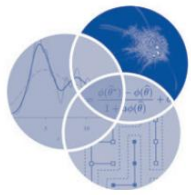
高级复习



通过主题提取进行文本挖掘的实用指南

Andrew Karl、James Wisnowski和 W. Heath Rushing

随着大量非结构化但非常有用的数据以无限的速度生成,文本分析继续激增。文本数据的向量空间模型（其中文档由行表示,单词由列表示）将这种非结构化数据转换为可以使用统计和机器学习技术进行分析的格式。这种方法在揭示共同主题、聚类文档、聚类单词以及将非结构化文本字段（例如开放式调查响应)转换为可用于预测建模的输入变量方面给出了出色的结果。在讨论了文本的收集和处理之后,我们探讨了文档术语矩阵（DTM)的属性和转换。我们展示了如何使用奇异值分解来大幅减小文档空间的大小,同时还为自动主题提取奠定了基础,这得益于最大方差旋转。这种潜在语义分析（LSA)方法产生与图形探索和高级分析兼容的因素。我们还探讨了用于主题分析的潜在狄利克雷分配。我们参考已发布的 R 软件包来实现这些方法,并总结了其他流行的开源和商业软件包。

如何引用本文:
WIREs Comput Stat 2015, 7:326-340。 doi: 10.1002/wics.1361

关键词:文本挖掘;文本分析;奇异值分解;潜在语义;分析;非结构化数据;潜在狄利克雷分配

介绍

人们越来越多地发现自己拥有大量数组织构成文本文档。他们花费大量资源来收集和存储这些数据,但大多数都未能利用分析中包含的丰富信息。如此大的文本集合（语料库)中的典型问题包括:

- 语料库可以分成相似文档的集群吗?
- 语料库中的主要主题是什么? · 特定词语是否组合在一起?
- 哪些词与特定词密切相关单词?
- 文档中的信息是否可用于预测数字响应? · 文档是否可以归类到现有类别中? · 文档采取正面还是负面的语气?

通过文本挖掘的过程,可以使用现有的统计和机器学习方法得出这些问题的答案。本文中文本挖掘的一般流程是:

1.定义要解决的问题陈述,简洁地。
2.收集适当的文本和结构化数据。
3.通过删除单词/字符（例如拼写错误、常见字符)来处理和过滤文本

*通讯作者:james.wisnowski@adsurgo.com Adsurgo LLC,美国科罗拉多州丹佛市

利益冲突:作者声明本文没有利益冲突。

单词（停用词）、从文本收集处理中继承的无关词、在单个或两个文档中出现的罕见词、太短或太长的单词、数字和非标准字符。此外，将同义词更改为单个代表性单词有时也很有用。

4.将文本转换为适当加权的矩阵，以便进行统计分析。

5.探索和发现主题和共同点
主题。

6.将相似的文档和单词分组。

7.从文本中创建新的结构化变量
用于预测分析。

将文本转换为数值的关键是将语料库转换为文档-术语矩阵 (DTM)，每行对应一个文档（页面、段落、记录、文件等），每列对应整个语料库中出现的每个单词。此 DTM 向量空间模型允许使用现有的数据挖掘方法（经过一些修改）分析语料库，将文档视为观察值，将单词视为变量。许多文本挖掘方法使用 DTM 的转置，称为术语-文档矩阵。但是，DTM 更适合我们的分析方法，因为它符合变量作为列、观察值作为行的预期格式。文本挖掘问题的一个主要挑战是 DTM 通常非常大。幸运的是，由于大多数单词在一行中出现的频率相对较低，因此 DTM 也是稀疏的。我们可以使用广泛使用的稀疏矩阵算法来有效地存储和操作 DTM。一种有用的 DTM 操作是奇异值分解 (SVD)，它为我们提供了表示文档空间的特征向量和表示单词空间的特征向量。潜在语义分析 (LSA) 使用来自 DTM 的降阶奇异值分解的这些因素来发现有用的关系。

常用的“词袋”方法假设文档中单词出现的顺序以及它们的词性和其他语法可以被忽略。这种方法对于从语料库中提取模式非常有效（参见参考文献 1-4）。这种对背景的忽视虽然看似极端，但实际上对上述问题产生了令人惊讶的良好结果。然而，根据研究目标，考虑上下文的复杂性可能会很有用。许多软件包（例如 SAS Text Miner、OpenNLP5）提供了解释词序和标记词性的功能。

我们将重点讨论使用“词袋”方法的 LSA。我们使用了 9466 篇摘要的运行示例，其中包括 1990 年国家科学基金会 (NSF) 资助奖，这些摘要是从加州大学欧文分校机器学习存储库获得的。6摘要的中位长度为 154 个单词，涵盖广泛的主题来自生命科学、工程、数学、地质学、海洋学和许多其他科学领域。

收集和处理文本

在明确定义研究要解决的问题后，文本挖掘进程的下一步是确定非结构化文本的适当来源以及如何提取它们。理想情况下，数据可能在企业内部可用，并且只需要少量的格式化和查询即可提取适当的字段，例如自由格式的调查评论。更常见的是，可能需要组合来自不同来源的文本，例如 Word 或 pdf 文档以及来自网站的文本。语音转文本也是一种常见的文本数据源，尽管它通常伴随着较高的数据翻译错误率。对于想要练习文本分析而不需要构建自己的语料库的新用户，Miner 等人的书 1提供了几个经过处理的示例语料库，并介绍了几个软件包的文本挖掘功能 -

年龄。

从文件中提取文本计算机本地存储的文
本通常为 Word 文档或 PDF 文件。虽然许多文本挖掘软件包都包含从这些文件转换文本的功能，但有时需要手动执行此操作以便在将文本发送到文本挖掘软件之前对其进行处理。由于文本嵌入 PDF 文件的特性，在转换这些文件时文本可能会混杂在一起。具有光学字符识别 (OCR) 功能的程序（例如 Adobe Acrobat Professional）往往能够更好地转换这些文件。虽然 OCR 最常用于从扫描文档中提取文本，但在转换使用非标准编码创建的数字 PDF 时也很有用。

另外，PDFBox 是一个开源库，已被证明具有良好的性能。3 Apache Tika （https://tika.apache.org）可以提取许多常见的文件类型，并可以自动检测内容，然后使用适当的库进行解析。

网络和 Twitter 爬行互联网提供了源源不断的新文本。

一些公司发现,通过网络爬取与其产品相关的材料可以比通过客户调查或焦点小组等更传统的渠道更快地获得反馈。Scrapy7是一个流行的开源网络爬虫框架。也可以在 R 中围绕RCurl包构建网络爬虫。8 网络文本分析的挑战是从大量嵌入的 HTML 代码中提取相关文本。由于 HTML 站点的构造方式各不相同,因此此过程很难自动化。但是,这种额外的结构也带来了一些好处。页面、日期、类别和标题等元数据可以清晰划分。在分析大型新闻网站时,使用页面标题而不是每篇文章的全文可能就足够了。

Twitter 数据是 Web 数据的特殊子集,需要最少的处理,因为它没有封装在 HTML 中。有多种应用程序编程接口 (API) 可用于连接 Twitter 来提取数据。 R 包StreamR9利用 Twitter Public Stream API,允许在用户指定的时间长度内实时监控 Twitter 的搜索短语。推文通常是非正式的,有许多“特殊”方式来表达同一件事,这可能需要在分析之前进行一些额外的清理。

预处理文本必须对来自不同来源

和格式的文本进行处理,以删除无关的细节(代码、注释等)并将其转换为纯文本格式。此步骤还涉及将所有文本转码为通用编码:许多 R 文本挖掘函数期望可以编码为 UTF-8 的字符。此外,在处理消息之前,可能需要从消息集合中删除公共标头(例如从电子邮件服务器中)。尽管有许多软件包能够执行此预处理步骤,但我们发现通过R 或 Perl10 中的grep函数使用正则表达式可以提供可靠的结果。

一旦相关文本被提取并以纯文本格式表示,单词本身需要进行处理。标点符号通常与相邻单词分开,并且大小写必须规范化,通常为小写。您需要仔细考虑是否删除数字,因为它们可以提供问题的背景,或者可能是语料库的重要组成部分。例如,社交媒体分析希望包括应有的数字“俚语”术语的激增(例如,be4 作为 before 的一种表示)。

