第四章 初步业务活动——文本分析

新道
seentao

目录.

CONTENT

新道
seentao

第一节

文本分析基础

一、大数据技术下的文本分析方法

概述

- 数字经济浪潮，导致社会生产的方方面面受到影响，人类社会产生的数据总量在不断增加。非结构化数据占据了人类数据总量的大部分，而且比重不断上升。

- 作为非结构化数据的重要组成部分，文本数据的类型丰富多样（如社交网络类文本、上市公司披露类文本、媒体报道类文本），对财税审领域具有较高的信息价值，因而**文本分析（TextualAnalysis）技术**异军突起，形成一个新的分析技术。

- 文本分析技术主要有主题分析、词典法、词袋法、监督学习、无监督学习与自然语言处理等六大类。从主题分析到自然语言处理，文本分析技术的自动化程度逐渐提高，使用的算法逻辑也逐渐复杂。

一、大数据技术下的文本分析方法

（一）主题分析

1.主题分析法的概念

主题分析（Thematic analysis）是一种专家方法，需要有经验的人员基于自身经验和理解，对研究数据进行挖掘。主题分析一般与扎根理论方法相结合，基于专家自身经验和对世界的理解产生对数据的见解，从而构建新理论。

主题分析是一个反复迭代的过程，在分析开始前研究人员尚不知道文本所属类别，需要对文献和数据不断进行比较，通常从参与者自己的语言开始（一阶编码），将相似编码归为一类（二阶编码），进而开发出一系列源自文本的编码和类别。主题分析常见于社会学与管理学领域。NVivo、ATLAS.ti等计算机软件能够简化相关过程，但文本分类仍主要依赖人类编码，计算机自动化程度较低。

2.主题分析法的优点与局限性

- 主题分析的优点在于使用参与者自身的认知来挖掘数据，对少量文本的理解更深入。
- 主题分析的局限性在于其属于时间、劳动力密集型任务，不适合用于大规模样本，同时编码人员的经历和偏好不同，编码标准不统一。

（二）词典法

1.词典法的概念

词典法（dictionary analysis）是一种运用词典对特定文本的词语（或词组）的词频统计计数，将定性的文本数据压缩成定量的词组频数的文本分析技术。运用词典法的关键是具备成熟且适合所分析领域的词典，否则需要分析师根据分析的问题与文本数据，结合领域的相关专业知识构建适配的词典加以验证。

2.词典法的种类

国外已先后形成多部比较成熟的英文文本词典，如LM词典、哈佛大学通用调查词典、文辞乐观与悲观词典。

国内大多数分析师在参考英文词典及其他词库的基础上，针对中文文本构建自己的词典并展开分析，如台湾大学NTUSD简体中文情感词典、知网Hownet情感词典、清华大学李军中文褒贬义词典。

一旦拥有适配的词典，则可以采用计算机软件协助文本内容分析，计算机自动化程度较高，人工工作量较低。

3.词典法的优点与局限性

- 词典法通常被用于管理学领域，如计算分析文本的语调情感。与主题分析相比，词典法的优点在于能够对所研究的文本信息进行定量分析，既提高了文本处理的效率，也加深了读者对文本含义或性质的理解与把握，同时还增强模型的可复制性。
- 词典法的局限性在于针对特定文本构建词典时，需要相关领域的专业知识，因而可能导致构建的特定词典与其他文本适配度不高。另外，词典法会忽略文档的上下文关系。

一、大数据技术下的文本分析方法

（三）词袋法

1.词袋法的概念

词袋法（Bag-of-Words）建立在文字词组语序不重要的假设之上，是一种将文本看作若干词语的集合，只计算每个词语出现次数的文本向量化的表示方法，可将非结构化的定性文本数据转换成计算机能理解和直接使用的向量。

