文本分析模块算法

目录

SPSSAU 中，文本分析模块支持对文本数据进行各类分析，包括切词（词定位）、绘制词云图、按词或者按行进行文本情感分析、按词或者按行进行文本聚类分析、社会网络关系图、LDA 主题分析、新词发现和自定义词库（包括新词词库、停用词和情感词），在 SPSSAU 中支持：

- ✓ 直接粘贴文本数据/txt 和 EXCEL 数据上传；
- ✓ 最高 10 个文本分析数据项目；
- ✓ 多样全面化的文本分析（包括情感、聚类、社会网络图和 LDA 主题分析等）；
- ✓ 批量自定义词库（包括新词词库、停用词和情感词）；
- ✓ 下载各类分析结果等。

spssau.com

SPSSAU 数据格式

进入 SPSSAU 文本分析模块之后，首先需要上传文本数据。上传数据涉及以下内容：支持直接粘贴文本进行上传数据，支持上传 txt 或 excel 格式数据，上传文件最高限制为 5M。

	A	B	C
1	近日，住房城乡建设部下发通知，要求各地深入贯彻落实习近平		
2	2023年，全国计划新开工改造城镇老旧小区5.3万个、涉及居民	A列放文本数据	
3	为贯彻落实《中共中央 国务院关于加强新时代老龄工作的意	不需要有标题	
4	为帮助各地更好学习并贯彻落实《住房城乡建设部关于全面推进	每行是1个分析文本	
5	为促进数字经济和实体经济融合，通过数字化赋能推动生活性服		
6	2023年12月18日，甘肃临夏州积石山6.2级地震发生后，住房城		
7	近日，住房城乡建设部安全生产委员会办公室约谈山西省住房城		
8	日前举行的中央经济工作会议强调，加快推进保障性住房建设、		
9	日前召开的全国住房城乡建设工作会议全面总结今年工作，深刻		
10	近日，住房城乡建设部党组书记、部长、安委会主任倪虹主持召		
11	2024年是中华人民共和国成立75周年，是实施“十四五”规划的		
12	12月21日~22日，全国住房城乡建设工作会议在京召开，住房城		
13	12月21日至22日，全国住房城乡建设工作会议在北京召开。会议		
14	近日，住房城乡建设部就六项国家标准《预制轻薄型热水辐射供		

spssau.com

提示：如果是通过 excel 格式（包括 csv/xls/xlsx 格式）时，只需要 1 列数据，该列数据中包括文本信息。将文本全部放置于 A 列中，A 列不需要有标题信息。每行（即每个单元格）存在 1 个分析文本。如果是 txt 文档或者粘贴文本进行上传，那么系统会自动过滤掉空行数据，并且以回车键作为每行（即每个分析文本）标志。

计算公式

1. 基本说明

SPSSAU 通过整合多个强大的 Python 包，如 jieba、HanLP、SnowNLP 和 gensim，为用户提供了全面且灵活的文本分析工具。这些工具不仅支持基本的分词和情感分析，还能进行复杂的聚类、社会网络分析和主题建模，使得用户能够深入挖掘文本数据中的信息，为决策提供有力支持。

2. 词云分析等

词云分析借助于 Jieba 包进行分词处理，该包已经内置很多停用词如“的”、“了”等，当然研究者可自行定义停用词或新词，以实现自定义分析。与此同时，SPSSAU 借助 antv 包实现可视化展示绘图。

提示：SPSSAU 进行文本分析时出现‘一键处理’是什么意思?

在进行文本分析时，如果切词过多（通常>50 万个切词时），其会导致系统内部出现存储问题。出现此情况时，SPSSAU 会建议进行‘一键处理’，即删除掉部分行（通常在 30%左右），即删除一部分数据后再次自动分析，一键处理即系统会自动计算大约删除多少数据并且重新进行分析。‘一键处理’后的数据行数会明显少于原始数据，建议研究者可自己下载 SPSSAU 系统分析的真实分析数据。