超过 30 个字符的单词通常会被删除:几乎所有英语单词都少于 30 个字母。任何更长的内容都可能是网址或一串乱码文本(在将 PDF 转换为文本时很常见)。由于单词可能根据其语法用途而采取不同的形式,因此对文本进行词干提取通常很有用。11词干提取可以有效地去掉单词的结尾。复数和单数名词被简化为相同的标记,共轭动词也被映射为单个代表词。

语料库(词典或词典)中术语的出现频率大致与其排名成正比,这是齐普夫定律所描述的行为。12这具有一些重要的实际意义。一小部分单词会如此频繁地出现(并且可能在如此多的文档中),以至于它们不具有任何区分能力。这些被称为停用词。软件包包含可以自动删除的常见停用词列表以及提供它们的选项。大部分不常见单词会出现在很少的文档中,因此在搜索语料库中的主导模式时它们没有用处。在对文档进行聚类或搜索主要主题时,最佳实践是删除出现在语料库中少于 1% 的文档中的术语。即使文本挖掘分析的目标是预测或分类,删除仅出现在单个文档中的单词通常也是安全的。

通常,会有一些中频词在区分文档时提供最大的灵活性。

图 1 显示了 NSF 语料库中前 100 个单词的频率,这些单词未经过词干提取或删除停用词。前三个单词是“the”、“of”和“and”。从图 2 可以看出,语料库中超过一半的术语只出现在一个文档中,因此对于检测语料库中的模式没有用处。

去除这些噪声项可以加快后续的 DTM 分析速度并提高分析质量。

DTM 中文本的表示

DTM 通常很稀疏(大多数条目为 0)。
即使对于中等规模的应用程序,如果将完整的 DTM 表示为密集矩阵,那么内存中也难以容纳。可以使用特殊软件和算法来存储和操作稀疏矩阵。13在 NSF 摘要示例中,DTM 的 4450 万个组件中有 99.8% 为 0。稀疏表示需要 18MB 来存储,而密集表示则需要 3.6 GB。矩阵

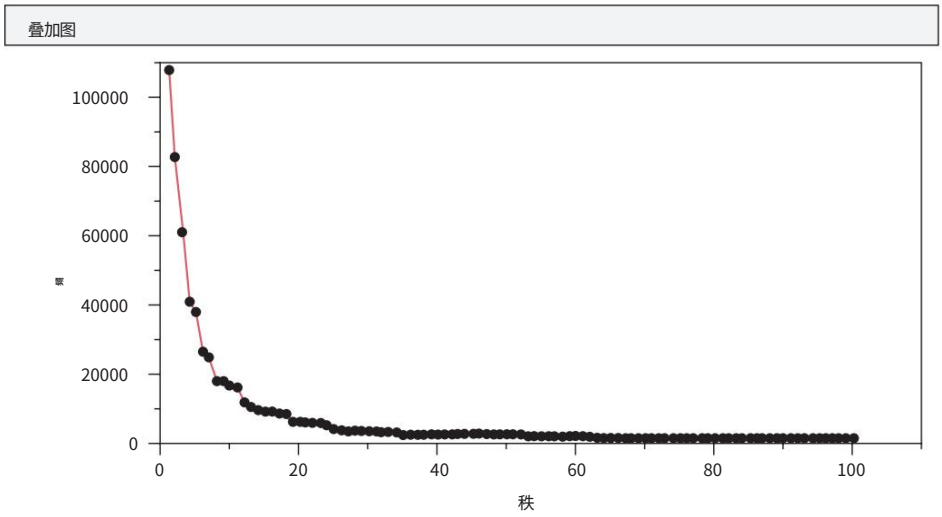


图 1 |美国国家科学基金会语料库的前四个词为“the”、“of”、“and”的词频。

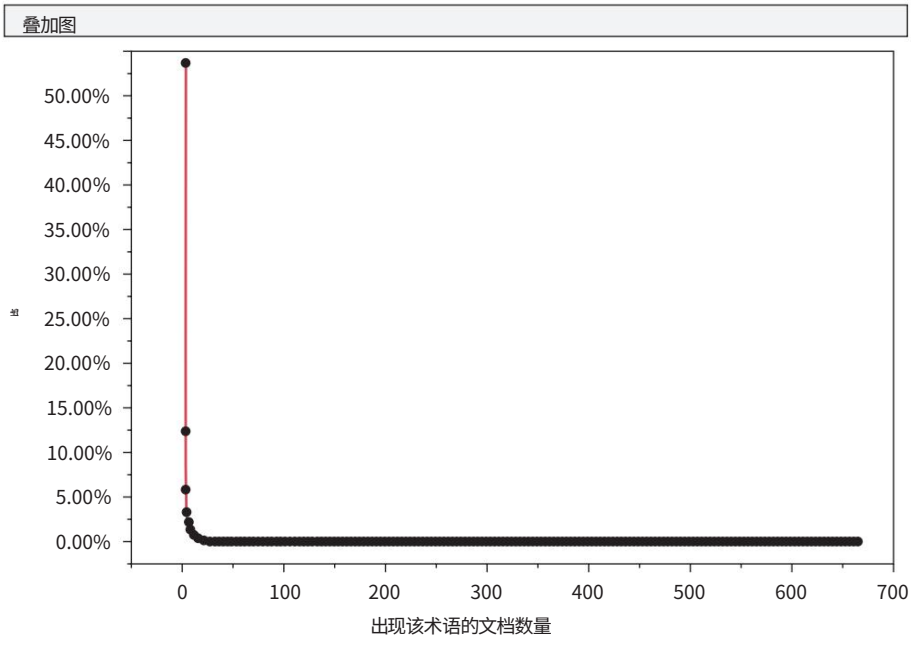


图 2 |按文档的词频分布。

涉及稀疏表示的乘法是
由于算法避免明确执行 0 乘法,因此速度很快。

DTM 中出现的原始术语计数
经常被改造。如果语料库包含混合
短文档和长文档,长文档
（具有较高的术语计数)可能会主导 DTM 的后续分析。此外,词

出现在许多文档中的词（停用词)将为 DTM 贡献包含
大量条目的列。如果
这种趋势没有被考虑,这些大柱
条目将导致文档聚类不理想
结果。变换可以是局部的（变换后的
DTM 中的组件仅依赖于原始 DTM 中对应的组件)或全局

（变换后的分量还取决于原始矩阵中的其他分量）。

DTM 的权重

DTM 中的原始词频条目的几种局部变换通常用于

抑制与计数数据相关的右偏。
根据分析目标,它可能有用
使用实际术语频率 (TF) 或变换
使用零一（不发生或发生

特定术语)二进制表示、TF 的对数或其中之一按逆文档频率的缩放版本。

二进制转换将 DTM 中的非零条目替换为 1。这在执行主题提取时非常有用,可确保文档中讨论的每个主题都获得相同的权重,无论该主题在该文档中出现多少次。这还增强了从 DTM 构建的分类和回归树 (CART) 等分析的可解释性。14也就是说,知道术语的存在与感兴趣的响应之间存在关联比知道响应与文档中出现次数超过任意次数的术语相关联更有参考价值。三元加权方案与二进制相同,但如果术语出现两次或更多次,则二进制表示的值为 2 而不是 1。14 由于计数过程通常遵循泊松分布,因此有时会对词频计数应用对数转换。这种转换会抑制较长文档中高计数的存在,而不会像二进制加权方案那样牺牲太多信息。

全局词频-逆文档14变换可能是最常用的频率(tf-idf) 。不同， tf-idf变换考与仅依赖于 DTM 单个组件的局部变换虑了整个语料库中每个单词的相对频率以及每个文档的长度。它缩小了出现在许多文档中的术语的权重,同时也增加了只出现在少数文档中的术语的权重。这通常可以提高预测和聚类性能。

术语的逆文档频率(idf)
t是:

$$\mathrm{idf}_t = \log_2\left(\frac{N}{\mathrm{df}_t}\right),$$

其中N是语料库中的文档数量， dft是包含术语 t 的文档数量。如果

如果词出现在每篇文档中,则itsidf为 0。出现在大多数文档中的词几乎没有判别能力（因此删除了停用词） 。为了解决这个问题,我们将原始 DTM 中的 TF 乘以逆文档频率,这会降低出现在许多文档中的单词的权重。也就是说,文档d和词t的tf-idf权重15为：

$$\mathrm{tf}-\mathrm{idf}_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t。$$

同样地,逆文档频率可用于转换二进制或对数转换的DTM。

无论 DTM 的加权方案如何,较长的文档将由比较短的文档具有更大幅度的行向量表示。此属性将在后续分析中为较长的文档赋予更大的权重,并可能产生次优的聚类结果。为了防止这种情况,可以对转换后的 DTM 中的行进行归一化,以使每个文档向量的总和为 1:归一化的 DTM 将每个文档表示为其中出现的术语的混合。例如,如果通过将文档 D 的两个副本粘贴在一起来创建文档D′ ,则归一化后 D 和D′将相同。15

