案例 7：玻森数据：从海量文本中挖掘商业价值

玻森数据是国内首家提供完整商用语义分析引擎的公司。通过整合企业内部客服、售后记录、第三方平台用户评论等数据，玻森数据的语义分析技术可以对企业数据进行层次化、多维度的分析挖掘，从而帮助企业实时、客观、全面地了解市场与产品。

互联网行业从不缺乏热词。近年迅速升温的，除互联网思维外，便是大数据。无论身在 IT 还是传统产业，企业都恨不得用大数据武装自己。不得不承认，这种奇异的全民「大数据」的现状，一方面反应了行业心态的浮躁，另一方面也折射出企业希望把握科技浪潮的焦灼。如今多达 80% 的商业数据都以文本、图像等「非结构化」的形式存在，如何挖掘出数据背后的价值，这对企业而言不仅是机遇，更是巨大的挑战。

在这轮大数据的浪潮中，许多企业已然在数据采集与储存上投入大量人力物力，却往往忽视了一个重要的问题：大数据意味着涉及资料规模巨大、数据结构复杂，无法仅通过传统分析工具和手段将其整合为直观、有用的分析结果。建造数据农场本身并不会为企业带来实际利益；与此相反，盲目抓取与储存数据只会增加运营成本。这也意味着，许多企业其实并未意识到，也不知应该如何在最大程度上发挥数据的价值。

大数据的战略意义并非掌握庞大的数据信息，而是对这些包含意义的数据进行专业处理。商家需要从数据中读出消费者对自己的反馈；拥有只是前提条件，

让机器读懂数据，处理、挖掘并提炼出数据中的价值，才是最终要达到的目的。这便是玻森数据正在解决的问题。

通过整合企业内部客服、售后记录、第三方平台用户评论等数据，玻森数据的语义分析技术可以对企业数据进行层次化、多维度的分析挖掘，从而帮助企业实时、客观、全面地了解市场与产品，譬如消费者对产品改进意见反馈、集中投诉的问题、以及客户流失原因等。

2013 年，玻森数据荣获国家科技型中小企业技术创新项目基金。

2014 年 3 月，玻森数据旗下中文语义开放平台 BosonNLP（www.bosonnlp.com）上线，首次将真正意义上的人工智能语义识别技术运用到商业服务中。「从核心的语义分析引擎角度讲，我们是国内第一家提供完整商用语义分析引擎的公司。」玻森数据联合创始人闵可锐谈到，虽然公司成立时间并不长，但在语义分析方向引擎和数据的积累已经超过 8 年。

@创之网

作为一家拥有核心人工智能技术的大数据公司，玻森数据专注于提供非结构化数据的分析引擎及解决方案。凭借自主研发的自然语义识别分析系统、情感分析系统、图像识别系统等人工智能技术，如今的玻森数据已经在网络监测、市场调研、精准营销等多个领域中崭露头角。

互联网解决了商家与消费者之间信息不对称的问题，也使得品牌掌控全局的时代成为过去。知己知彼方可百战不殆，在这个竞争极度激烈的市场，企业唯有了解消费者对产品的评价，并根据其反馈调整市场策略方为上上之策。

而要深入分析网上众多的用户评价，语义分析是不可或缺的关键技术。我们可能用过类似「围脖关键词」的应用来自动提取并显示文本的核心词语，从而达到辅助分析的目的。玻森在此基础上迈出了一大步。作为一家专注中文语义分析

领域的公司，玻森拥有整套自主研发的自然语义分析系统，可通过机器学习方法对海量互联网文本进行分析学习建模，从而从分词、词性、句法等一系列角度对文本实现综合分析。

如此看来，摆在玻森面前的是一个广阔的市场：退可帮助公司新闻监测、品牌分析产品口碑；进可作为技术与解决方案提供商，面向不同行业提供定制服务，与第三方公司和企业携手建立产品分析模型。举例来讲，商家想树立品牌形象，就必须时刻了解顾客反馈、从多种维度分析顾客心理，尤其是跟踪顾客的负面评价内容；这便需要用到文本情感分析。

简单来说，情感分析指对文本中情感的倾向性和评价对象进行提取的过程。

传统广告、监测公司往往低估该任务的难度，通常的方法是采用「正负面词典」对一篇文章进行判定，即通过简单计算文本中出现的正面以及负面词语的个数判断文章的情感倾向。这种方式可能导致忽略词语间搭配、上下文、一词多义等情况，准确率难以达到商业应用的级别。闵可锐表示，想做到精确辨析文本含义，就要着眼于整体句法，而非简单依赖几组关键词。玻森的智能学习算法可以让机器自动学习词语相关性并挖掘语义关联词汇；这也是它与传统分析算法的最大区别。

谈及机器学习，许多人对其概念存在误读。机器学习并非让机器模仿人类大脑思考，而是教会机器逐级处理信息，并根据上下文进行相应修正。在文本处理方面，这一过程指让机器「明白」词语间的相互关联，最终理解词语在句中的真实含义。

玻森使用「大规模训练数据标注+高效特征挖掘算法」的模式，试图在大量文本训练的基础上让机器最终实现自主识别。数据标注相当于为机器提供学习样例，譬如「我今天很高兴→正面」，「我今天不开心→负面」；而算法可以在众