词袋法是计算科学领域对文本数据的简化和压缩方法，后续可以据此进行进一步监督学习和无监督学习，常被用于管理学领域，如计算分析文本相似度。词袋法同样可以采用计算机软件协助文本内容分析，计算机自动化程度较高，人工工作量较低。

2.词袋法的种类

词袋法主要包括独热表示法和词频 - 逆文档频率法。独热表示法将多个文档组织成一个文档特征矩阵，矩阵中每行代表一个文档，每列代表一个特征词，每个元素衡量每个文档中对应特征词的出现频数。词频 - 逆文档频率法的核心理念是如果某个词或短语在一篇文章中出现的频率高，但在其他文章中很少出现，则认为该词或短语具有很好的类别区分能力，适合用来分类。

3.词袋法的优点与局限性

- 词袋法的优点在于编码标准稳定统一，具有统计学特性与强扩展性。
- 但由于编码过程忽略了词语先后顺序，无法反映上下文含义，牺牲了文本的很多信息，变通性弱，文本分析深度不足；当文档中词语数量过多时，向量维度过高则可能产生维度“灾难”。

（四）监督学习

1.监督学习的概念

监督学习（Supervised Leaming）是指将输入数据x和对应标签y作为训练集，建立合适的模型来学习两者之间的关系，并使用确定的模型来预测未知样本。监督学习中预设的标签可以明确分类目标，而算法则可映射输入与输出之间的联系。

监督学习常被用于计算机科学、政治学与管理学领域，可采用Python中的scikit-leam、gensim、nltk或R中的topicmodels、stm等实现。

2.监督学习模型预测效果评价指标

监督学习算法包括支持向量机、线性回归、逻辑回归、朴素贝叶斯、线性判别分析、决策树、K-近邻等。其中，朴素贝叶斯和支持向量机技术是文本分析中常用的监督学习算法。衡量监督学习模型预测效果可采用的评价指标有准确率、查准率、查全率、F得分、AUC等。

3.监督学习的优点与局限性

- 监督学习的优点是允许研究者事先定义编码规则，逻辑简单，目标明确，同时可用于海量数据的研究，精确度较高。
- 其局限性在于，不仅需要高质量的标签变量，且训练的模型容易由于特征词太多导致过拟合。

（五）无监督学习

1.无监督学习的概念

无监督学习（Unsupervised Learning）是指根据类别未知（没有被标记）的训练样本解决模式识别中的各种问题。与监督学习相比，无监督学习不需要为数据打标签，缺乏具有明确目的的训练方式，无法提前预知结果，也很难量化预测效果。无监督学习常被用于计算机科学、政治学与管理学领域，可采用Python中的scikit-learn、gensim、nltk或R中的 topicmodels、stm等实现。

2.无监督学习的常用算法技术

聚类和降维技术是文本分析中常用的无监督学习算法。

（1）聚类。聚类的关键是计算相似度，目的在于将相似的东西聚在一起，具体又可分为划分和层次两种方法。

（2）降维。常用的文本降维方法是隐含狄利克雷分布主题模型。LDA是一种文档主题生成模型，包含词、主题和文档三层结构，可以用来识别大规模文档集或语料库中潜藏的主题信息。

3.无监督学习的优点与局限性

- 无监督学习能加速数据的标注与分类，降低人工工作量。
- 局限性为“标注”是机器按照数字特征进行的分组，需要研究者解读才可以赋予“标注”的意义，同时训练过程需要大量的调参。

（六）自然语言处理

1.自然语言处理的概念

自然语言处理（Natural Language Processing，NLP）是文本分析技术中自动化程度最高的类型。NLP以语言为对象，利用计算机技术模拟人类行为理解和处理语言，包括自然语言理解和自然语言生成，关注计算机和人类语言的相互作用。

NLP考虑单词或文字的先后顺序，可以标记句子中单词的词性（如名词、形容词），将文档从一种语言翻译成另一种语言，甚至能使用句子的上下文来阐明词语的词义。NLP是一个完全计算机自动化的过程，不需要人类的理解或解释，具有丰富的用途，如采用深度学习和多模式等尖端技术进行情感分析。