SVD DTM 往往
有几列非零项很少（Zipf 定律） 。其中许多列代表噪声,对于聚类文档或构建预测模型没有用处,如图 3 中的 NSF 摘要示例所示。对于较小的语料库,DTM 中的术语（列)通常比文档（行)多,从而导致后续 DTM 分析出现问题。因此,通常对加权 DTM 应用降维程序。

降秩 SVD 是文本挖掘中使用的标准降维技术。16 SVD 将 DTM 简化为列数更少的密集矩阵。新的（正交)列是原始 DTM 中列的线性组合,选择它是为了保留尽可能多的方差结构

文件名	1990 年 NSF 奖摘要文本	基础	啊啊	ab 放弃 abelian aberr				能力	阿比奥特	可	烧蚀 异常 绝上 中止 比比皆是				
a9000006	过去两百多年的商业开发.....	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a9000031	对鸡的研究提供了血清学和...	0	0	0	0	0	0	0	0	1	0	0	0	0	0
a9000038	这项研究是...正在进行的计划的一部分	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a9000040	该 SBIR 提案旨在 1 合成新的...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a9000043	奇泽姆博士将调查.....的基本方面	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a9000045	这项研究将研究计算的复杂性杜克大学将操作 RV CAPE HATTERAS.....	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a9000046		0	0	0	0	0	0	0	0	0	0	0	0	0	0
a9000048	斯克里普斯海洋研究所将运营...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a9000049	百慕大生物站将操作 RV WEAT....	0	0	0	0	0	0	0	0	0	0	0	0	0	0

图 3 |非常稀疏的国家科学基金会文档术语矩阵的一部分。

文件名	文本	SVD1	SVD2	SVD3
a9000006	过去两百年来的商业开发.....	0.0221022908	0.0011192494	0.0000365208
a9000031	对鸡的研究提供了血清学和.....	0.0265216975	0.0046012733	0.0127392977
a9000038	这项研究是主要项目的一部分...	0.0048169574	0.0295638292	0.0103079577
a9000040	该 SBIR 提案旨在 1 合成新的...	0.0178877164	0.0117685537	0.0073317583
a9000043	Chisholm 博士将研究……的基本方面。	0.0312545193	0.009322851	0.0066186216
a9000045	这项研究将研究杜克大学将操作 RV CAPE HATTERAS 的计算复杂性.....	0.0185357152	0.0231188763	0.015040584
a9000046		0.0073200182	0.0100655772	0.0113419483
a9000048	斯克里普斯海洋研究所将运营四个百慕大生物站,并将运营 RV WEATHI....	0.0063642026	0.0094867337	0.0114474443
a9000049	0.0084351651		0.0088749983	0.0124502687

图 4 稀疏的国家科学基金会文档术语矩阵简化为密集的三维表示。

尽可能保留原始 DTM。对于 DTM X,降阶 SVD 分解 16为：

$$X \approx UDV^t,$$

其中U是具有正交列的密集d × s （其中d是文档数， s 是减少的维数)矩阵（生成文档降维描述的映射）， D是非负对角矩阵条目（ s最大奇异值），V是一个稠密w × s （w是项数)矩阵,具有正交列,其中s作为 SVD 分解的秩（对于s∈{1, ⋯, min(d, V为我们提供了术语的降级描述的映射。U 和V的s列按相应奇异值的大小降序排列。

例如,设置s=2 将分别给出文档和术语空间的最佳可能二维表示。 s的适当值是一个有争议的问题,并且取决于应用程序。

s值越小,表示降维程度越大,但代价是原始 DTM 的结构会丢失。通常使用 30 到 300 之间的值。在实践中,使用累积碎石图来选择s ,以便恢复大约 75-80% 的原始方差,往往会取得良好的平衡。

计算降阶 SVD 的一个重要计算工具是增强隐式重新启动 Lanczos 双对角化方法。 17该例程允许近似s最大奇异值及其相关奇异向量,而不需要计算所有的奇异值。可通过R中的irlba包获得,18该例程需要 3.8 分钟来近似 NSF 语料库的前 300 个奇异值以及左右奇异向量（同样,大约 9500 行每个 150 个单词）,而标准 R svd函数需要 7.1 分钟来计算相同的向量（以及所有奇异值）。随着维数的增加,这种时间差距被放大。

寻找主题和主题在自然语言处理中,使用降秩 SVD 称为 LSA。15 LSA 可有效提取主题并进行建模。我们还将讨论使用 SVD 的替代方法,称为潜在狄利克雷分配 (LDA)。一种流行的 LSA 技术是使用 DV矩阵产生的前两个向量绘制语料库词典。将相似的单词（在同一文档中频繁出现的单词,或在整个语料库中与常见单词集频繁出现的单词)绘制在一起,并且通常可以将粗略的解释分配给图中出现的维度,具体取决于 DTM 的加权方案。图 4 中,SVD 将 NSF 示例的 DTM 降低到 s=3 维。

SVD 与主成分分析 (PCA) 方法相关。这两种方法的区别在于,我们不会在计算 SVD 之前减去X的列均值,因为这样做会产生一个密集矩阵,而对这个密集矩阵进行 SVD 计算的计算量会大得多。虽然计算 m × n矩阵X的主成分的标准表示涉及协方差矩阵 XTX 的特征分解，但该过程在数值上可能不稳定。相反， X的右奇异向量产生 XTX 的特征向量,而X的奇异值产生XTX特征值的平方根。PCA的许多直觉适用于降秩 SVD 表示。

因子分析通常在主成分分析之后进行,用于研究原始变量之间的关系

以及所产生的成分（现在称为因子）
来自 PCA。 “因子载荷”描述了成对的
原始变量（项)与
因子,并通过将特征向量乘以其相关特征值的平方根来产生。因此UD
产生的因子载荷描述

文档空间， VD为
术语空间。当VTV =I 时,近似值XV ≈ UD
允许我们将新文档（使用原始字典)映射到同一空间,而无需

重新运行 SVD。也就是说,DTM 的降维表示是通过将 DTM 与其右
奇异向量(V)后乘而形成的

对应于s个最大奇异值。什么时候
新文档被添加到语料库中,它们可能
通过删除单词映射到同一个空间
没有出现在原始语料库中,并将新文档的 DTM 与右侧的

原始 DTM 的奇异向量V。虽然这个
避免重新计算 SVD,它还允许
用于预测建模的组件，
因为否则组件的组成
会随着每个新的 SVD 的变化而改变。

DTM 的 SVD 的方差最大旋转

当
几个术语与每个因素高度相关。它
结果表明,表示
V跨越的线性子空间可能会导致
一个更容易理解的表示。
选择旋转来产生加载矩阵
结构简单,19其中每个因素都将强烈
仅与少数原始变量相关，
而其他相关性应该接近于零。20
“刚性旋转”只是基本向量的变化
通过乘以表示线性子空间
正交矩阵。
我们可以保留方差结构
载荷矩阵L= XV ≈表示的数据
UD通过后乘任意正交矩阵 T 来实现。LT的协方差矩阵为

T =LTTTLT =LLT,与协变量LT(LT)相同
L的 ance 矩阵。因此， XVT的表示
文档空间保留了原始相关性
XV所表示的文档空间的结构
同时提高结果的可解释性
因素。或者,可以应用旋转来
术语空间VD。
不同的因子旋转施加不同的标准-21方差最大旋转22
选择T时在LT上进行 ria 。
是最广泛使用的因子旋转之一,因为它
有效性记录。方差最大标准选择
T使得LT中的大多数条目都接近 0 或 1
通过最大化列内方差之和

LT:这使得大负载变大,小负载变小。进一步讨论请参阅
参考文献 23-26
方差最大旋转,包括参考文献 27 的简明
方差最大目标函数的陈述。参考文献
21 和 28 讨论了方差最大旋转

文本挖掘。方差最大标准可用于
选择一个旋转矩阵,调整为
SVD 生成的加载矩阵： UD或VD。
前者将尝试创建将文档清晰分开的因素,而后者将寻求

干净地区分术语的因素。这些表示对于主题提取比原
始表示更有用
SVD 输出。
为了说明这一点,我们获得前两个 SVD 因子
NSF Abstract 标准化二进制 DTM 的载荷
（在词干提取之后,删除停用词和
出现在少于 15 篇摘要中,并减少
排名到s=300)。
第一个组成部分由术语主导
将要 。接下来是“研究”、“使用”和“学习”
最大的术语。第二个组成部分对“奖项”和“学生”有
较大的正权重。“研究”
在两者上都有相对较大（在数量级上)的权重
因素。图 5 中这些因素的双变量图
没有对 1990 年的主要研究主题给出太多的见解。

