文本分析实录：原理、应用与操作

Text Analysis: principle, application and operation

左祥太 (Shutter Zor)

厦门大学管理学院会计学系
Department of Accounting
School of Management
Xiamen Univeristy

30, July, 2023

厦門大學
XIAMEN UNIVERSITY

个人介绍

简历

- Bilibili: 拿铁一定要加冰
- 微信公众号: OneStata
- Stata 第三方包: oneclick与onetext
- 文章发表于 *Journal of Cleaner Production*、*Technology Analysis & Strategic Management*、*Heliyon*、*Plos One*、《科研管理》等
- 邮箱: *shutter_z@outlook.com*

目录

本节内容

背景介绍

常见方法

参考文献

什么是文本分析

文本分析（Text Analysis）也称文本挖掘（Text Mining），是从非结构化的原始文本中提取有效信息并生成相关数据的分析范式 [1]，是自然语言处理（Natural Language Processing）的具体应用方式。

自然语言处理起源于 Alan Turing（1950）[2], [3]，在会计、公司金融领域的实际应用当中，该方法主要是对相关文本的关键信息进行提取，并逐渐演化出了相对固定的分析方法。

实证写作中的应用场景

在实证论文写作过程中，通过文本分析方法构建指标的方式主要包括基于统计的文本挖掘与基于神经网络的词向量构成。前者主要通过基于词典构造文本向量的方法对文本进行量化，如数字化转型 [4]、管理者语调 [5]–[8]、环境规制程度 [9], [10]、文本相似性 [11], [12] 等；后者主要通过神经网络生成词语向量的方法对词语进行向量化，以实现词语类比与困惑度计算，如扩充经济政策不确定性词语、计算文本可读性等 [13]–[15]。

本节内容

词频统计

词频分析是分析文本中特定词语的出现频率，常见于文本情绪分析，以及类似于数字化转型、环境规制等指标的构建。

词频统计-人工判读

以文本的情绪分析为例，在文本分析技术得到广泛应用以前，学者们一般通过人工阅读报告的方式判断文本中所蕴含的情感，而人工阅读的方法存在一些缺陷，比如，不同人对同一文本的情绪判断可能存在差异。为了消除这种差异，一般采用多人阅读同一文本判断情绪后计算情绪均值的做法，但该方法也会受限于阅读者个人学历、认知、阅读习惯等因素，并且实施成本较高。

词频统计-机器识别

计算机虽然不能对文本产生直接的情绪判断，但通过告知计算机哪些词语具有积极情绪，哪些词语具有消极情绪的方式，能够让计算机在特定文本中寻找这些词语出现的频率。在检索方面，计算机具备非常高的处理效率，并且误差较低。通过分别检索含有积极情绪信息的词语和消极情绪信息的词语，然后选取一些简单的算法，就能实现通过计算机判断文本情绪的功能 [8], [16]–[22]。

词频统计-详细步骤

以文本情绪的量化为例（其余方法类似），具体而言，利用计算机软件进行文本情绪计算的主要步骤为：

1. 选择恰当的词典。如果分析文本情绪，则需要确定适用于特定文本的情绪词典，包括积极情绪词典与消极情绪词典。常见的情绪词典包括 Loughran-McDonald 词典 [8]、中文情感极性词典 [16]、清华大学李军中文褒贬义词典、知网情感分析词典以及 Henry 词典 [17]。
2. 计算每个词语在对应文本中出现的次数，并将其赋予给某一变量进行存储
3. 通过一些公式计算相应的指标，常见的一些公式如下：
   - 3.1 $Sen = \frac{Pos-Neg}{Pos+Neg}$，该方法应用于 [5]–[7], [17], [18]
   - 3.2 $Sen = \frac{Pos(Neg)}{Tot}$，该方法应用于 [19], [20]
   - 3.3 $Sen = \frac{Pos-Neg}{Tot}$，该方法应用于 [21], [22]