2.自然语言处理的技术实现

NLP在计算机科学、信息科学、语言学、心理学等多个领域被用作文本分析工具，可采用Python中的 nltk实现。主要有词嵌入、长短期记忆网络、深度自注意力网络、基于语义理解的深度双向与训练模型等算法。

3.自然语言处理的优点与局限性

- NLP的优点在于计算机自动化程度高，系统性强，能够分析语义，且执行速度快。
- 局限性在于，大多数模型是黑箱状态，虽然模型使用简单，但解读原理困难，此外模型训练较为耗时。

二、通过Python实现文本分析

➢ JIEBA库简介

JIEBA库是一款优秀的Python第三方中文分词库，是使用词典法对中文文本进行分析时的首选工具，支持三种分词模式。

01 精确模式

精确模式将句子最精确地切开，切分后的字符串长度与原文一致，适合进行词频类的文本分析。

02 搜索引擎模式

搜索引擎模式是在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

03 全模式

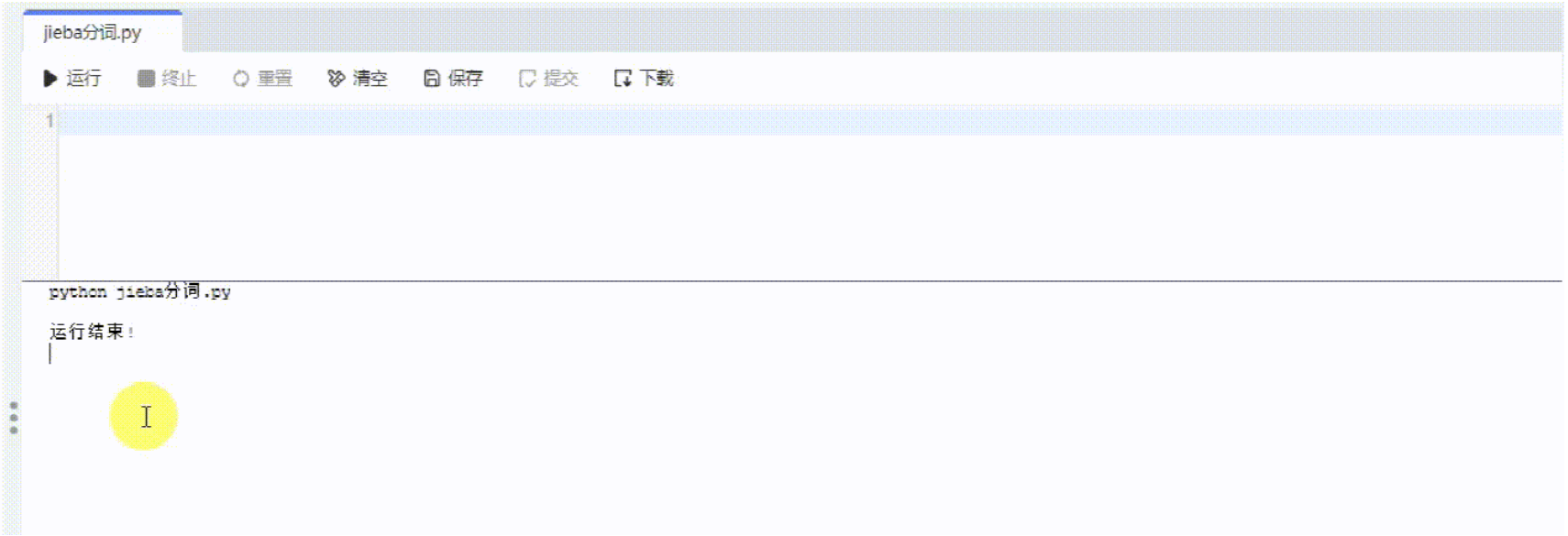
全模式是扫描句子中所有的可以成词的词语全部，但是不能解决歧义的问题，切分后的字符串长度大于原文长度。