图 6 清楚地说明了在应用方差最大旋转（调整为
UD)后,最大的成分
涉及遗传学研究的主题。第二个
组件可能被解释为涉及计算机算法的开发,包括并行处理和模拟。由
于方差最大旋转

减少了项具有大载荷的程度
在多个组件上,我们可以将s 个组件视为单独的主题（图
7）,而不必
像在 LSA 中通常做的那样检查双标图。这
促进主题提取和
预测模型,因为重要因素可能
指定了一个解释。

概率主题模型
概率主题模型（例如潜在 Dirich-let 分配 (LDA))提供了 LSA
的替代方案
使用 SVD。 LDA 仍然使用“词袋”
与 DTM 的方法相同,但使用不同的方法
用于分析。一般的想法是,有一组术语定义一个主题,每
个文档
包含一个或多个具有一定概率的主题。 Blei29定义 LDA 的目标是
自动发现隐藏（潜在)的主题

一个统计模型来分配主题的概率
包含在特定文件中。文本数据是
被视为来自具有隐藏变量的生成过程,其中存在联合分布

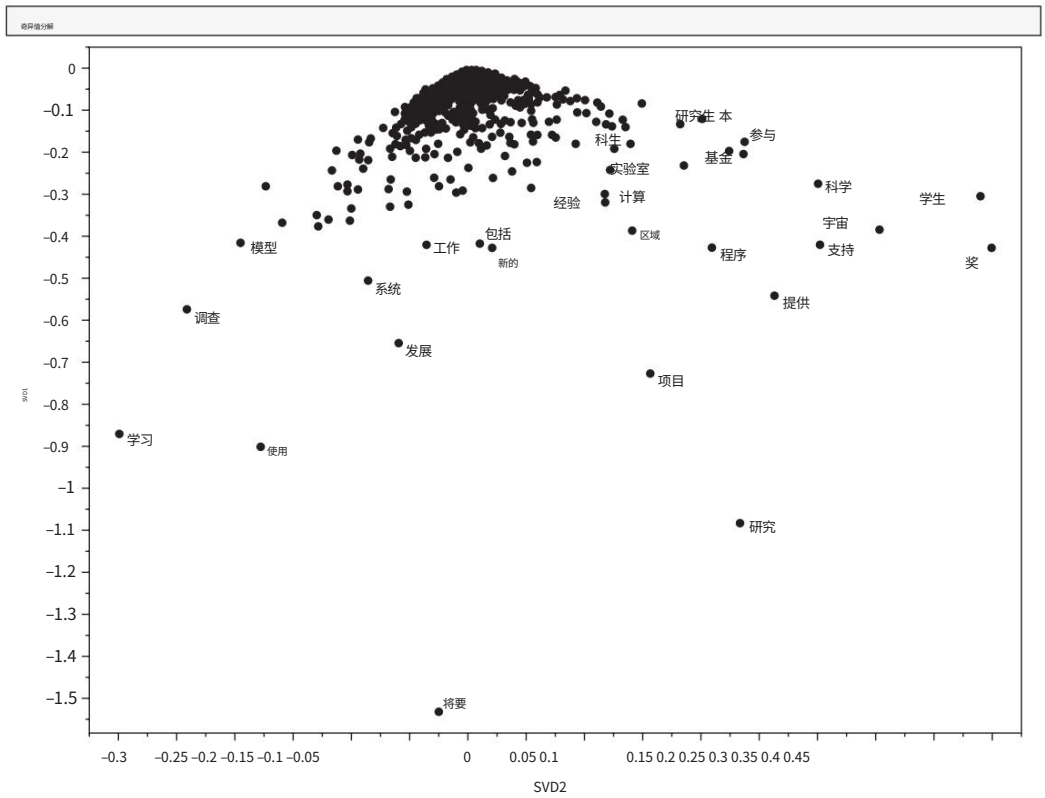


图 5 |美国国家科学基金会摘要的前两个奇异值分解因子图没有显示真正的见解。

观察到的（文档中的单词)和隐藏的（随机变量）。目标是在贝叶斯上下文中找到后验分布（给定观察到的文本），这将允许主题提取、信息检索、相似文档聚类和单词分组探索。

感兴趣的隐藏随机变量是哪些单词被分配给哪个主题、每个文档中主题的比例以及语料库中的主题分布。狄利克雷分布是多项式的共轭，其特征在于正向量的总和为 1。例如，特定文档的主题比例需要总和为 1，就像主题的单词成员资格一样。请注意，LDA 模型的一个显着特征是它们是混合成员模型，允许对每个文档多个主题进行建模，并假设主题是独立的。仅基于 DTM 估计这些隐藏随机变量的后验分布很棘手，需要先进的近似算法。

Blei 和 Lafferty30建议采用变分期望最大化和马尔可夫链蒙特卡罗（例如吉布斯采样)方法等。

Blei29指出 LDA 是一种正在发展的方法论，并且有许多有用的实际应用

现在可以通过放宽一些严格 LDA 公式的假设来实现。相关主题模型 (CTM) 放宽了后验近似中主题独立性的要求。在我们的 NSF 示例中，CTM 表示单词 parallel可能与计算机主题的关联性大于与海洋学主题的关联性。动态主题模型可以通过结合文档输入 DTM 的顺序来显示随时间变化的差异，就像在经度研究中一样。Griffiths 等人 31通过对语法进行建模而不是对一般的“词袋”进行建模来扩展 LDA 方法。

Grun 和 Hornick4开发了主题模型包，该包增强了 R 函数tm 和lda ，以允许后验分布的多种近似方法。该软件包还可以使用 CTM 提供估算。尽管您可以通过使用训练和验证集迭代地求解最佳数量，但必须事先指定主题的数量。此外，需要为吉布斯采样应用程序中的贝叶斯估计指定主题术语先验分布的估计。输出包括每个文档最有可能的主题、每个文档上术语的权重分布、后验分布以及对新文档进行分类的能力。