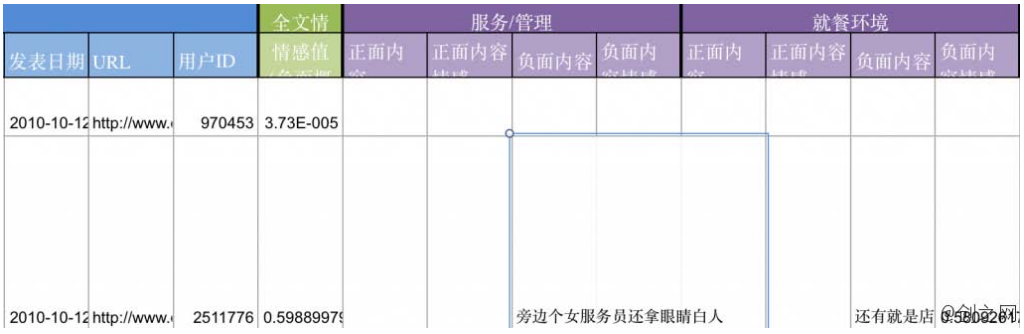
多语言特征量中找到对情感判断最有效的信号。通过机器自动学习与初始人工标注，目前玻森的通用情感分析准确率可达 85%。

这并非情感分析技术的全部。玻森的情感引擎并不仅是给出非黑即白的正面或负面结果，而且会提供其正负面程度，从海量评价中自动提取最正面和最负面的内容，方便客户快速作出改进。

当然，企业需要的数据分析维度往往不仅是单纯的情感分析。这便涉及到了玻森的第二类业务，即非结构化数据解决方案。

虽然目前仍处于起步阶段，但非结构化数据价值挖掘已是一个需求广泛的领域。传统行业从未面对像如今这样复杂的消费者群体。借助互联网，消费者彼此之间分享信息，一条评价中的褒贬可能被数十甚至数百倍放大。有别于传统调研问卷的固定模式，如今消费者在网上写下的评价更为复杂。缺乏引导的反馈往往缺乏条理性，一条评论中，产品、服务、环境等多个方面相互交织，积极与负面反馈参杂不清，网络热词等非规范化语言也可能大量出现。这种由自然语言写就的文本为分析过程带来了极大挑战。

针对这种情况，玻森在掌握客户分析需求与目的基础上，通过综合运用文本分类、聚类、情感、信息提取等多种分析引擎并建模，提炼特定对象中的有效数据，使用户洞察更为真实。例如，企业希望了解消费者网络反馈意见，需要首先从海量数据中总结顾客集中反馈的方面，将其归为不同类别；再通过信息抽取将用户点评归类到所属内容中，最后通过情感分析进一步区分其正面或负面内容。



以某品牌反馈为例，不难看到顾客评价中糅合了多种内容。经过综合分析，其中「旁边有个女服务员还拿眼睛白人」被归为「服务管理」类别，并被判定为服务管理的负面内容。

以这种方式，海量数据被转换为简单的统计表，企业可以直观挖掘出潜藏的问题，从而推出对应的市场宣传与营销策略。这才是大数据的真正优势：让产品与服务更加契合用户的真实需求。

「玻森是一家技术与数据驱动的公司。」闵可锐表示，「数据是玻森的核心资产，我们内部专门有平台和团队不断进行数据完善。」无需多言，要实现精准的机器分析，除了准确高效的引擎外，背后必定离不开庞大的数据与样例支持。玻森数据的竞争力之一就在于其自主研发的千万级中文语料库，其中包括微博、新闻语料、广播语料和论坛语料四个部分，可有效覆盖常见的词语和语法结构。

我们今年三月份才转到分析业务方向，如今已经得到了不少客户的认可，也与多家分析机构展开合作。如今人们每年在网络上分享的内容相当于过去几年的数量；对大数据与人工智能的谈论已久，而市场需要有公司拿出真正有价值的产品。谈及行业背景，玻森表示，国内公司大多直接做到产品端，例如提供舆情分析软件等，而鲜有公司提供底层技术和解决方案。即便是拥有处理结构化数据的 BI 团队的大型电商，在处理非结构化数据上也仍处于初级阶段。

「许多公司在看过我们的应用案例后，才知道数据的价值以及可挖掘的深度。」总结其原因，闵可锐认为当前市场仍处于启蒙时期。因此，玻森的近期目标并非是与其他公司展开竞争，而是帮助市场了解非结构化数据的价值以及实现流程。如今的玻森正致力开发品牌标准化报告，即在不经人工干预的前提下采集品牌相关数据，并利用语义分析引擎自动生成数据分析报告，将内容分类、情感分析、流失原因等以最简明直观的方式呈现给客户。未来还将针对不同行业开发标准化分析报告，在短时间、低成本的前提下让客户了解品牌相关情况，形成决策信息。这只是玻森应用的冰山一角。背靠大数据和云计算，通过可灵活扩展的语义解决方案，自然语义分析还可以实现相似话题聚类、典型意见抽取、过滤噪音歧义等多种功能。未来无论在市场研究、舆情监测，还是电子商务、金融投资领域，语义分析都将拥有不可替代的位置。

注：案例来源于“玻森数据”

思考题：

（1） 从“玻森数据”的成长历程探讨你对文本挖掘这一技术的价值认识。

（2） 你觉得文本挖掘技术的难点在哪里？

（3） 文本挖掘技术如何才能形成产业？有什么样的市场应用前景？