词频统计-代码实现

接下来分别以Python与Stata为例，展示在两款软件中分别计算文本情绪的方法。本部分代码改编自我的公众号推文：

- OneStata: 【NLP】文本情绪指标综述及 Python 实现
- OneStata: 「Stata」词频统计下的数字化转型

向量空间模型

向量空间模型（Vector Space Model）由 Salton 等（1975）提出 [23]，是文本分析的另一个方向，与基于词频统计的文本分析方法不同，该方法侧重与以向量形式将文本进行“翻译”，并期待通过向量之间的关系重构文本之间的关系。将一系列的文本向量化后，便可以通过向量之间的数学运算关系，生成文本之间的联系，比如通过向量之间的相似度关系来实现相似文本之间的识别。同样地，在该方法出现以前，学者们也多采用人工方式对文本进行相似辨别，当文本长度过长时，人工误判的概率就会大大增加。

向量空间模型-向量生成

首先，也就是最重要的一步，在对文档进行向量化之前，需要首先确定用于向量化的词典（特有词典或者全部词语）。文档向量的维度上限取决于所选词典的词语数量。假设我们有如下词典：

$$dict = a, b, ..., z$$

该词典记录了 26 个小写英文字母，并将作为我们的词典。需要注意的是，这个词典中的词语数理论上是无上限的，也可以是任意专有名词所有构成的。

向量空间模型-向量生成

那么，对于我们的文本：

$$Sequence_1 = This\ is\ file\ 1$$

$$Sequence_1 = This\ is\ file\ 2$$

可以构造两种不同的基本向量，一种是Bool-based vector，另一种是Count-based vector。

- Bool-based vector，根据词典中的词是否在文本中出现，构造 1 和 0 的向量。
- Count-based vector，根据词典中词语出现的频率，构造基于词语频率的向量。

向量空间模型-向量生成-一个例子

假设我们有以下文本，文本来自 [12]：

$$Text_A = We\ expect\ demand\ to\ to\ increase.$$

$$Text_B = We\ expect\ worldwide\ demand\ to\ increase.$$

对于 $Text_A$ 与 $Text_B$，可以构造如下词典：

$$Dict = [we,\ expect,\ worldwide,\ demand,\ to,\ increase]$$

通过含有 6 个词的词典，可以对 $Text_A$ 与 $Text_B$ 分别构建一个 6 维向量：

- Bool-based vector for $Text_A$: $BV_A = [1,\ 1,\ 0,\ 1,\ 1,\ 1]$
- Count-based vector for $Text_A$: $BV_A = [1,\ 1,\ 0,\ 1,\ 2,\ 1]$

向量空间模型-权重选取

$TF-IDF$，即Term Frequency - Inverse Document Frequency，是一种比较常见且合理的权重选取方法，其优点在于结合了单个文本与全部语料库的的词语权重信息，具体如下：

$$TF-IDF = TF \times IDF$$

$$IDF = \log \frac{n}{df}$$

其中，TF 为词频，是某一词语在某一文本中出现的频次；IDF 为逆文档频率，是语料库中的文档总数 n 除以包含该词语的文档数 df 后的自然对数。为避免分母为 0 的情况，通常使用 $df+1$ 作为分母。

更多权重选取方法见我的公众号推文OneStata: 【NLP】从向量空间模型到词嵌入模型

向量空间模型-归一化处理

归一化处理（Normalization），即将数据放缩至区间 $[-1, 1]$。这部分在经管领域的实证分析中不常被提及，单在机器学习等领域较为重要。常见的归一化方法主要从采用L1和L2范数实现，具体如下：

$$V_{L1} = \frac{V}{||V||_1} = \frac{V}{\sum_1^{i=n} |v_i|}$$

$$V_{L1} = \frac{V}{||V||_2} = \frac{V}{\sqrt{\sum_1^{i=n} v_i^2}}$$

归一化方法的不同将会影响后续的相似性计算结果。但是，一些主流的会计实证期刊，并未提及归一化处理的相关方法 [12]。

向量空间模型-文本相似性

确定文本的特征向量之后，我们往往会通过相似性的计算来判断两个文本之间的相似性，最典型的相似性判别方法就是余弦相似性，如下：

$$CosineSimilarity = \frac{V_1 \cdot V_2}{||V_1|| \times ||V_2||}$$

在这部分的计算上，很多期刊的做法都有问题。根据我发表于*Journal of Cleaner Production*的文章 [10]，我认为可以取语料库向量均值作为对照组，并依次计算不同文档与均值之间的相似关系。

特别地，常见的相似性度量方法还有杰卡德相似度（Jaccard Similarity）、最小编辑距离（Minimum Edit Distance）以及欧氏距离（Euclidean Distance）等。