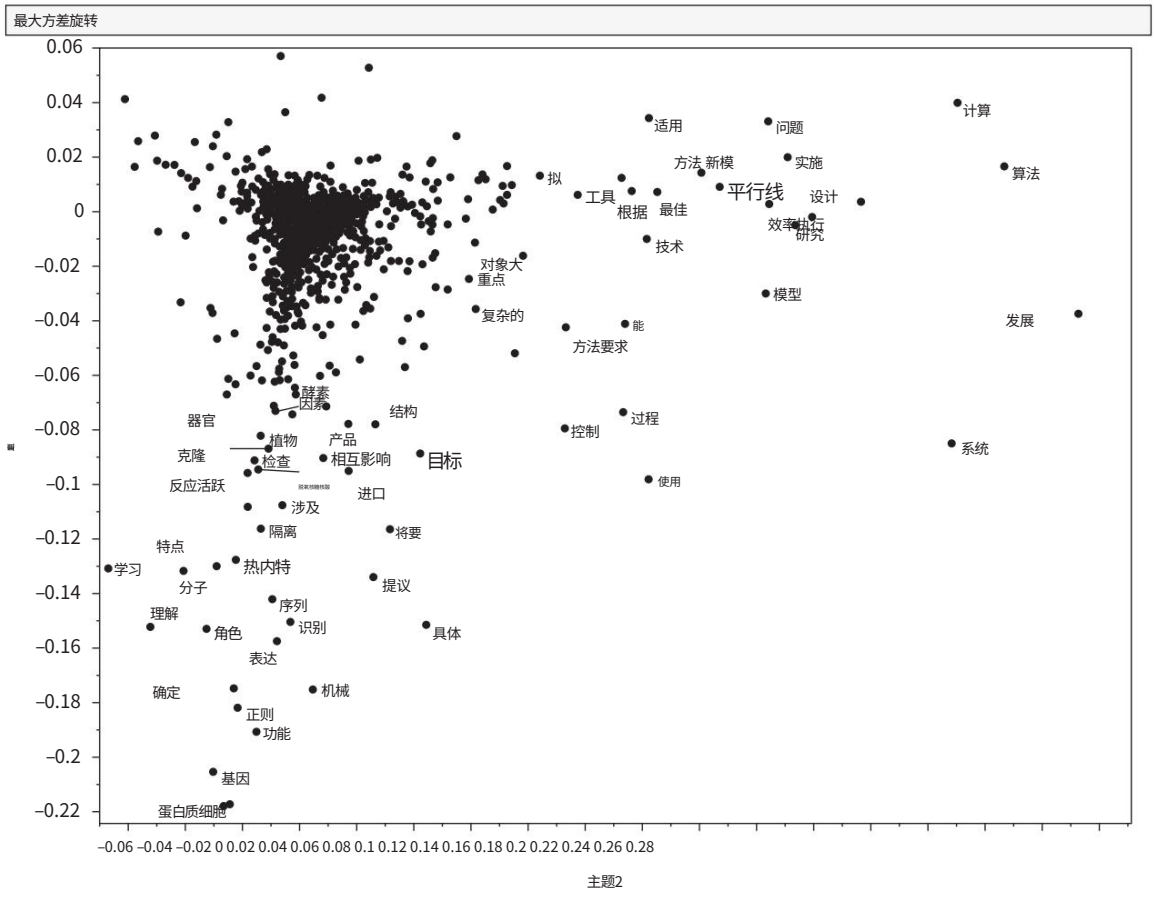


图 6 |前两个最大方差旋转奇异值分解因子的图清楚地显示了遗传算法和计算机算法研究。

前 5 个主题的前 10 个术语来自使用向量
期望最大化选项包括：

	主题 1	主题 2	主题3	主题4	主题5
[1,]	“政治”	“车间”“平行”		“代数”“女性”	
[2,]	“配体”	“次要”“机器”“多项式”			“矩阵”
[3,]	“血”	“海浪”	“图”“逻	“机器人”	“日程安排”
[4、]	“手性”	“宣传”	辑”	“戒指”	“家庭”
[5、]	“苏联”“加强”“工作状态”“单数”				
[6，]	“钙”	“禁用”“语言”“组合器”			“工作”
[7，]	“合法的”	“承诺”“愿景”		“算术”	“矩阵”
[8、]	“态度”	“联盟”	“编译”“定理”“省略”		“性别”
[9,]	“微管”	“博士”	“逆”	“一致”	“逻辑”
[10，]	“派对”	“荣誉”	“猜想”		“道德”

类似于方差最大旋转特征向量使用 SVD 的 LSA,这些主题应被视为探索性大于确定性。它们通常提供但可能不像
希望。例如,本例中的主题 1 似乎有几个不同的重点领域。首先三篇 NSF 摘要被认定为具有主要的“主题 1”加载,它们都集中于化学领域比政治态度更重要和苏联研究。主题3好像比较多前10个单词的定义是同质的。这些主题是使用默认值生成的总共 30 个主题。如果该值增加,则有每个主题上的歧义较少,但应谨慎尽量不要有太多主题。

同样,前 5 个主题的前 10 个术语使用吉布斯抽样的方法是：

	主题 1	主题2	主题3	主题4	主题 5
[1,]	“海洋”	“地震”“变形”“同位素”“流形”			
[2，]	“循环”	“土壤”	“拉紧”	“岩石”	“拓扑”
[3，]	“印度”	“地震”	“陶瓷”“矿工”“单一”		
[4,]	“浮游植物”“损害”“微结构”“构造”“超级计算机”				
[5、]	“颜色”	“过错”	“熔化”	“曼特尔”“对称”	

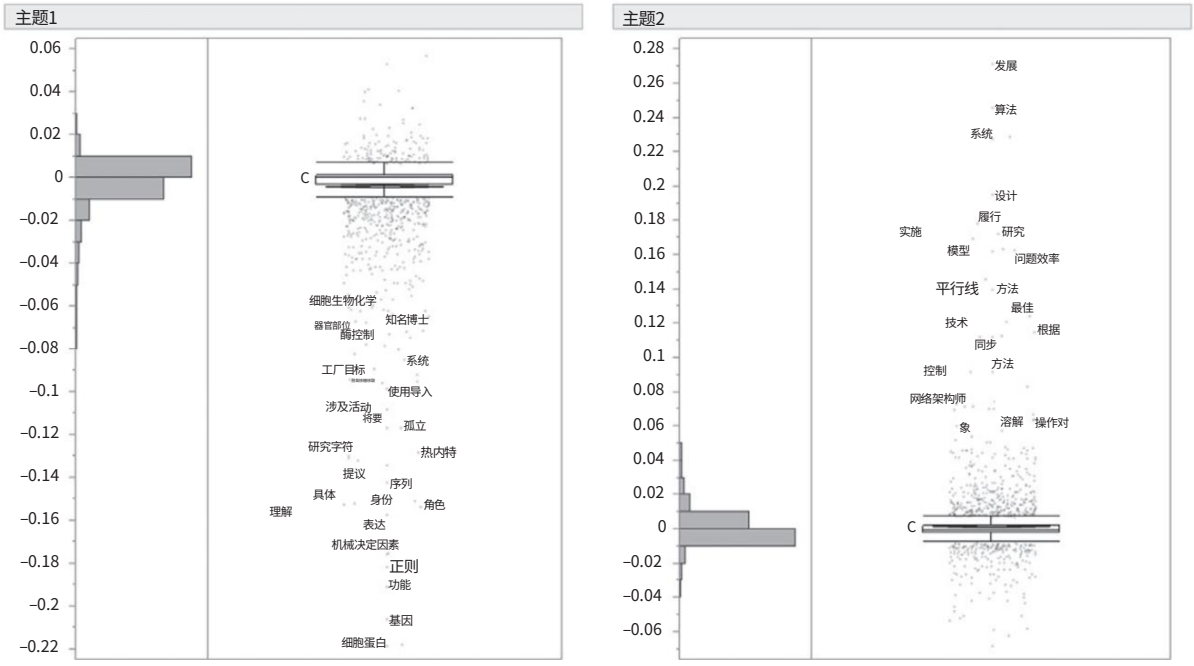


图 7 使用最大方差旋转的奇异值分解因子的独立主题。

	主题 1	主题 2	主题 3	主题 4	主题5
[6、]	赤道	“普里埃塔”	“失败”		“国家安全网”
[7，]	“印度”	“洛马”	“合金”	“ridge” “plate” “harmon”	
[8、]	“陷阱”	“岩土工程”	“粮食”	“火山”	“省略”
[9，]	“臭氧”	“旧金山”“塑料”		“地壳”	“轨道”
[10、]	“追踪者”	“墙”	“折叠”	“脆皮”	“否定”

从前五个主题的样本和我们的
与其他例子的经验,吉布斯采样器
具有更多可解释且分离的主题。

最后是不需要的CTM选项
线性独立主题显示每个主题的热门词的相当同质的集合。

	主题 1	主题2	主题 3	主题4	主题 5
[1,]	“冰”	“膜”“翻新”		“次要的”	“粒子”
[2、]《沉积物》		“大脑”“标本”		“土壤”	“太阳的”
[3、]	“盆地”	“受体” “博物馆” “森林”			“加速器”
[4、]	“tecton”	“激素”“arfmp”		“视频”	“等离子体”
[5，]	“南极”	“神经元”	“公司”	“筛选”	“RNA”
[6，]	“海洋”	“神经” “数据库” “强化”			“床”
[7，]	“北极”	“肽”	“演讲”	“雷乌”	“电离层”
[8,] “沉积” [9,]		“神经”	“植物”	“离心机”“气溶胶”	
	“动物群”	“紧张的”	“巢”	“联盟”	“探测器”
[10、] “流”		“游戏”	“馆长”	“博士”	“风”

聚类的应用和
树模型

回到带有降级 SVD 的 LSA,我们可以
使用 DTM 和特征向量来发现其他
结构。既然语料库的文本已经
转换为数据矩阵,其中包含文档（或段落、调查回复、事件报告等）
行和术语作为列,我们打开了
使用标准统计和机器的可能性
学习工具来回答所提出的问题
在介绍中。完整的 DTM X确实有
一些特殊性（稀疏性,所有条目都是非负的）,需要对现有的进行一些修改
方法,14将在下面讨论。相比之下，
降秩 SVD 提供了一种稠密的表示
XV更适合的文档空间
现有的惯例。

聚类
虽然主题提取过程可能很有用
为了识别语料库中的流行主题，
文档很可能同时在多个主题上“活跃”。聚类分析提供了同时检查所有因素的文档相似性的能力。28聚类分析的结果

