文本分析：从文本到论文-王菲菲 Day3 问答 （2023/11/25）

> 全文由王老师解答的问题、助教解答的问题两部分共同构成。

王菲菲老师答疑

Q1.Text-based innovation*Non patenting firm的交互为啥都是不显著，应该是要显著啊？没有专利的会有创新

Q2. 老师想请教一下，您觉得App自己的描述可以提取情感吗？您觉得有意义吗？

Q3.请问主题模型可以做中文吗，应该在何处修改代码，可否请老师分享一下

Q4. 老师写motivation您一般会涉及哪些方面，通过哪些角度去写？（想老师回答，谢谢）

Q5.这边的innovation会不会有内生性问题？

Q6.老师想请教一下关于参考的文本一般通过什么途径去找啊？常用的网站都有哪些呀？谢谢老师

Q7.老师想问一下，我用摘要做LDA，结果困惑度是递增的，想问一下如何调整。想问一下如果调不出来的话是否只看一致性最高的来选择主题数

Q8.老师感觉找参考文本有难度，如果是找相对应期刊的话（大约几篇以上如何）。如果真的没有怎么办？感觉难度还是很大。

Q9.在用R做jieba分词时如果想要在自定义词典的基础上再加一个同义词词典，加了 `synonym="./同义词.txt"` **之后，** `cutter = worker(bylines = TRUE,stop_word=stopwords, user = dictpath)` **这里的代码怎么处理呢（如何设定同义词词典）。**

Q10.请问回归时控制变量如何选取？关于创新对企业绩效的影响研究，不同的文章在控制变量的选取上有所不同。

Q11.多品牌文本评论联合主题模型（3个不同品牌进行LDA）。然后每个产品与组合LDA求相似度?

Q12.公共主题数（困惑度虽然是2最好，但是从一致性看出最高的是7，这如何去取舍）特有主题数的数量如何来确定?

Q13.老师想请教一下，联合主题建模在产品创新方面是否也适合？就是App开发商，他们会开发不同类别的App，比如同一个开发商同时会开发娱乐类App同时也会有游戏类。但是可能他们主题是差不多的。想问一下开发商新产品开发定位是否也可以拓展。（想老师来回答，谢谢！）

Q14.在跑LDA，超参数的两个变量如何来调参。一般怎么设置?

> LDA模型中的α和β值的选择取决于具体的文本体裁、主题数量和词汇量，因此在实际应用中，alpha和beta的设置需要根据具体的任务和数据进行调整。一般而言，alpha值越大，生成的文档中的主题越多，即文档-主题分布更均匀。相反，alpha值越小，生成的文档中的主题越少，即文档-主题分布更稀疏。对于beta值，越大表示每个主题中的词越多，即主题-词分布更均匀；而越小表示每个主题中的词越少，即主题-词分布更稀疏。
>
> 关于alpha和beta参数，我个人的经验建议是：以使用和你研究类似语料的文献中的推荐值为起点，然后在自己的语料上进行反复迭代（在起点值附近一个较小窗口，基于这两个超参数使用网格搜索，强烈建议以最终的主题模型效果优劣进行反复迭代，根据最终聚类的主题效果选择最佳的超参数组合，因为经管领域的文本分析更加注重主题模型的可解释性。）
>
> 推荐文献，Huang et al.(2018) 在电话会议文本和分析师报告文本上进行LDA主题建模，他们选择的α和β值分别为0.1和0.01。在实际应用中，这可以作为一个合理的参考，但最终的超参数选择应该根据你的具体任务和数据来调整。
>
> 参考文献：*Huang A H, Lehavy R, Zang A Y, et al. Analyst information discovery and interpretation roles: A topic modeling approach[J]. Management science, 2018, 64(6): 2833-2855.*
>
> （回复人：范思妤；整理人：范思妤）

Q15.想问一下老师目前在研究文本分析的什么方向。对于博士生在哪方面进行努力去挖掘有推荐。老师现在平台经济很火，老师有没有推荐的方向？谢谢。

Q16.老师想问一下，如果自己去给文本打标签，想问一下这个怎么操作?