二、通过Python实现文本分析

➢ 实训演示-对一句话进行分词

请对：“同学你好，请给我一份小饼干，谢谢你呀，你可真是个小可爱！”进行分词

```
import jieba
text = "同学你好，请给我一份小饼干，谢谢你呀，你可真是个小可爱!"
cutwords = jieba.lcut(text)     #用jieba分词
print(cutwords)
```



二、通过Python实现文本分析

新道 seentao

➢ 使用停用词库，对代码进行优化：

停用词库

我们日常使用的代词、介词、语气词、标点符号等脱离句子后信息含量较低，为了提高分词后的使用效果，将此类词组或符号从分词结果中剔除。

剔除过程

1.建立停用词库。
本章实战练习中提供了一份通用的停用词库，如有特殊需剔除的词组，可以在词库中添加。

2.使用“isin”函数的反运算，将停用词库中的词剔除。

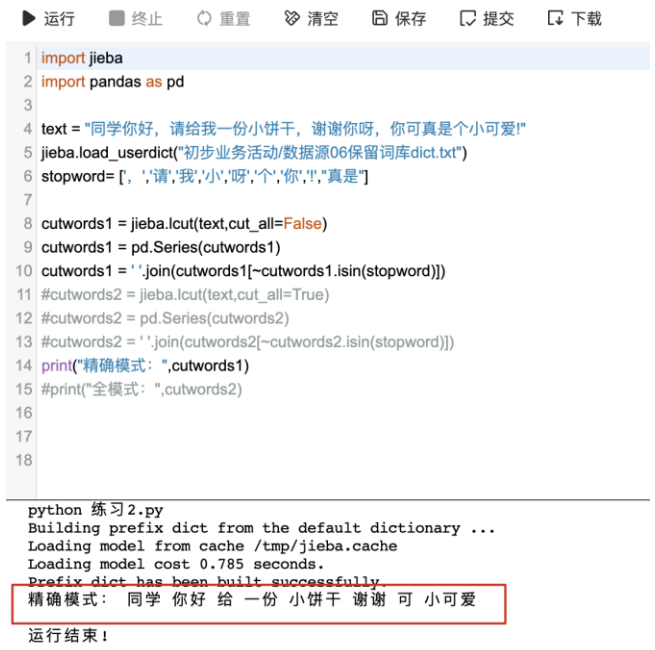
二、通过Python实现文本分析

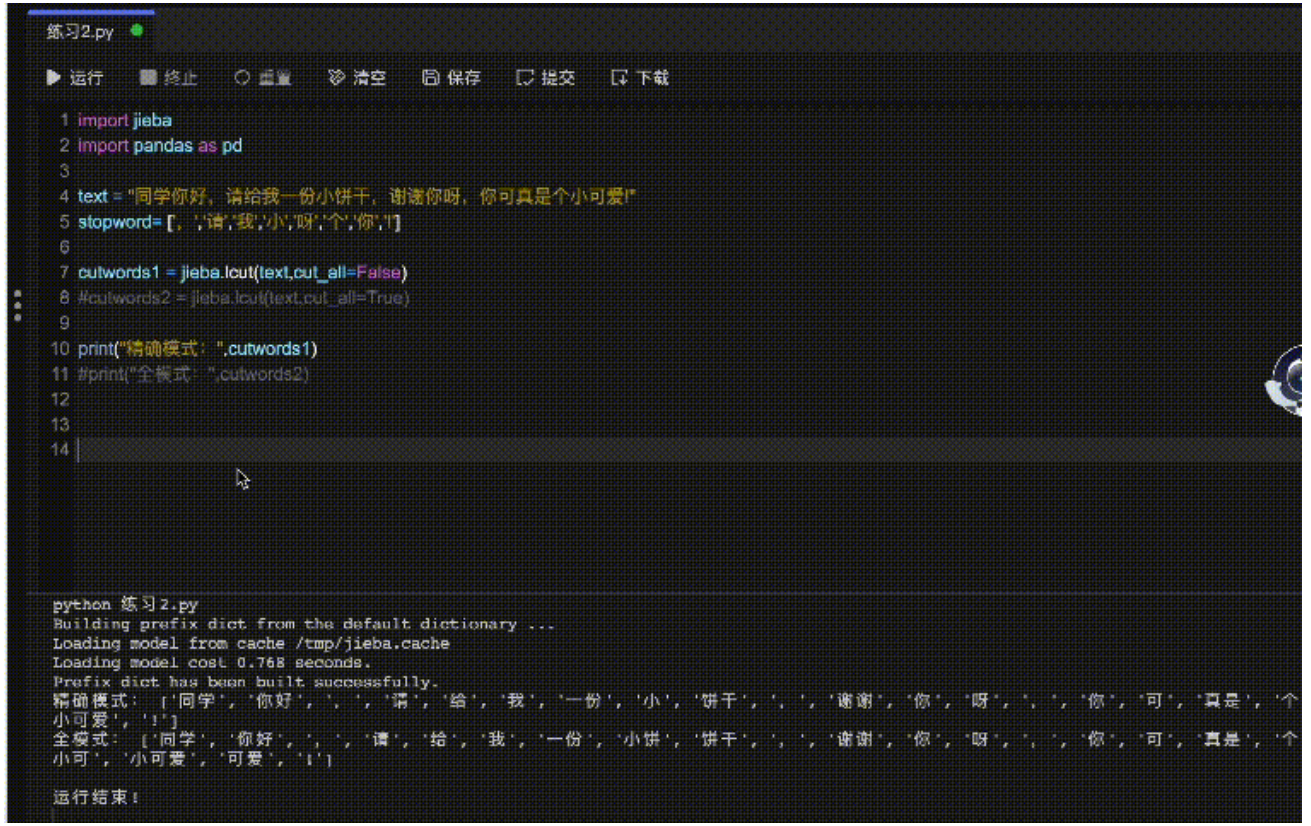
➢ 实训演示-使用停用词库进行优化

```
import jieba
import pandas as pd
text ="同学你好，请给我一份小饼干，谢谢你呀，你可真是个小可爱！“
stopword= [ '，  ','请','我','小','呀','个','你','!',"真是"]

cutwords1 = jieba.lcut(text)
cutwords1 = pd.Series(cutwords1)
cutwords1 = ' '.join(cutwords1[~cutwords1.isin(stopword)])

print("精确模式：",cutwords1)
```