也可用于提供语料库的代表性子样本。每个簇中一份文档的分层样本将提供以下样本：

比整个语料库中的简单随机样本（相同大小）更能代表语料库。

完整 DTM 的聚类分析需要一些特殊的调整。存在不相关变量时，聚类算法结果的质量会受到影响。一个相关问题是 DTM 的稀疏性：并非所有统计软件包都提供稀疏矩阵功能。此外，正条目的重叠（两个文档之间的共享单词）比零条目的重叠（两个文档中都不存在的单词）包含更多信息。欧几里得度量对这些度量给予相同的权重，并且与余弦和 Jaccard 等其他度量相比，通常表现不佳。32 相比之下，DTM 的降级 SVD 表示可以与标准聚类软件一起使用。与其他聚类应用程序一样，Ward 的方法对于最多大约 30,000 个文档（取决于计算机内存可用性）往往表现良好，之后可以使用 K 均值。没有必要对列进行标准化，因为 SVD 已经对它们进行了适当的缩放。另一个优点是 SVD 提供的降维消除了通常会导致聚类算法出现问题的冗余/不相关变量。

对于包含短篇文档的申请，可以阅读每个集群中的文档样本，以便为集群分配一个主题。对于较长的条目，例如 NSF 摘要，可以对最大的集群进行单独的分析，以确定存在的主题（使用主题提取程序，或者在许多情况下，通过简单地检查 TF）。在 NSF 示例中，层次聚类表明大约有 400 个集群。最大的三个集群每个集群包含不到 90 篇摘要，涉及会议资助、举办教师研讨会和暑期项目以及本科生暑期研究项目。

分类和预测在某些情况下，能够预测未来观察的响应可能很重要。文本分析分类问题的一些示例包括识别欺诈性保险索赔、贷款违约、交叉销售或追加销售的潜在客户以及电子邮件垃圾邮件。同样，文档可用于预测索赔金额、维修成本或产品评级。处理新观察时必须特别小心：它们需要转换为训练数据上 SVD 所跨越的空间。如果X0是由具有相同列的新文档集合形成的 DTM 由于原始 DTM X ≈ UDVt （排除原始语料库中未出现的单词，并在必要时引入 0 列），则X0V （或X0VT，如果将旋转T应用于原始语料库）提供新文档的降级表示。

事实证明，从回归和朴素贝叶斯14到支持向量机33的复杂模型在文本挖掘中非常有用。由于文本分析是从筛选过程开始的，因此应将几种不同的建模技术与 DTM 的不同选项（加权、降级等）结合使用。这个过程可以说是一门艺术：经验有助于培养一种不易教授的直觉。虽然在拟合响应（资金、欺诈指标等）时使用多个设置拟合多个模型会增加误报的数量，但主题知识通常会很快消除虚假结果。剩余的重要结果可以通过交叉验证或理想情况下通过对新数据进行假设检验来进一步研究。

当（旋转的）SVD 向量用作预测因子时，随机森林提供了一种有用的方法：尤其是当对文档空间应用旋转时，文档通常会在因子内聚类为两个聚类。这种分组与 CART 方法创建的切点配合得很好，但与回归配合得不太好。方差最大旋转通常可以使 SVD 产生的因子的解释更加清晰，这在构建预测模型时非常有用。如果发现 SVD 因子与响应之间存在显著关联，那么就有可能为结果分配解释。我们将 Miner 等人提供的 3235 份飞机事故报告语料库 34 处理成tf-idf DTM，并获得了 150 个 SVD 因子的（词空间调整）方差最大旋转。我们使用 SVD 因子拟合二元死亡率指标。

柱贡献图表明主题 36 (SVD36) 与死亡率指标的相关性最强。相关主题图表明，主题 36 涉及飞行员继续按照目视飞行规则 (VFR) 飞行，而实际上情况已经恶化，需要仪表飞行规则 (IFR)。这与主题专业知识是一致的，即通用航空死亡事故在 VFR 和 IFR 之间的过渡中最为突出，其中飞行员可能对以低能见度为特征的仪表气象条件没有足够的资格或熟练程度。在后续的预测分析中，我们使用这些源自文本字段的重要 SVD 因素以及结构化变量（例如，一天中的时间、月份、位置、风速）。

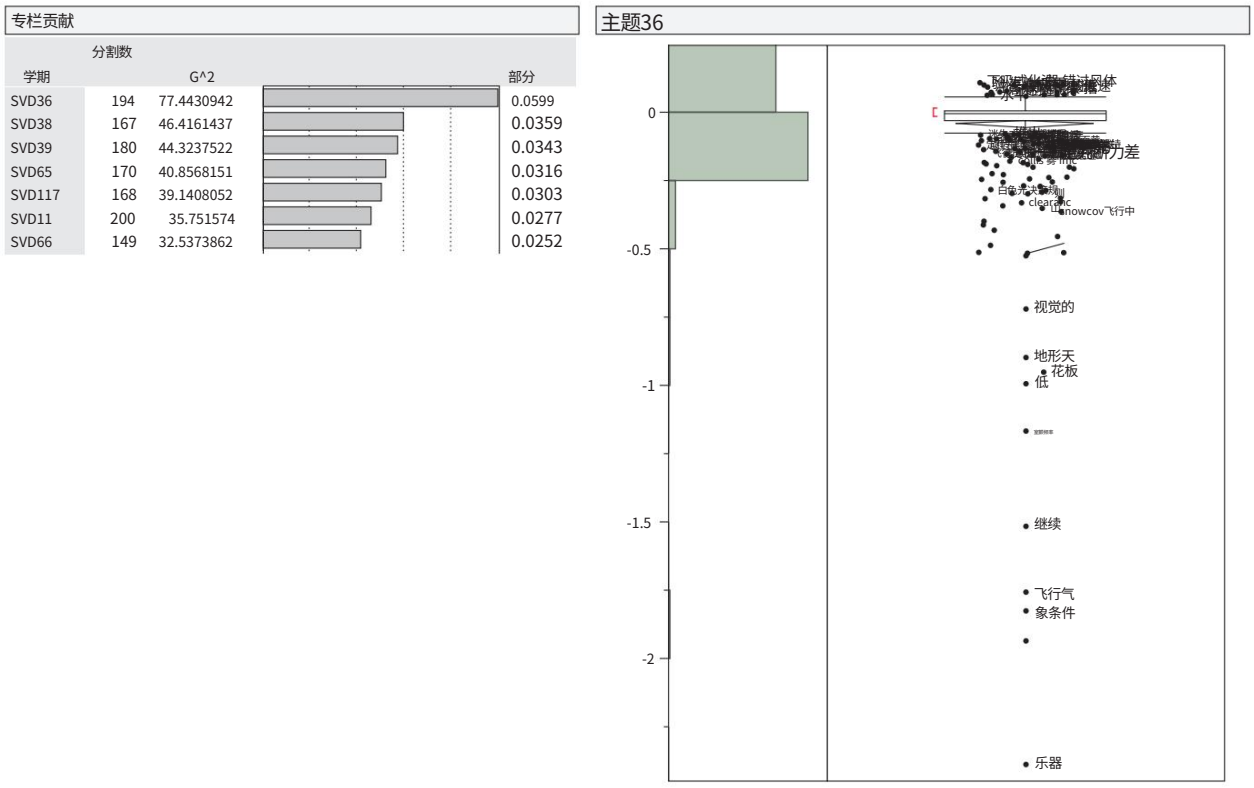


图 8 |与死亡最相关的奇异值分解因素。

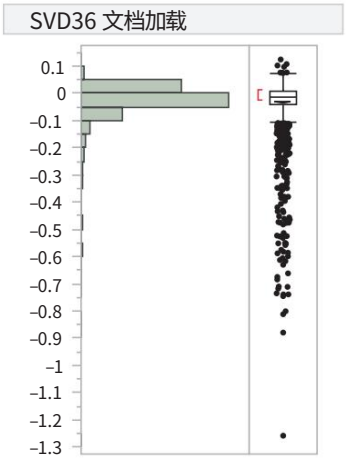
图 8 中主题 36 的文档加载表明哪些文档在该主题上处于活动状态。
根据主题 36 个文档加载进行排序并阅读前几个条目证实了我们对主题图的解释（图 9）。

图 10 中使用 DTM（二进制加权)中的单词列作为输入变量的死亡响应分类树提供了术语之间有趣的关系。每个节点的死亡人数比例以及记录数（计数)以蓝色表示。如果“土地”出现在叙述原因文本中,那么预计死亡人数会很少。那些包含“山”的内容与死亡有关。如果文本中没有“陆地”和“低地”,那么死亡的可能原因是失速或旋转。