Q17. 老师想请教一下关于主题的选取可以详细讲一讲吗？主题选取的个数取决于什么呀？谢谢老师。

Q18. 请教老师和各位助教，文本数据样本量不到100条的情况下（特定人群的面对面访谈数据）这种样本量可以做LDA、词向量、词嵌入、深度学习方法吗？（这些方法分别建议最低样本量有多少呢）

> 对于文本数据样本量不到100条的情况，使用一些复杂的自然语言处理方法可能会面临一些挑战。以下是对于不同方法的建议最低样本量：

> LDA（Latent Dirichlet Allocation）：LDA是一种用于主题建模的统计方法。虽然LDA可以在较小的样本量上工作，但是为了获得更好的结果，建议至少有几百条样本。
>
> 词向量（Word Embeddings）：词向量是将单词映射到连续向量空间的表示方法。对于基于预训练的词向量模型（如Word2Vec、GloVe等），建议至少有1千条样本，以获得更准确的词向量表示。
>
> 词嵌入（Word Embedding）：词嵌入是将单词映射到低维向量空间的表示方法。与词向量相似，建议至少有1千条样本，以获得更好的词嵌入表示。
>
> 深度学习方法：深度学习方法通常需要大量的数据来训练复杂的神经网络模型。对于基于深度学习的方法（如循环神经网络、卷积神经网络等），建议至少有2千条样本，以获得较好的性能。（回复人：马洪栋；整理人：范思妤）

Q19.请教老师和各位助教，我的文本数据量很大，到了TB量级，在Python中无法实现并发编程，请问R是否可以实现多线程处理？如果R也难以处理的话，是否有合适的服务器推荐呢。

> 至于R语言，它有多线程的支持。R中的parallel包提供了一些函数，例如mclapply，可以用于并行化处理任务。这里的并行化通常是通过多进程来实现的，而不是多线程，因为和Python一样，R在解释器层面也有一些全局的限制。
>
> 对于服务器的选择，云服务提供商如华为云、阿里云等都提供了强大的计算资源，可以方便地扩展到TB级别的数据处理。这些云服务还提供了各种计算和存储服务，以及支持分布式计算框架的工具和服务。
>
> 具体选择哪种工具和服务取决于你的需求、预算和对平台的偏好。在进行大规模数据处理时，通常考虑到可扩展性、性能、易用性以及成本等方面的因素。（回复人：马洪栋；整理人：范思妤）

Q20. 老师好，想请教一下目前chatpgt在编写代码，建模，包括文本分析和论文写作方面的能力很强大，如何在chatgpt广泛使用的大环境下提高自己的科研竞争力？目前的研究方向也是文本分析，感谢！

>

Q21. STM和LDA的区别是啥？哪个更推荐，是否STM更新?

>

Q22. 主题模型最开始的时候并不知道具体k个主题是什么，那模型内部用的是什么方法将所有的词在k个主题上分配概率的？是根据词向量在空间中的距离吗？还是其他什么机制？网上看到一句话"主题模型就是从一篇或多篇文章(文本)中，找出关键词，并依照关键词之间的相似度提炼成主题"，所以这句话对吗？这里的相似度指的是什么相似度?

>

Q23. 感觉Term- selection term听是听得懂，BUT如果要自己用还是不会用。一般是在评论里会使用?

>

Q24. 老师不同的产品评论数不同，如果后面去跑面板的话是不是直接用非平衡面板就行?

>

助教答疑

Q1.为什么说相对固定的文本结构更有利于LDA建模?

> **A:** LDA（Latent Dirichlet Allocation）是一种常用的主题建模算法，用于发现文本数据中的潜在主题。相对固定的文本结构对LDA建模有利的原因如下：
>
> 1. 主题一致性：相对固定的文本结构通常具有一致的主题分布。例如，新闻文章通常会按照标题、导语、正文等固定结构组织，这意味着不同的新闻文章在主题上可能存在一定的一致性。LDA可以利用这种一致性来发现并建模这些主题。
> 2. 上下文关联：相对固定的文本结构可以提供上下文关联的线索。例如，在一篇科技文章中，引言部分可能提供了关于主题的背景信息，而正文部分可能详细讨论了该主题的各个方面。LDA可以利用这些上下文关联来更好地理解和建模主题之间的关系。
> 3. 分割和过滤：相对固定的文本结构可以帮助分割和过滤文本数据。例如，在一篇博客文章中，可以将标题、正文、评论等部分分开处理。这样的分割和过滤可以帮助LDA更好地聚焦于特定的文本部分，提高建模的准确性和效率。
> 4. 先验知识：相对固定的文本结构可以提供先验知识。例如，在法律文件中，常见的结构包括引言、定义、条款等。这些结构可以为LDA提供有关主题的先验知识，帮助更好地解释和建模文本数据。
>
> 总的来说，相对固定的文本结构提供了一种有序和一致的方式来组织文本数据，这对于LDA建模非常有利。它可以帮助发现主题一致性、利用上下文关联、进行分割和过滤，并提供先验知识，从而提高LDA建模的准确性和解释能力。
>
> （回复人：马洪栋；整理人：范思妤）

