授。他拥有信息系统学士和硕士学位以及博士学位。德国明斯特大学经济学博士。奥利弗的研究目标是帮助组织和个人通过大数据和分析创造价值。在这方面,他特别关注从互联网和企业内部来源的大量非结构化文本数据中提取知识。他的研究成果发表在《欧洲信息系统杂志》、《信息系统协会杂志》、《信息系统协会通讯》、《计算机与教育》等杂志上。

Iris Junglas是佛罗里达州立大学信息系统副教授。她的研究兴趣涉及广泛的主题,最突出的是电子商务、移动商务和移动应用商务、医疗保健信息系统、IT 消费化和商业分析领域。她的研究成果发表在《欧洲信息系统杂志》、《信息系统杂志》、《信息系统协会杂志》、《管理信息系统季刊》、《战略信息系统杂志》等各种杂志上。她是《管理信息系统季刊》和《战略信息系统杂志》的编辑委员会成员,也是《欧洲信息系统杂志》的高级副主编。

Jan vom Brocke是列支敦士登大学信息系统教授。他是喜利得业务流程管理主席、信息系统研究所所长、信息系统国际硕士项目联席主任、博士生导师。列支敦士登大学商业经济学项目和研究与创新副校长。在他的研究中,他专注于数字创新和转型,尤其是业务流程管理、设计科学研究、绿色信息系统和神经信息系统。他的研究成果发表在《管理信息系统季刊》、《管理信息系统杂志》、《商业与信息系统工程》、《信息系统协会通信》、《信息与管理》等杂志上。他是影响深远的书籍的作者和编辑,其中包括《业务流程管理国际手册》以及《BPM – 数字世界中的驱动创新》一书。他在信息系统研究和教育领域担任过各种编辑职务和领导职务。