3. 文本情感分析

在文本情感分析方面，SPSSAU 提供了按词和按行两种方式进行情感评估的功能：

✓ 按词情感分析

该方法基于一个庞大的情感词库，该库由 SPSSAU 团队收集和整理，包含约 13 万条情感词（包括 BosonNLP、台湾大学、清华大学、知网等情感词库整理得到）。用户也可以通过“我的词库”功能添加自定义情感词，以增强情感分析的准确性。SPSSAU 将情感词全部压缩在 -1~1 之间，并且在输出结果时设置情感方向如下表格：

情感分值区间	情感方向
[-1, -1/3)	负向
[-1/3, 0)	偏负向
[0, 1/3)	偏正向
[1/3, 1]	正向

spssau.com

没有分值时	情感词典中无该词

如果研究者自定义情感词，设置分数时建议介于-1~1 之间，如果并非这样，SPSSAU 会事先对情感分进行压缩处理，便于分值具有可对比性。情感分压缩处理公式如下：

$$a + (b - a) \times \frac{(x - x_{\text{Min}})}{x_{\text{Max}} - x_{\text{Min}}}$$

其中：

a为-1，b为 1

x为情感分，x_{Max}表示情感分最大值，x_{Min}表示情感分最小值

✓ 按行情感分析

按行情感分析时，SPSSAU 首先利用情感词典计算出该行文本中切词的情感得分（每个切词的情感分值均介于-1~1 之间），并且进行求和得到情感分值 x，接着对分值进行反正弦处理，以便对情感分进行压缩 在-1~1 之间。计算公式如下：

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

其中：

x 为对文本各切词的情感得分求和值

4. 文本聚类分析

SPSSAU 支持文本聚类分析，用户可以选择按词或按行进行聚类：

✓ 按词聚类

该方法将选中词（默认为词频 top20 词，研究者可自定义选择）转换为 word2vec 向量表示，然后使用 K-means 算法对这些向量进行聚类。通过这种方式，可以将具有相似语义的单词归为同一类，从而帮助理解文本中的主题。与此同时，SPSSAU 计算选中词的共词矩阵并且对其处理，进而进行 MDS 多维尺度变换，得到按词聚类可视化图形。关于共词矩阵的处理，如下述：

$$A = \begin{bmatrix} f_1 & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & f_2 & c_{23} & \cdots & c_{2n} \\ c_{31} & c_{32} & f_3 & \cdots & c_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & f_n \end{bmatrix}$$

其中：

A：共词矩阵

f_i：第i个词的词频

c_{ij}：词汇w_i和词汇w_j之间的‘共现次数’，其基于行计算，其指两词在‘同一行’时出现的次数。SPSSAU 得到共词矩阵后，为绘制可视化图形，进行如下处理：

spssau.com

$$A' = \begin{cases} 1, \text{if } i = j \\ \frac{1}{N_{ij}} \times S, \text{if } i \neq j, N_{ij} \neq 0 \\ 0.5 \times S, \text{if } i \neq j, N_{ij} = 0 \end{cases}$$

其中，A'是处理后的矩阵，其数字越大表示两词之间的共现越少即距离越远。

N_{ij}是共现词数，即词汇w_i和词汇w_j的共现次数

$S = \sum_{k=1}^{n} f_k$是主对角线数据之和，即所有词频的总和

✓ 按行聚类

此方法时，SPSSAU 首先获取得到每行文本的切词信息，并且每个切词均有其 TF-IDF 值，比如有 1 万行文本，共切词 10 万个，那么就形成 1 万*10 万的数据矩阵；

接着针对‘TF-IDF’值进行 Kmeans 聚类，默认为 3 个类别，当然用户可切换聚类个数；得到聚类别信息即为每行的聚类类别。

5. 社会网络分析

社会网络分析是共词矩阵的可视化（默认是针对词频 top20 词，研究者可自定义选择），关于‘共词矩阵’，说明如下：

$$\begin{bmatrix} f_1 & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & f_2 & c_{23} & \cdots & c_{2n} \\ c_{31} & c_{32} & f_3 & \cdots & c_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & f_n \end{bmatrix}$$

其中：

A：共词矩阵

f_i：第i个词的词频

c_{ij}：词汇w_i和词汇w_j之间的‘共现次数’，其基于行计算，其指两词在‘同一行’时出现的次数。

社会网络图绘制时，词频体现节点的大小，节点之间的关系为‘共现词数’的体现。

6. LDA 主题分析

SPSSAU 利用 gensim 包进行 LDA（Latent Dirichlet Allocation）主题分析，提供了一系列功能以帮助用户深入理解文本数据中的主题结构。以下是该分析模块的详细描述和功能优化：

