Stata 文本分析：可能性与局限性

第八届 Stata 中国用户大会 · 青年学者论坛

左祥太 (Shutter Zor)

厦门大学会计系
Accounting Department
Xiamen Univeristy

Aug. 19-20, 2024

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 包: oneclick, onetext, econsig, 和 wordcloud
- 文章发表于:
  1. *Journal of Cleaner Production* [Sole author $\times 2$]
  2. *Technology Analysis & Strategic Management* [XMU T2]
  3. 《科研管理》[自科 A、CSSCI]
  4. 其他 SCI、ESCI、EI，学报等
- 审稿: *Journal of Business Ethics*、*Nature Communications*、*Journal of Cleaner Production*、*Nature Communications*、*BMC Public Health*、*Corporate Social Responsibility and Environmental Management*、*Humanities and Social Sciences Communications*、*Technological Forecasting & Social Change*、*Polish Journal of Environmental Studies* 等

目录

本节内容

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]、文本相似性[10-11]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[12-14]。

具体而言

- Zor S. Conservation or revolution? The sustainable transition of textile and apparel firms under the environmental regulation: Evidence from China[J]. Journal of Cleaner Production, 2023, 382: 135339.
  1. 利用词频统计测算地方政府的环境规制强度
  2. 利用文本相似度测算模仿式创新的强度（专利 IPC 分类号）
- Zor S. A neural network-based measurement of corporate environmental attention and its impact on green open innovation: Evidence from heavily polluting listed companies in China[J]. Journal of Cleaner Production, 2023: 139815.
  1. 利用 Word2Vec 神经网络扩充环境词集
  2. 利用词频统计测算上市公司的环境注意力强度（管理层讨论与分析文本）

本节内容

全文词频统计

全文词频统计是一种文本分析技术，它通过计算文本中每个词出现的次数来分析文本内容。这种技术可以帮助我们了解文本中的关键词分布，主题倾向，以及语言使用的特点。

本部分代码来自我的工作论文：

Zuo X. Comparing with Python: Text Analysis in Stata[J]. arXiv preprint arXiv:2307.10480, 2023.

全文词频统计

优点

- 直观性：词频统计提供了一种简单直观的方式来了解文本中哪些词汇使用得最频繁。
- 快速分析：可以快速对文本进行初步分析，不需要深入理解文本内容。
- 辅助理解：帮助读者或研究人员快速把握文本的主要内容和主题。
- 关键词提取：通过识别高频词汇，可以辅助提取文本的关键词或主题词。
- 文本比较：可以用来比较不同文本之间的词汇使用差异。
- 趋势分析：在大量文本数据中，可以分析词汇使用的趋势和变化。

全文词频统计

缺点

- 忽略语境：词频统计不考虑词汇在文本中的语境和语义，可能导致误解。
- 忽视语法：不区分词汇的语法角色，如名词、动词等。
- 无法理解复杂概念：不能理解由多个词汇组成的复杂概念或短语。
- 重复统计问题：对于形态变化丰富的语言，不同形态的同一词汇可能被统计为不同的词。
- 忽略重要性：所有词汇的权重相同，无法区分哪些词汇对文本的理解更为关键。
- 可能的误导：高频词汇可能是常用词或停用词，不一定代表文本的核心内容。

全文词频统计

Stata 工作流

- 读入文本文件：*import delimited* or *fileread()*
- 去除停用词：
  - *subinstr()* or *ustrregexra()*？
  - 思考：如果 he 出现在 the 前面会怎样？
- 去除标点符号：
  - *subinstr()* or *ustrregexra()*
  - 需要注意，现在还没有方法去除‘ 符号，即 Stata 的 Macro Marks 引用符号。
- 以空格为界分词：
  - 对于英文文本：无需过多处理，具有天然的分词优势
  - 对于中文文本：需要额外处理，*ustrwordcount()*、*ustrword()*
- 统计全文各单词的词频
- （可选项）绘制词云图：*ssc install wordcloud*