由于文本中没有“land”一词,因此当叙述中出现“low”时,dark（黑暗)和autorotate（自动旋转)就是用来解释死亡的主动词。

情感分析虽然词袋方法丢弃了
有关单词出现的语法上下文的信息,但单词仍然通过其定义携带信息。心理学家开发了哈佛 IV-4 词典35,其中包含大约 1600 个积极词汇和 2000 个消极词汇。对于每个 DTM 行

（对应于特定文档、段落、记录等)统计正面单词实例的计数和负面单词的计数。这些情绪列表在来自广泛学科的语料库中表现得相当好。例如,它们可用于根据主题讨论的基调对新闻文章或 Twitter 帖子进行排序。然而,当通过开发自定义情感列表、根据领域知识或给定一组已分类（例如,正面、负面)或评级（例如,总体)的训练数据来分析来自特定源的文本时,可以获得更好的性能。满意度）。36–38为了说明:使用车主评论语料库以及数字总体满意度评分,39可以构建一个自定义情感分析列表,用于预测新评论的总体评分。通过构建套索回归并另外拟合 CART 以使用二进制 DTM 的列（不是 SVD 表示)对总体用户评分（从 1 到 5 缩放)进行建模,可以找到哪些单词与总体评分相关联。列表中的单词{传输、柠檬、吱吱声、廉价、塑料、修复、问题、可怕、经销商、发送、索赔、窗口等}与较低评级的相关性最强,而列表中的单词{好、手
柄,平稳,加速,伟大,可靠,



narr_cause~飞行	来自 U 的 SVD36
员继续目视飞行进入仪表气象条件。	1.2586525754
飞行员继续在仪表气象条件下飞行。造成这种情况的因素包括雾、低云层、……	0.879509973
飞行员在机动时继续按照仪表气象条件进行目视飞行规则飞行。事故发生的因素...	0.8118630715
飞行员持续目视飞行规则飞行进入仪表气象条件导致飞机相撞...	0.8011472086
飞行员继续按照目视飞行规则飞行,进入仪表气象条件,随后与地形相撞.....	0.745395256
飞行员的判断失误,故意在仪表气象条件下启动 VFR 飞行。共...	0.7393123011
巡航飞行期间,飞行员继续 VFR 飞行进入仪表计量条件。与此相关的因素...	0.7390912097
飞行员继续在仪表气象条件下进行目视飞行,并且未能保持地形净空……	0.7351538413
在仪表气象条件下,飞机在飞行中因不明原因失控。	0.7176443542
飞行员遭遇仪表气象条件并随后失去对飞机的控制。	0.7114139375
飞行员继续在仪表气象条件下目视飞行,随后与树木相撞……	0.7101342894
飞行员继续在仪表气象条件下进行目视飞行规则飞行,随后在飞行中与飞机相撞……	0.6842974747
飞行员在飞行过程中决定继续目视飞行进入仪表气象条件,从而导致了事故发生.....	0.6748188452
飞行员有意继续按照目视飞行规则 (VFR) 飞行进入仪表气象条件 (IM...	0.6617060754
飞行员继续目视飞行,进入仪表飞行条件,导致飞行中与地形相撞。.....	0.6300856776
飞行员故意飞行进入仪表气象条件并且未能保持与地形的间隙...飞行员在巡航期间无意中目视飞行规则（VFR)飞行进入仪表气象条件	0.6168120814
（IMC）...	0.6126443124
飞行员未能与障碍物（天线塔)保持足够的距离,并继续飞行……飞行员无意中按照目视飞行规则飞行,进入仪表气象天气条件,导致飞行中	0.6056780679
发生碰撞……	0.5989421647
飞行员继续在目视飞行规则 (VFR) 条件下飞行,进入仪表气象条件,导致飞机与山区地形相撞。	0.588182416

图 9 |与主题 36 的低文档负载相关的记录。

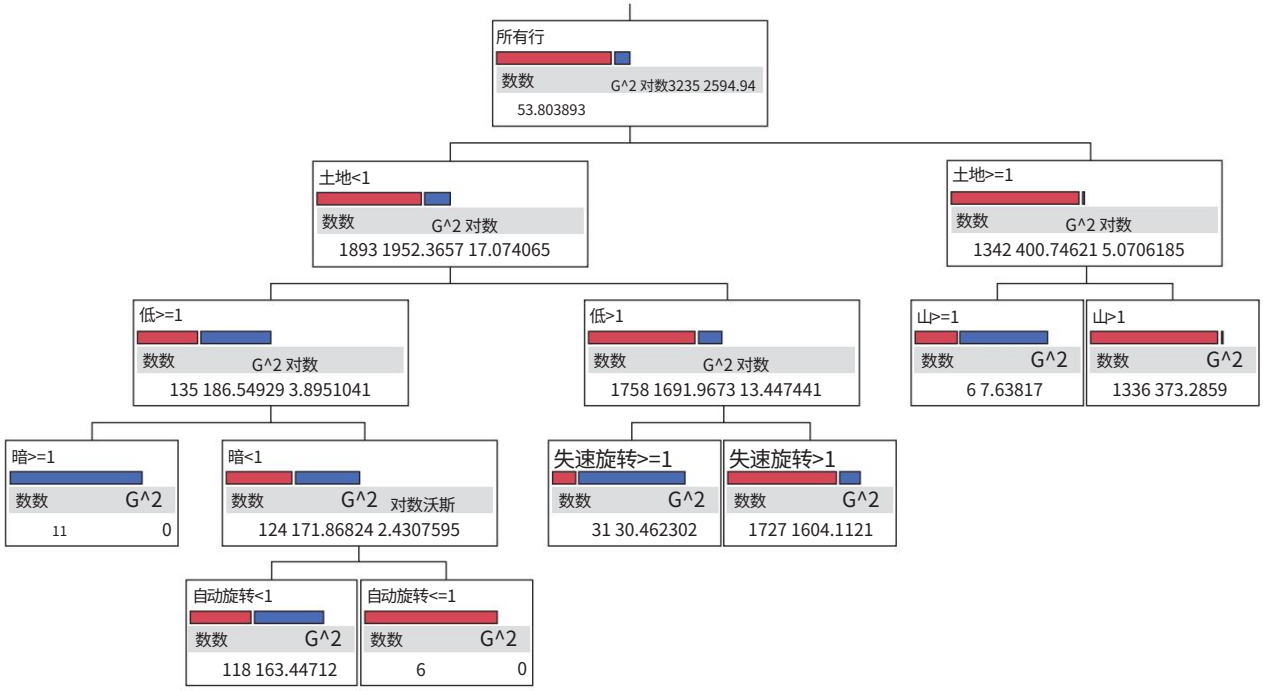


图 10 |以词语为因素的死亡率反应变量的分类树。

导航、低音、快速、舒适等} 与更高的评分相关。否定可能会对这种方法造成问题（例如，“不可靠”）。一个快速的解决方案是在形成 DTM 之前将语料库中所有出现的“不”连接到下一个单词,尽管也有更复杂的方法可用。40,41

使用 Perl、Java 和 Python 等开源实用语言修改或开发了特定工具。英格索尔等人。给出示例和 Java 代码来实现通过流行的开源 Apache 软件基金会提供的许多高级文本分析工具。3 R 编程语言提供了多个用于文本挖掘的开源包。9,42该语言的稀疏矩阵库维护良好;新发布的方法通常首先通过综合 R 存档网络提供;程序员能够将 C 代码编译成 R 包以加快处理速度,例如

软件

用于文本挖掘的软件选择有很多,每种软件都必须根据成本、功能和用户学习曲线进行权衡。许多组织

由 wordcloud 包完成。43然而,虽然 R
为定制程序提供了巨大的灵活性
行为。结果分析（尤其是图形分析）
与商业软件包相比,它可能具有挑战性且受到限制。我们发现导入

DTM 及其 SVD 分解后导入标准统计软件包（如 IBM SPSS 或 SAS JMP）
在 R 中构建矩阵。自动化此过程
使用插件自定义输入选项可以
快速探索和发现,同时利用用户现有的统计能力

应用程序。本文中的图形由
来自 SAS JMP。
一些流行的全面商业文本挖掘程序包括 Clarabridge、SAS Text Miner
与 Enterprise Miner、ANGOSS、IBM 集成
SPSS Modeler Premium 和 StatSoft 的 STATISTICA。
Chakraborty 等人 2介绍了 SAS
文本挖掘器使用了许多在
本评论以及概念链接、词性标记和本体的高级组件