Q2. 删除不足100词或超过5847词的研究报告（这边的100词和5847次的依据是什么？是自己定义），后面情绪处于前1/4的报告。（1/4也有依据吗，还是自己定义？）

> **A:** 低于100的话是字数太少会导致分析结果的有偏差。5847是第98%的分位数。论文的原文是这么写的：为了集中分析同质的分析师报告，我们删除了清理后剩余字数少于 100 字或超过 5,847 字（第 98 百分位数）的报告。经过文本处理和与 Compustat 的识别匹配，我们最终得到了 665,714 份报告样本，并在此基础上进行了文本分析。
>
> 前1/4的原因是选取了比较积极向上的调性的文本。原文中是这么写的：对于我们的主要衡量指标，我们不考虑分析师报告中情绪低于第 75 百分位数的创新语言。
>
> 我已经把论发到了微信群里面了。
>
> （回复人：马洪栋；整理人：范思妤）

Q3. 具体是如何把基于文本分析构建的新指标放入传统计量模型中的，是如何作为变量放进去回归的

> **A:** 将基于文本分析构建的新指标纳入传统统计模型中的具体步骤可以根据具体情况而异，但以下是一般的指导原则：
>
> 1. 确定目标：首先，明确你希望通过文本分析构建的新指标在传统统计模型中的作用和目标。确定你希望该指标对模型的哪个方面进行改进或增强。
> 2. 数据准备：收集和准备用于文本分析的数据。这可能包括文本文档、社交媒体帖子、新闻文章等。确保数据的质量和完整性，并进行必要的预处理步骤，如文本清洗、分词、去除停用词等。

3. 特征提取：使用适当的文本分析技术从原始文本数据中提取有用的特征。这可以包括词袋模型、TF-IDF、词嵌入等。根据你的需求和数据的特点，选择适合的特征提取方法。
4. 构建新指标：根据你的目标和特征提取结果，设计并构建基于文本分析的新指标。这可能涉及计算文本相似度、情感分析、主题建模等技术。确保新指标能够提供有关文本数据的有用信息。
5. 整合到传统统计模型中：将新指标与传统统计模型进行整合。具体方法取决于你使用的统计模型类型。例如，如果你使用线性回归模型，可以将新指标作为额外的自变量添加到模型中。如果你使用决策树模型，可以考虑将新指标作为分裂节点的依据。
6. 模型评估和调整：将整合了新指标的传统统计模型进行评估和调整。使用适当的评估指标来评估模型的性能，并根据需要进行调整和改进。

需要注意的是，将基于文本分析构建的新指标纳入传统统计模型中可能需要一定的领域知识和技术经验。确保在实施过程中进行适当的验证和验证步骤，以确保新指标的有效性和可靠性。（回复人：马洪栋；整理人：范思妤）

Q4. 控制替代解释下LDA模型的稳健性那里（分析师使用的revenue,growth，technology）这些词是我对掉再做吗?

把这些词作为因变量，替换掉原来的因变量ROA，自变量和控制变量不变，然后进行回归即可。（回复人：马洪栋；整理人：范思妤）

Q5. 请问如何对上万份上市公司PDF年报提取供应链金融相关的词，并进行词频分析？目前示例代码都是Excel数据？想请问PDF文档怎么处理？很多份文档下，如何循环处理？是否有这方面的代码可供学习？谢谢。

pdf提取txt / csv这个任务有多个 R/ Python包可以完成。

1. 该网址对应的教程中，用Python的 `PyPDF2` 库和 `pdfminer.six` 库读取PDF年报内容，用Python的Beautiful Soup解析年报中的文本内容。具体代码请参考：https://wenku.csdn.net/answer/eee5ee72e70711edbcb5fa163eeb3507
2. 下面提供的代码示例是使用Python中的 `pdfplumber` 包。