全文词频统计

示例一：Harry Potter 全文统计，并绘制词云

```
. sum max_value, d

                        max_value
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            1              1
10%            1              1       Obs                22,011
25%            1              1       Sum of Wgt.        22,011

50%            3                      Mean              17.83172
                      Largest         Std. Dev.         138.4349
75%            9           2387
90%           31           2911       Variance          19164.22
95%           67           5019       Skewness          90.01196
99%          249          17134       Kurtosis          10720.94
```

全文词频统计

示例一：Harry Potter 全文统计，并绘制词云

全文词频统计

示例一：Harry Potter 全文统计，并绘制词云（提取人名）

dobby	482
ron	501
moody	507
dudley	526
fudge	546
vernon	563
mcgonagall	737
ginny	775
george	776
neville	777
lupin	808
sirius	1159
voldemort	1302
malfoy	1372
weasley	1728
snape	1858
hagrid	2048
dumbledore	3374
hermione	5381
harry	18620

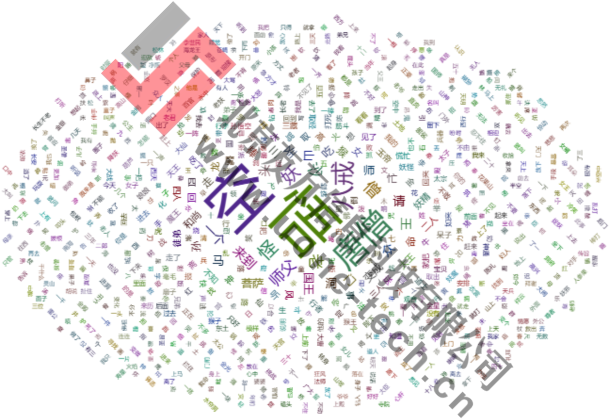
全文词频统计

示例二：西游记全文统计，并绘制词云

```
. sum Total, d

                            Total
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1            1
 5%            1            1
10%            1            1       Obs                9,049
25%            1            1       Sum of Wgt.        9,049

50%            2                    Mean               6.925185
                      Largest       Std. Dev.          36.29849
75%            5          664
90%           13          959       Variance           1317.58
95%           23         1930       Skewness           38.14896
99%           74         1978       Kurtosis           1907.014
```

全文词频统计

示例二：西游记全文统计，并绘制词云



特定词语词频分析

特定词语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。

特定词语词频分析

Stata 工作流

- 将原始文本，通常为多个 txt 文件，导入 Stata 并形成 dta 文件
  - 通常使用 *local*、*dir*、*tempfile* 进行文件夹遍历、合并等操作，参考：「Stata」遍历文件夹与批量追加合并
- 使用 *preserve* 与 *restore* 框架读取词袋文件，并以 *levelsof* 的形式将特定词语记录在 *local* 中
- 使用 *onetext* 配合循环统计词频，并生成新变量
- 是否需要分词？我认为不需要，理由如下：
  - 特定词语专业性一般都比较强，使用 Stata 分词可能导致词语被强制分开，从而造成统计误判
- 接下来以Stata为例，展示在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:「Stata」词频统计下的数字化转型