管理。
http://www.kdnuggets.com/software/text.html
提供完整的文本挖掘软件列表
包。它们被总结和分解
商业应用与开源应用。
此外,还提供了每个应用程序的链接
网站以供进一步审查。

结论

非结构化文本包含有价值的信息,可以
使用广泛可用的统计数据可以轻松利用
方法。只需将来自可能不同的来源的文本转换为可用的格式,并将行作为
将文档和专栏视为文字可能是一个挑战。我们展示了由 LSA 补充的词袋文本挖掘过程,将 DTM 转换为

通过 SVD 有意义的因素可以迅速导致
发现。方差最大旋转导致一般主题
从语料库中可以更清楚地加载 SVD 因子。 SVD-U 行的层次聚类

矩阵合并具有相似主题的文档
对 SVD-V 列进行聚类时会收集相似的单词。概率主题模型,例如

因为 LDA 在寻找共同点方面也非常有效
主题和分组文档。 SVD 本身可以用作与结构化数据相结合的因子来进行预测分析。决策树和

回归方法可以帮助识别解释响应变量的最显着的 SVD。情绪

分析确定阳性与阳性的比例
每个文档的否定词,尤其是
对于社交网络分析等新兴领域非常有用。我们的经验表明,这些方法

使用开源软件以低廉的成本实现
解决方案往往就是所需要的,因为语法和语义元数据的复杂性可能会提供
仅提供边际额外见解。

参考

1. Miner G、Elder J、Hill T、Nisbet R、Dursun D、Fast A.文本挖掘与统计分析实用指南
非结构化文本数据应用。牛津:Aca-demic Press;2012 年。

2. Chakraborty G、Pagolu M、Garla S.文本挖掘和分析:使用 SAS 的实用方法、示例和案例研究。北卡罗来纳州卡里:SAS Institute Inc;2013 年。

3. Ingersoll G、Morton T、Farris A。驯服文本。庇护所纽约州艾兰:曼宁； 2013年。

4. Grun B,Hornik K。主题模型:R 包拟合主题模型。统计软件杂志2011 年,40:1–30。

5. Apache 软件基金会。 OpenNLP。 2015年。位于： https://opennlp.apache.org/ （已访问 2015 年 3 月 24 日）。

6. Lichman M. UCI 机器学习库。加州大学欧文分校信息与计算学院计算机科学。2013 年。网址:http://archive.ics。uci.edu/ml（2015 年 2 月 13 日访问）。

7. Scrapinghub S。一个快速而强大的网络爬行框架。 2015。可在:http://scrapy.org/ （访问 2015 年 3 月 10 日）。

8. Temple Lang D. RCurl:通用网络（HTTP/FTP/ …) R. CRAN 客户端接口。 2015年。可用位于:http://cran.r-project.org/web/packages/RCurl/index。html（2015 年 3 月 10 日访问）。

9. Barbera P.streamR:访问 Twitter 流 API来自 R.2014。可在:http://CRAN.R-project 获取。org/package=streamR（2015 年 3 月 10 日访问）。

10. Bilisoly R.使用 Perl 进行实用文本挖掘。霍博肯,新泽西:约翰威利父子公司;2008 年。

11. 波特 MF。一种后缀剥离算法。程序1980 年,14:130–137。

12. 杨 C. 谁害怕乔治·金斯利·齐普夫?或者:儿童和黑猩猩有语言吗?意义2013,10:29–34。

13. Bates D, Maechler, M. 矩阵:稀疏和密集矩阵类和方法。CRAN。2015 年。可从以下网址获取:

http://CRAN.R-project.org/package=Matrix（2015年 2 月 13 日访问）。

14.Weiss S、Indurkhya N、Zhang T、Damerau F。文本挖掘：分析非结构化信息的预测方法。纽约：施普林格； 2005年。

15. Manning CD、Raghavan P、Schutze H。信息检索简介。纽约：剑桥大学出版社；2008年。

16. Albright R.使用 SVD 驯服文本。北卡罗来纳州卡里：SAS Institute Inc；2004年。

17. Baglama J、Reichel L. 增强隐式重启 Lanczos 双对角化方法。SIAM J Sci Comp-put 2005，27：19-42。

18. Baglama J、Reichel L. irlba：通过隐式重启双对角化实现快速部分 SVD。Lanczos
CRAN。2015 年。网址：http://cran.r-project.org/web/packages/irlba/index.html（2015 年 2 月 13 日访问）。

19. Thurstone LL. 《多因素分析》。伊利诺斯州芝加哥：芝加哥大学出版社；1945年。

20. Kline P.因子分析简易指南。纽约：
劳特利奇；1993 年。

21. Visinescu LL，Evangelopoulos N.潜在语义分析中的正交旋转：一项实证研究。Decis Support Syst 2014，62：131-143。

22. Kaiser HF。因子分析中解析旋转的方差最大标准。Psychometrika 1958，23：187-200 。

23.里奇曼MB。主成分的旋转。
气候学杂志1985 年，6：293–335。

24.邓特曼GH。主成分分析。加利福尼亚州纽伯里帕克：Sage Publications, Inc；1989年。

25. Ramsay J、Silverman BW.功能数据分析。
纽约：Springer；2005 年。

26. Pett MA、Lackey NR、Sullivan JJ。理解因子分析：在医疗保健研究中使用因子分析进行仪器开发。加利福尼亚州千橡市：Sage Publications, Inc； 2003年。

27. Khattree R，Naik DN。使用 SAS 软件进行多变量数据缩减和区分。北卡罗来纳州卡里：SAS Institute，Inc；2000 年。

28. Evangelopoulos N，张X，Prybutok VR。潜在语义分析：五项方法建议。欧洲信息系统杂志2012 年，21：70–86。

29. Blei D. 概率主题模型。 Commun ACM 2012，55：77–84。

30. Blei D、Lafferty J. 主题模型。收录于：Srivastava A、Sahami M 编。文本挖掘：分类、聚类、
和应用程序。 Chapman & Hall/CRC 数据挖掘和知识发现系列。佛罗里达州博卡拉顿：查普曼和霍尔； 2009年。

31. Griffiths T、Steyvers M、Blei D、Tenenbaum J。整合主题和语法。高级神经信息处理系统， 2005 年 17 期，537–544。

32. Strehl A、Ghoash J、Mooney R。相似性度量对网页聚类的影响。见：网络搜索人工智能研讨会， AAAI， 2000 年，58-64。

33. Jiang EP。使用机器学习算法进行基于内容的垃圾邮件分类。收录于：Berry MW、Kogan J 编。文本挖掘：应用与理论。西萨塞克斯：John Wiley & Sons；2010年。

34. Nisbet R、Elder J、Miner G. 《统计分析与数据挖掘应用手册》。马萨诸塞州伯灵顿：Academic Press；2009 年。

35. Kelly E、Stone P. 探究者类别描述及探究者词典的使用。网址： http://www.wjh.harvard.edu/~inquirer/homecat.htm （2015 年 3 月 10 日访问）。

36. 唐华, 谭世杰, 程晓燕. 评论情绪检测研究综述. Expert Syst Appl 2009, 36: 10760–10773.

37. da Silva NFF、Hruschka ER、Hruschka Jr. ER。使用分类器集成进行推文情绪分析。Decis Sup-port Syst 2014，66：170-179。

38. Prabowo R、Thelwall M. 情绪分析：一种组合方法。J Informetrics 2009，3：143-157。

39. Lahoti S，Mathew K。文本挖掘：使用 STATISTICA 数据挖掘器和文本挖掘器进行汽车品牌审查。
引自：Nisbet R、Elder J、Miner G 编。《统计分析和数据挖掘应用手册》。马萨诸塞州伯灵顿：Academic Press；2009 年。

40. Cruz Diaz N、Vazquez J、Alvarez V. 一种用于临床文本中否定和推测检测的机器学习方法。J Am Soc Inf Sci Technol ，2012，63：1398-1410。

41. Carrillo-de-Albornoz J，Plaza L。基于情感的否定、增强器以及极性和强度分类模态模型。 《美国社会信息科学技术杂志》 2013 年，64：1618–1633。

42. Feinerer I、Hornik K、Meyer D. 文本挖掘基础设施，R. J Stat Softw ，2008，25：1-54。

43. Fellows I. wordcloud：词云。CRAN。2014 年。
网址：http://cran.r-project.org/web/packages/wordcloud/index.html（2015 年 2 月 13 日访问）。