向量空间模型-相似度计算的代码实现

接下来以Python为例，展示计算文本相似度的方法。本部分代码改编自我的公众号推文：OneStata: 【NLP】文本相似性指标综述及 Python 实现

注意：没有用Stata并不是表示Stata不行，而是用Stata相当地麻烦，会额外增加很多不必要的工作量。

神经网络模型

向量空间模型一经提出便得到了广泛地应用，即便在最近，中英文的会计、公司金融领域的顶刊都不缺乏对它的使用 [4], [12], [20]。但是，经由特定词典（或全文本）生成的向量空间是维度受限的，并且在维度提升时会出现维度灾难（Curse of Dimensionality），同时向量中数据的稀疏性（Sparse）也会大大降低高维向量的表现。

为了解决这个问题，文本分析发展出了两条路径：一种是借助传统的机器学习方法使用主成分降维（Principal Component Analysis）、奇异值分解（Singular Value Decomposition）等方法对原有的高维数据进行降维处理；另一种则是基于神经网络应用而发展起来的词向量（Word Embedding）。

神经网络模型-Word2Vec 介绍

Word2Vec is a widely used algorithm based on neural networks, commonly referred to as "deep learning" (though word2vec itself is rather shallow)

神经网络模型-Word2Vec 含义

通过神经网络的转换，词向量模型（CBOW与Skip-gram）能够将赋予每一个重要词语在相同维度上的不同向量，从而解决了维度灾难以及零膨胀的问题。此外，采用神经网络方法生成的词向量，仍保留了词语之间的类比关系，比如已知某语料库中代表男人的词向量是 $[1,\ 0,\ 0]$、代表女人的词向量是 $[0,\ 1,\ 0]$，代表国王的词向量是 $[2,\ 1,\ 0]$，则可以根据条件“国王-男人 $=$ 女王-女人”，计算出女王的词向量约为 $[1,\ 2,\ 0]$。

神经网络模型-Word2Vec 应用

最后，词向量模型一般不会单独使用，而是会结合基于词典的文本分析方法共同使用。例如：刘瑶瑶和路军伟（2022）[15] 利用 Word2Vec 扩充了胡楠等（2021）的有关管理者短视的词集 [24]，并衍生出了有关前瞻性信息披露的词集。王新光（2022）[25] 扩充了胡楠等（2021）的有关管理者短视的词集，并验证管理者短视对数字化转型的影响。

神经网络模型-Word2Vec 注意事项

以胡楠等（2021）[24] 为例，对Word2Vec的使用注意事项进行一些说明。

- 如使用，则尽量给出自己的参数，并说明使用的是何种方法。比如，是CBOW还是Skip-gram，最好也说明其他参数。
- 认识到Word2Vec的不完美性，避免明显的错误，见下页。

神经网络模型-Word2Vec 注意事项

《管理者短视主义影响企业长期投资吗？——基于文本分析和机器学习》附录

1.Word2Vec相似词结果示例

附表1 Word2Vec相似词结果示例(取相似度排序位于前10)

中心词	相似词	相似度(基于年报语境)	词频
尽快	尽早	0.861	3744
尽快	早日	0.771	6732
尽快	抓紧	0.532	4934
尽快	及早	0.517	854
尽快	力争	0.469	34473
尽快	全力	0.467	17251
尽快	尽力	0.457	3309
尽快	立即	0.451	10129
尽快	争取	0.438	34450
尽快	加紧	0.438	3212

神经网络模型-Word2Vec 代码实现

接下来以Python为例，展示计算文本相似度的方法。本部分代码改编自我的公众号推文：【NLP】Word2Vec 文本相似词扩充

注意：没有用Stata并不是表示Stata不行，而是用Stata相当地麻烦，会额外增加很多不必要的工作量。

本节内容

背景介绍

常见方法

参考文献

文献详情 I

[1] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining," *Journal for Language Technology and Computational Linguistics*, vol. 20, no. 1, pp. 19–62, 2005.

[2] A. M. Turing, *Computing machinery and intelligence*. Springer, 2009.

[3] A. M. Turing, "Computing machinery and intelligence (1950)," *The Essential Turing: the Ideas That Gave Birth to the Computer Age*, pp. 433–464, 2012.