特定词语词频分析

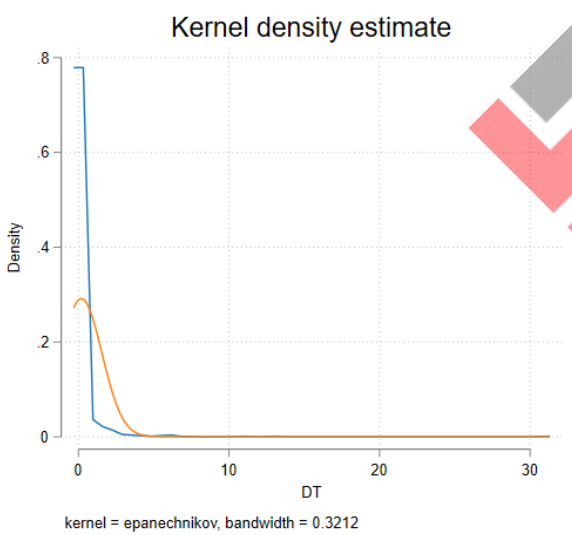
示例一：数字化转型 - 分析代码

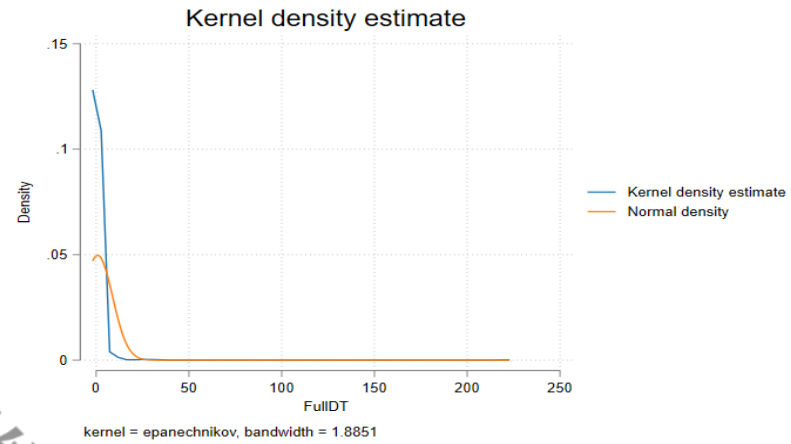
```
*- 读入文本
local txtFiles : dir "resources/files" files "*.txt"

local N = 1
foreach singleFile in `txtFiles' {

    import delimited "resources/files/`singleFile'"
            , delimiter("shutterzor", asstring)
            varnames(nonames) encoding(UTF-8) clear

    gen stkcd = ustrregexs(0) if ustrregexm("`singleFile'", "\d+")
    gen year = ustrregexs(0) if ustrregexm("`singleFile'", "_\d+-")
    replace year = substr(year, 2, 4)

    tempfile file`N'
    save "`file`N''"

    dis as result "file `N' has been finished"
    local N = `N' + 1
}

use "`file1'", clear
local file_total_num = `N' - 1
forvalues fileNum = 2/`file_total_num' {
    append using "`file`fileNum''"
}
rename v1 content
save "MDAText.dta", replace
```

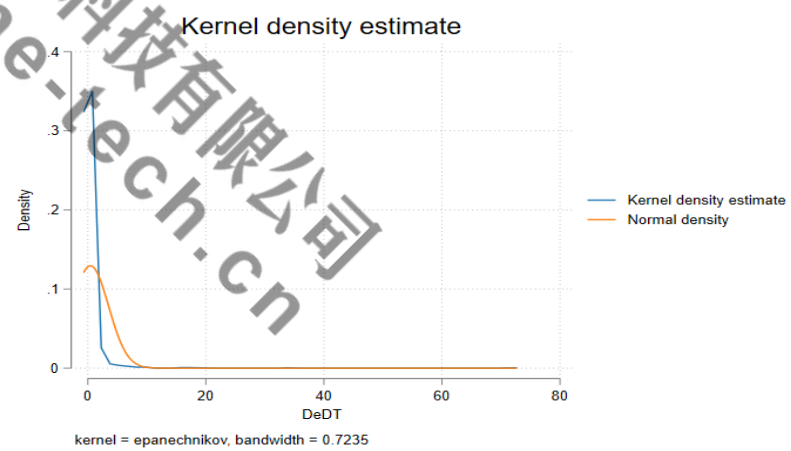
```
*- 统计特定词语词频
use MDAText.dta, clear

    *- 积极词汇
    preserve
        import delimited "resources/DigitalWords.txt", encoding(UTF-8) clear
        levelsof v1, local(DigitalWords)
        local totalNum = _N
    restore

    local tempCount = 1
    foreach DigitalWord of local DigitalWords{
        quietly onetext content, k("`DigitalWord'") m(count) g(dw`tempCount')
        local tempCount = `tempCount' + 1
        dis as result %4.2f (`tempCount'-1)/`totalNum'*100
    }

        *- 计算积极情绪，并删除无用变量
        egen DT = rowtotal(dw*)
        keep stkcd year content DT

    *- 保存结果
    save "result.dta", replace
```