二、通过Python实现文本分析

➢ 使用保留词库，对代码进行优化：

保留词库

对于特殊领域可能存在专有名词，当我们不希望某些长词被分割时，可以将其放入保留词库，保留词库中的词将不被拆分。

保留过程

1.建立保留词库。

本章实战练习中提供了一份财务保留词库，如有特殊需保留的词组，可以在词库中添加。

2.读入保留词库

二、通过Python实现文本分析

➢ 实训演示-使用保留词库进行优化

```
jieba.load_userdict("数据源06保留词库dict.txt")    #载入保留词库
```

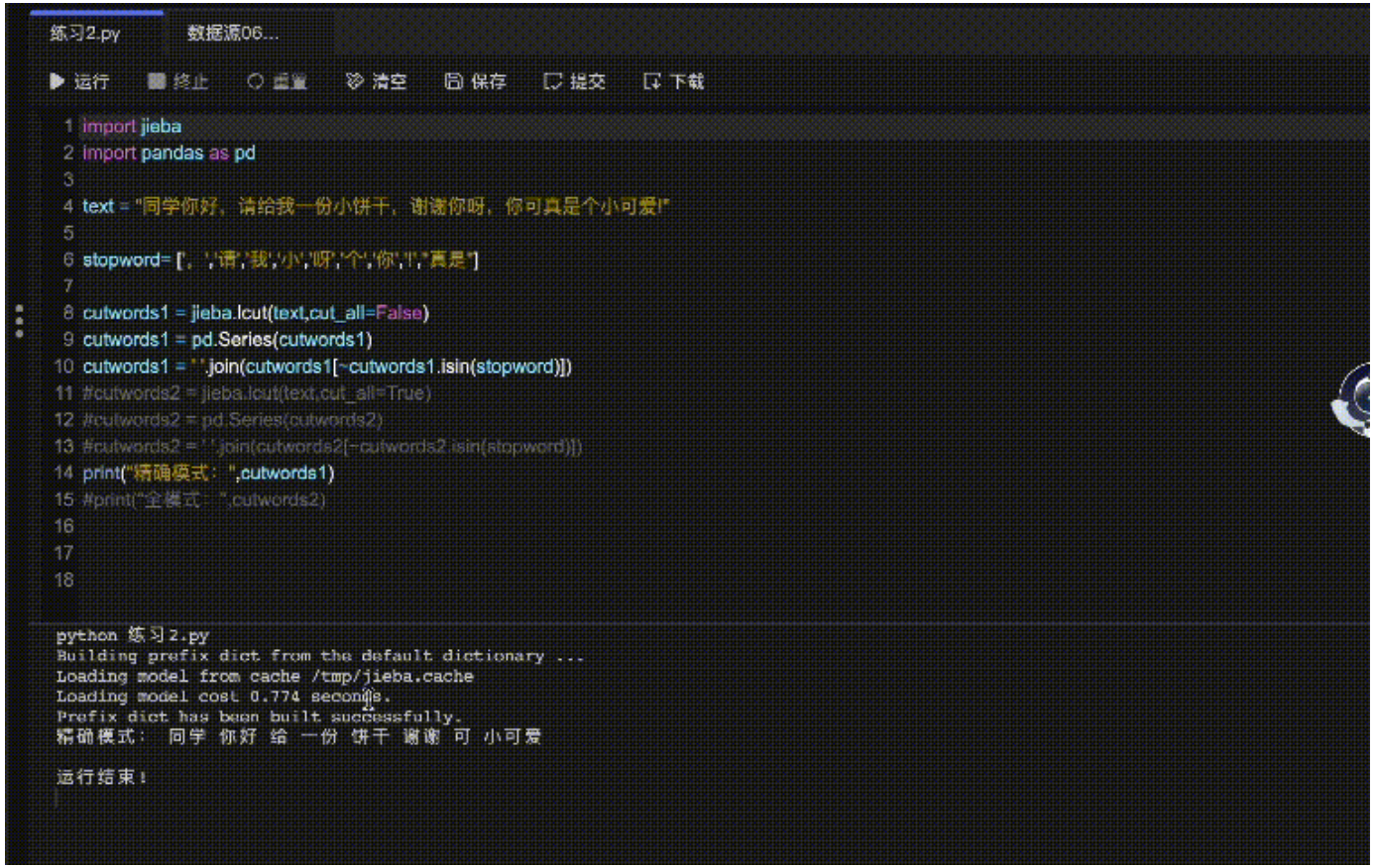
练习2.py

运行 终止 重置 清空 保存 提交 下载

```
1  import jieba
2  import pandas as pd
3
4  text = "同学你好，请给我一份小饼干，谢谢你呀，你可真是个小可爱!"
5  jieba.load_userdict("初步业务活动/数据源06保留词库dict.txt")
6  stopword= [',','请','我','小','呀','个','你','!',"真是"]
7
8  cutwords1 = jieba.lcut(text,cut_all=False)
9  cutwords1 = pd.Series(cutwords1)
10 cutwords1 = ' '.join(cutwords1[~cutwords1.isin(stopword)])
11 #cutwords2 = jieba.lcut(text,cut_all=True)
12 #cutwords2 = pd.Series(cutwords2)
13 #cutwords2 = ' '.join(cutwords2[~cutwords2.isin(stopword)])
14 print("精确模式：",cutwords1)
15 #print("全模式：",cutwords2)
16
17
18
```

```
python 练习2.py
Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
Loading model cost 0.774 seconds.
Prefix dict has been built successfully.
精确模式：  同学 你好 给 一份 小饼干 谢谢 可 小可爱

运行结束！
```