主题与词权重及视化图

SPSSAU 生成一个表格展示各主题与关键词的权重值。其包括词、词频及对应主题的权重信息等。通过可视化工具展示各主题与关键词的权重值。研究者可以点击特定主题的气泡，以仅展示该主题下某词的权重，并按权重从大到小排序。这种

spssau.com

交互式图形能够帮助用户直观理解不同主题之间的关系以及每个词词在特定主题中的重要性。

除此之外，可视化图还展示各主题与词的权重值，采用气泡图形式呈现。气泡的大小与权重成正比，气泡越大表示该词在对应主题中的权重越高。这种可视化方式使得用户能够快速识别出最重要的词及其在不同主题中的分布情况。

- 主题分布表格

SPSSAU 还提供一个 主题分布表格，展示各行文本隶属的主题编号。此表格将有助于用户了解哪些文本内容属于哪些特定主题，从而为后续分析提供依据。

以及用户可以根据实际分析需要确认并修改每个主题的名称。修改后，系统将重新展示更新后的表格和图形信息，以确保所有输出结果都能准确反映用户对主题的理解和命名。这一功能增强了模型结果的人性化和可解释性，使得分析结果更符合实际应用场景。

- 一致性系数

SPSSAU 输出的结果包括 UMass 一致性系数，该系数用于评估主题模型的质量。UMass 一致性系数越高，表明主题的内部一致性越强，从而反映出该主题在文本中的重要性和相关性。

- 总结

通过以上功能，SPSSAU 为用户提供了一个全面且互动性强的 LDA 主题分析工具，这些功能不仅帮助用户深入挖掘文本数据中的潜在信息，还提升研究成果的可读性和应用价值。

7. 新词发现

新词发现算法针对中文文本分析才有意义，SPSSAU 借助于 Hanlp 包完成新词发现。新词发现算法基于信息熵和互信息，具体可参考文章《基于信息熵和互信息的新词提取实现》或《互联网时代的社会语言学：基于 SNS 的文本数据挖掘》。

- 信息熵：信息熵越高，表示词汇与其他词组合成新词的可能性和稳定性更强。
- 互信息：互信息值越高，表明词汇之间的关联性更强，增加新词出现的可能性。

通常情况下，新词更可能在信息熵较高（约 0.5）且互信息值较高（50~200）时出现。研究者可以根据需要自行调整这两个指标的标准，并重新进行分析。研究者如果某些词为‘新词’，可直接将其加入‘新词词库’或者批量加入‘新词词库’中。

8. 我的词库

SPSSAU 中提供自定义‘新词’、‘停用词’和‘情感词’，每个词库最多 5000 词。

提示：在词云分析和词定位分析等处，均有加入或者移出停用词的功能。但通常情况下建议一次性将停用词批量处理，在处理完成后，重新进行分析，重新进行

spssau.com

分析的方式为：点‘我的项目’-》点击‘重新分析 ICON’。‘新词’或‘情感词’的处理类似。

参考文献

【1】The SPSSAU project (2024). SPSSAU. (Version 24.0) [Online Application Software]. Retrieved from https://www.spssau.com.

【2】Scikit-learn: Machine Learning in Python (Version 1.4.2). Available at: https://github.com/scikit-learn/scikit-learn.

【3】Jieba: A Chinese Text Segmentation Library (Version 0.42.1). Available at: https://github.com/fxsjy/jieba.

【4】HanLP: A Natural Language Processing Toolkit for Chinese (Version 1.8.4). Available at: https://github.com/hankcs/HanLP.

【5】Gensim: Topic Modelling for Humans (Version 4.3.2). Available at: https://github.com/piskvorky/gensim.

【6】SnowNLP: A Simple Library for Processing Chinese Text (Version 0.12.3). Available at: https://github.com/isnowfy/snownlp.

【7】antv: A Visualization Framework for Data Visualization (Version 4.0.0). Available at: https://antv.vision.

spssau.com