   假设我们的多个pdf文件已经下载储存在我们电脑本地的文件夹folder中了，下面这段代码的任务目标是遍历文件夹中储存的多个pdf，依次将其中的文本依次抽取出来，并写入一个csv文档，csv文档中的一行对应文件夹中的一个pdf：：

```
import os
import pdfplumber
import pandas as pd

# 定义文件夹路径，此处假设我们把pdf存在了电脑本地的folder文件夹
folder_path = r'./folder'

# 初始化一个空的 DataFrame
# 它有两个列，一个是文件名File_Name，一个是文本Text
df = pd.DataFrame(columns=['File_Name', 'Text'])

# 遍历文件夹下的所有 PDF 文件
for filename in os.listdir(folder_path):
    if filename.endswith('.pdf'):
        pdf_path = os.path.join(folder_path, filename)
```

```
        with pdfplumber.open(pdf_path) as pdf:
            pdf_text = ""
            for page in pdf.pages:
                try:
                    page_text = page.extract_text()
                    pdf_text += page_text
                except Exception as e:
                    print(f"Error extracting text from {pdf_path}: {e}")
                    pass

            # 将提取的文本添加到 DataFrame 中
            df = df.append({'File_Name': filename, 'Text': pdf_text},
ignore_index=True)

# 写入 DataFrame 到本地 CSV 文件
csv_path = os.path.join(folder_path, 'pdf_text_data.csv')
df.to_csv(csv_path, index=False, encoding='utf-8')
```

（回复人：马洪栋，范思妤；整理人：范思妤）

Q6. 请问如何对上万份上市公司PDF年报提取供应链金融相关的词，并进行词频分析？目前示例代码都是Excel数据？想请问PDF文档怎么处理？很多份文档下，如何循环处理？是否有这方面的代码可供学习？谢谢。

R语言分析方法：

```
1.调入包
library(pdftools)
library(jiebaR)
library(tidyverse)
2.读入并提取所有 PDF 文档的内容
fs::dir_ls('pdf') %>%
as_tibble() %>%
set_names('file') %>%
mutate(text = map_chr(file, function(x){
pdf_text(x) %>%
paste0(collapse = "") %>%  str_remove_all("[\\s\\n\\t\\d[a-z].]")
})) -> textdf
```

这时候的 textdf 是一个数据框，里面是每个文档对应的文本内容，可以对其进行文本分析。（回复人：曹昊煜；整理人：范思妤）

Q7. 2021年2022年间的券商研报，这边是面板数据吗？

可能不是常见的面板数据形式，因为可能存在同一家企业的不同研究报告。但是可以通过汇总到 "企业-年份" 层面来构造面板数据。（回复人：曹昊煜；整理人：范思妤）

Q8. 2293家公司来自于哪里？A股，科创板，创业板？

东方财富网的A股（例如中国化学601117），创业板（例如晶瑞电材300655）和深圳主板（例如盐津铺子002847）。我通过查询老师PPT的截图中的公司代码，找到的上述公司来源。（回复人：马洪栋；整理人：范思妤）

Q9. 老师能否给出实现KL散度计算并画图的代码?

> 老师发的程序中"7 综合实践1"中的"计算KL散度，选择创新主题" 提供了这个代码。（回复人：马洪栋；整理人：马洪栋）

Q10. 研报和中国知网的内容如何导入 Excel？应该不是手工复制粘贴吧?

> 一般来说，研报和中国知网的内容并不能直接导入Excel，但可以通过爬虫实现。具体的操作方法可能需要查阅资料，这两篇推送比较细，可参考：
>
> https://zhuanlan.zhihu.com/p/599579339?utm_id=0
>
> https://www.zhihu.com/question/49292600/answer/2472050342?utm_id=0
>
> 另外，对于一些可以下载的研报，你可以先将其保存为文本文件或者CSV文件，然后再通过数据处理来导入。对于中国知网，一些文章有提供数据表可以直接下载为Excel文件。（回复人：邱一崎；整理人：范思妤）

Q11. 请问上午代码案例中的dict.txt文档在worker时起到什么作用？是不是为了让dict.txt中的词语不会被分解?

> 是的，dict.txt 中一般添加一些专用的词汇，确保分词的时候不会把它们分开。

Q12. 什么叫做多模态?

> 多模态（Multimodal）指的是在一个系统或环境中同时使用多种不同的感知模态或信息来源。这些感知模态可以是视觉、听觉、触觉、语言等不同的感知方式。
>
> 举个例子，考虑一个智能语音助手，如Amazon Alexa或Google Assistant。这种智能助手结合了多种感知模态，包括语音识别、自然语言处理、语音合成、图像识别等。用户可以通过语音与助手进行交互，而助手则通过语音识别将用户的语音转换为文本，然后使用自然语言处理技术理解用户的意图，并通过语音合成将回答转换为语音输出给用户。此外，助手还可以通过图像识别来理解和处理与图像相关的请求，例如识别物体、扫描二维码等。
>
> 在这个例子中，语音、文本和图像是多模态系统中的不同感知模态。通过结合多种感知模态，智能助手能够提供更丰富、更全面的交互体验，并能够更好地理解和满足用户的需求。
>
> 多模态技术在许多领域都有应用，如人机交互、智能交通系统、医学诊断等。通过整合多种感知模态的信息，可以提供更全面、准确和丰富的数据，从而改善系统的性能和用户体验。（回复人：马洪栋；整理人：范思妤）