[4] 吴非, 胡慧芷, 林慧妍, and 任晓怡, "企业数字化转型与资本市场表现——来自股票流动性的经验证据," 管理世界, vol. 37, no. 7, pp. 130–144, 2021.

[5] 谢德仁 and 林乐, "管理层语调能预示公司未来业绩吗?——基于我国上市公司年度业绩说明会的文本分析," 会计研究, no. 2, pp. 20–27, 2015.

文献详情 II

[6] S. M. Price, J. S. Doran, D. R. Peterson, and B. A. Bliss, "Earnings conference calls and stock returns: The incremental informativeness of textual tone," *Journal of Banking & Finance*, vol. 36, no. 4, pp. 992–1011, 2012.

[7] 曾庆生, 周波, 张程, and 陈信元, "年报语调与内部人交易: "表里如一"还是"口是心非"?" 管理世界, vol. 34, no. 9, pp. 143–160, 2018.

[8] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of finance*, vol. 66, no. 1, pp. 35–65, 2011.

[9] 李哲 and 王文翰, "'多言寡行'的环境责任表现能否影响银行信贷获取——基于"言"和"行"双维度的文本分析," 金融研究, vol. 498, no. 12, pp. 116–132, 2021.

文献详情 III

[10] S. Zor, "Conservation or revolution? the sustainable transition of textile and apparel firms under the environmental regulation: Evidence from china," *Journal of Cleaner Production*, vol. 382, p. 135 339, 2023.

[11] 张勇 and 殷健, "会计师事务所联结与企业会计政策相似性——基于 tf-idf 的文本相似度分析," 审计研究, no. 1, pp. 94–105, 2022.

[12] L. Cohen, C. Malloy, and Q. Nguyen, "Lazy prices," *The Journal of Finance*, vol. 75, no. 3, pp. 1371–1415, 2020.

[13] S. B. Bonsall IV, A. J. Leone, B. P. Miller, and K. Rennekamp, "A plain english measure of financial reporting readability," *Journal of Accounting and Economics*, vol. 63, no. 2-3, pp. 329–357, 2017.

[14] 李春涛, 张计宝, and 张璇, "年报可读性与企业创新," 经济管理, vol. 10, pp. 156–173, 2020.

文献详情 IV

[15] 刘瑶瑶 and 路军伟, "前瞻性信息披露与分析师盈余预测——基于文本分析和机器学习的证据," 外国经济与管理, pp. 1–15, 2023.

[16] L.-W. Ku and H.-H. Chen, "Mining opinions from the web: Beyond relevance retrieval," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 12, pp. 1838–1850, 2007.

[17] E. Henry, "Are investors influenced by how earnings press releases are written?" *The Journal of Business Communication (1973)*, vol. 45, no. 4, pp. 363–407, 2008.

[18] 付文博 and 曾皓, "非处罚性监管能约束管理层语调操纵吗——基于年报文本的经验证据," 当代财经, no. 3, pp. 89–101, 2022.

文献详情 V

[19] 张小慧, 孙晓玲, 张璇, and 李万峰, "管理层语调会影响股价暴跌风险吗——基于业绩说明会的文本分析," 产经评论, vol. 13, no. 4, pp. 113–129, 2022.

[20] 王帆 and 邹梦琪, "关键审计事项披露与企业投资效率——基于文本分析的经验证据," 审计研究, no. 3, pp. 69–79, 2022.

[21] 沈菊琴, 李淑琴, and 孙付华, "年报语调与企业财务绩效: 心口如一还是心口不一," 审计与经济研究, vol. 37, no. 1, pp. 69–80, 2022.

[22] 梁日新 and 李英, "年报语调与审计费用——来自我国 a 股上市公司的经验数据," 审计研究, no. 5, pp. 109–119, 2021.

[23] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

文献详情 VI

[24] 胡楠, 薛付婧, and 王昊楠, "管理者短视主义影响企业长期投资吗?——基于文本分析和机器学习," 管理世界, vol. 37, no. 5, pp. 139-156+11+19–21, 2021.

[25] 王新光, "管理者短视行为阻碍了企业数字化转型吗——基于文本分析和机器学习的经验证据," 现代经济探讨, vol. 6, pp. 103–113, 2022.