特定词语词频分析

示例一：数字化转型 - 与 CNRDS 比对的统计结果







左边是按照上述代码从管理层讨论与分析文本中提取出来的数字化转型词语的总数的核密度图；右上是 CNRDS 的年报全文数字化转型词语总数的核密度图；右下是 CNRDS 的年报去表后全文数字化转型词语总数的核密度图。

特定词语词频分析

示例二：管理层语调/情绪分析

将上述代码中，用于计算数字化转型的词典，依次替换为积极、消极情绪词典后，就能分别计算管理层的积极、消极情绪/语调，随后按照参考文献的公式进行情绪计算即可。常见的情绪词典包括 Loughran-McDonald 词典[8]、中文情感极性词典[15]、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[16]。

一般而言，如下公式都可以用于计算管理层语调/情绪：

1. $Sen = \frac{Pos-Neg}{Pos+Neg}$，该方法应用于[5-7,16-17]
2. $Sen = \frac{Pos(Neg)}{Tot}$，该方法应用于[18-19]
3. $Sen = \frac{Pos-Neg}{Tot}$，该方法应用于[20-21]

进阶文本分析

主题聚类-原理

全文的词频统计在一定程度上揭示了文本的重点走向，但却难以反映文本的主题。被统计文本所展现的具体主题，仍需要人为地、机械地根据已有词语的出现频次与概率进行总结。

主题聚类则能够很好地解决这一问题，从而更方便地从词语中提炼出潜在的关系与主题。

使用 Stata 做 LDA 可以参考如下文章：

Schwarz C. ldagibbs: A command for topic modeling in Stata using latent Dirichlet allocation[J]. The Stata Journal, 2018, 18(1): 101-117.

进阶文本分析

主题聚类-示例

举个  :

- 在某段文本中，词语与词频的分布构成了如下集合：$\{A:100,\ B:50,\ C:10\}$
- 绘制词云图，可以发现 A 跟 B 可能更加紧密
- 殊不知，C 出现的每一次，周围三个词以内都有 A，而 B 出现的 50 次中只有 15 次有 A

从词语共现的概率来看，A 跟 B 同属一个主题的概率会小于 A 跟 C 同属一个主题的概率。当然，实际的主题聚类需要用到的 LDA 方法会比上述例子复杂更多。

一个误区： LDA 并不能告诉我们某段文本中包含了哪些具体的主题，而是会依据“主题 123”的形式，将不同的词语归纳至该主题下。使用 LDA 后，仍需要人为根据该算法生成的词语合集判定该主题的类别。

进阶文本分析

向量空间模型-原理

向量空间模型（Vector Space Model）由 Salton 等（1975）提出[22]，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。

进阶文本分析

向量空间模型-示例

假设我们有以下文本，文本来自 Cohen et al.(2020)[11]：

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

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

对于 $Text_A$ 与 $Text_B$，可以构造如下词典：

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

进一步地，可以对 $Text_A$ 与 $Text_B$ 分别构建一个 6 维向量：

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

进阶文本分析

向量空间模型-示例

从而，我们可以根据两个向量计算文本之间的余弦相似度：

$$BV_A = [1,\ 1,\ 0,\ 1,\ 1,\ 1]$$

$$BV_B = [1,\ 1,\ 1,\ 1,\ 1,\ 1]$$

$$\begin{aligned} CS_{BV_A, BV_B} &= \frac{BV_A \cdot BV_B}{||BV_A|| \times ||BV_B||} \\ &= \frac{1 \times 1 + 1 \times 1 + 0 \times 1 + 1 \times 1 + 1 \times 1 + 1 \times 1}{\sqrt{1^2 + 1^2 + 0^2 + 1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2}} \\ &= \frac{5}{\sqrt{5 \times 6}} \approx 0.913 \end{aligned}$$