三、大数据审计的文本分析实现

(一)财务报表审计中的文本分析实现

1.通过文本分析梳理会计摘要

企业会计在进行账务处理时应规范地填写凭证摘要，审计师在取得企业明细账的同时也获得了会计凭证摘要的文本数据。这些文本数据隐含着大量信息，通过对摘要分析，可以帮助审计师找到被审计单位隐藏的问题。

例如

审计师对某生产制造企业审计时所做的摘要分析。审计师对被审计单位进行摘要分析，通过汇总摘要，发现该企业会计摘要中“收货款”出现了134次、“产品销售收入”出现了123次。对于生产制造企业，这样的出现频次属于正常情况。同时审计师发现“魏诗彰报销”出现了10次，属于较为异常的情况。审计师调查发现，魏诗彰是被审计单位财务部的一名员工，根据其所处的岗位性质，并不应该发生频繁的报销情况。通过进一步的调查，审计师发现该员工几乎每月都会进行一笔金额较大的报销。通过检查这些报销原始凭证，审计人员发现原始凭证中的发票与会计账务内容不相符，乃该员工通过虚构报销内容套取公司资金。审计师将这一情况与管理层进行沟通，该员工受到了相应的惩处。

在财务报表审计中，由于企业业务量巨大，生成了海量的凭证，审计师无法对所有凭证进行一一检查，通过对摘要进行文本分析可以帮助审计师定位异常凭证，有针对性地进行检查，从而高效地发现被审计单位存在的问题。

三、大数据审计的文本分析实现

(一)财务报表审计中的文本分析实现

2.通过文本分析检查购销合同

- 在传统的审计过程中，审计师需要对被审计单位的各类合同进行人工复核，这个过程消耗了审计师大量的时间，同时也会由于审计师人为判断的失误导致错失重要信息。
- 在大数据时代可以通过文本分析技术，将合同中重点信息进行自动化提取和汇总，审计师只需对汇总后的合同内容进行复核和检查，审计的效率和质量都大幅提升。

例如

审计师对被审计单位130份销售合同进行检查，对所有销售合同进行扫描和文本转换，将纸质合同资料转化为文本数据，对文本数据进行进一步的提取，重点提取“销售方”“收款方”“合同金额”“收款条件”“合同日期”等信息，发现这130份合同中有5份“合同日期”与被审计年份不相符，企业存在提前确认收人的情况。审计师经过与管理层讨论，将这5份合同对应的收入金额进行了审计调整。

文本分析技术帮助审计对大量文本信息进行初筛，使审计师有更多的精力投入具体疑点的解读和处理中，从而提升审计师发现审计疑点的效率和准确性。

三、大数据审计的文本分析实现

(一)财务报表审计中的文本分析实现

3.通过文本分析进行数据预测

文本信息能够用来预测市场反应，股票留言板的信息并非噪声，而是与财务相关的信息，通过文本分析能够预测股票交易量和波动性。

例如

市场指数MSH与前一日的市场情绪显著相关。Loughran和McDonald（2011）构建了五个关于积极词汇、不确定性、诉讼、强势语调、弱势语调的词典，发现采用这些词典度量的语调与超额收益、交易量、回报波动率、未预期盈余等显著相关。前瞻性语句的情感积极程度与分析师报告发布后的投资者反应显著正相关。

通过文本分析，审计师可以脱离财务账表数据进行预测，将预测结果与企业实际财务数据进行比对从而发现审计疑点。