此外，可用于计算相似度的公式还有：Jaccard 相似度、Hamming 距离、Manhattan 距离、Mahalanobis 距离、Chebyshev 距离，以及Euclidean 距离等。

进阶文本分析

向量空间模型-缺点

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，可以考虑使用主成分降维（Principal Component Analysis）或奇异值分解（Singular Value Decomposition）等方法对原有的高维数据进行降维处理。

注意： Stata 暂时没有可供实现向量空间模型的现成命令。

进阶文本分析

神经网络模型

这里提到的神经网络模型主要是指 Word2Vec 一类的词嵌入模型（Word Embedding）。该方法可以生成词与词之间的类比关系，并且通常被用于拓展原始词典中的词汇[23]。

注意： Stata 暂时没有可供实现Word2Vec的现成命令。

结语

总之，要想在 Stata 中实现完备的文本分析功能，还需要广大开发者的共同努力，期待在不久的将来能开发出相关代码包。

本节内容

文献详情 I

[1] HOTHO A, NÜRNBERGER A, PAASS G. A brief survey of text mining[J]. Journal for Language Technology and Computational Linguistics, 2005, 20(1): 19-62.

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

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

[4] 吴非, 胡慧芷, 林慧妍, 等. 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[J]. 管理世界, 2021, 37(7): 130-144.

[5] 谢德仁, 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[J]. 会计研究, 2015(2): 20-27.

文献详情 II

[6] PRICE S M, DORAN J S, PETERSON D R, et al. Earnings conference calls and stock returns: The incremental informativeness of textual tone[J]. Journal of Banking & Finance, 2012, 36(4): 992-1011.

[7] 曾庆生, 周波, 张程, 等. 年报语调与内部人交易: "表里如一" 还是 "口是心非"?[J]. 管理世界, 2018, 34(9): 143-160.

[8] LOUGHRAN T, MCDONALD B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks[J]. The Journal of finance, 2011, 66(1): 35-65.

[9] 李哲, 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[J]. 金融研究, 2021, 498(12): 116-132.

[10] 张勇, 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[J]. 审计研究, 2022(1): 94-105.

文献详情 III

[11] COHEN L, MALLOY C, NGUYEN Q. Lazy prices[J]. The Journal of Finance, 2020, 75(3): 1371-1415.

[12] BONSALL IV S B, LEONE A J, MILLER B P, et al. A plain English measure of financial reporting readability[J]. Journal of Accounting and Economics, 2017, 63(2-3): 329-357.

[13] 李春涛, 张计宝, 张璇. 年报可读性与企业创新[J]. 经济管理, 2020, 10: 156-173.

[14] 刘瑶瑶, 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[J]. 外国经济与管理, 2023: 1-15.

[15] KU L W, CHEN H H. Mining opinions from the Web: Beyond relevance retrieval[J]. Journal of the American Society for Information Science and Technology, 2007, 58(12): 1838-1850.

文献详情 IV

[16] HENRY E. Are investors influenced by how earnings press releases are written?[J]. The Journal of Business Communication (1973), 2008, 45(4): 363-407.

[17] 付文博, 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[J]. 当代财经, 2022(3): 89-101.

[18] 张小慧, 孙晓玲, 张璇, 等. 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[J]. 产经评论, 2022, 13(4): 113-129.

[19] 王帆, 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[J]. 审计研究, 2022(3): 69-79.

[20] 沈菊琴, 李淑琴, 孙付华. 年报语调与企业财务绩效: 心口如一还是心口不一[J]. 审计与经济研究, 2022, 37(1): 69-80.

[21] 梁日新, 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[J]. 审计研究, 2021(5): 109-119.

文献详情 V

[22] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[23] 胡楠, 薛付婧, 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[J]. 管理世界, 2021, 37(5): 139-156+11+19-21.