Q13. 请问一下需要批量进行文本翻译,英文翻译成中文使用什么API。

> 1.Google Cloud Translation API:
>
> Google Cloud提供了翻译API，支持多种语言翻译，包括英文到中文的翻译。你需要创建一个Google Cloud账户，并设置API密钥来使用该服务。
>
> 2.Microsoft Azure Translator API:
>
> 微软Azure提供了翻译服务的API，可以用于批量翻译文本。你需要创建一个Azure账户，并设置相应的认证信息。
>
> 3.百度翻译API:
>
> 百度翻译开放平台 提供了中文翻译服务的API。你需要在百度翻译开放平台注册一个账户，获取API密钥。

4.腾讯云翻译API:

腾讯云提供了文本翻译的API服务，支持多种语言。你需要在腾讯云注册并创建一个API密钥。

API调用的代码逻辑基本相似，但是具体的API调用代码需要去参考每个平台的官方文档，一般官方文档里面会给出详细的调用代码和相关参数说明。此外，在使用这些API之前，需要了解每个服务的使用限制、费用和服务条款。一般来说，各个平台的API都会有请求速度/总量的限制，需要按照自己的需求进行充值。（回复人：范思妤；整理人：范思妤）

Q14. 机器学习做简单的文本情感分析结合交叉验证，最低样本量需要多少，几十条文本数据可以吗?

通常来说，机器学习在进行文本情感分析时，需要相对较大的数据集来训练模型，以便让模型学习到足够泛化的规律。对于简单的情感分类任务，几十条文本数据可能会不够。

交叉验证是一种评估模型性能的技术，通常用于验证模型的泛化能力。然而，如果数据量太小，即使使用交叉验证也难以保证模型在未见过的数据上表现良好。

一般来说，对于文本情感分析，建议的最低样本量是数千条甚至更多，特别是如果希望构建泛化能力较强的模型。这样才能确保模型学习到各种情感表达的差异和特征。

如果只有几十条文本数据，可能会出现以下问题：

过拟合：模型可能会记住训练数据中的特定模式，而无法泛化到新的数据。

代表性不足：文本数据量少可能无法涵盖各种情感表达和语境，导致模型偏向某种特定类型的数据。

如果你的数据量确实非常有限，可以尝试使用一些预训练的模型或者基于迁移学习的方法，利用在大规模数据上训练过的模型来进行微调，有时候这种方法能够在少量数据上取得一定效果。但要注意，即便如此，仍然可能无法获得非常好的泛化能力。

总体来说，对于机器学习任务来说，样本量越大，模型往往表现越好，尤其是对于自然语言处理任务这种需要理解语境和情感的任务

（回复人：马洪栋；整理人：范思妤）

Q15. 文本数据样本量不到100条的情况下（特定人群的面对面访谈数据）这种样本量可以做LDA、词向量、词嵌入、深度学习方法吗？（这些方法分别建议最低样本量有多少呢）

对于文本数据样本量不到100条的情况，使用一些复杂的自然语言处理方法可能会面临一些挑战。以下是对于不同方法的建议最低样本量：

LDA（Latent Dirichlet Allocation）：LDA是一种用于主题建模的统计方法。虽然LDA可以在较小的样本量上工作，但是为了获得更好的结果，建议至少有几百条样本。

词向量（Word Embeddings）：词向量是将单词映射到连续向量空间的表示方法。对于基于预训练的词向量模型（如Word2Vec、GloVe等），建议至少有1千条样本，以获得更准确的词向量表示。

词嵌入（Word Embedding）：词嵌入是将单词映射到低维向量空间的表示方法。与词向量相似，建议至少有1千条样本，以获得更好的词嵌入表示。

深度学习方法：深度学习方法通常需要大量的数据来训练复杂的神经网络模型。对于基于深度学习的方法（如循环神经网络、卷积神经网络等），建议至少有2千条样本，以获得较好的性能。（回复人：马洪栋；整理人：范思妤）

Q16. 情感词典法进行中文文本情感分析时，想要把程度副词考虑进去，比如"非常好"的正向情感得分比"挺好"更高，如何为情感词的程度引入权重系数呢，在软件中具体如何操作?

一种常见的做法是使用一个预定义的程度词典，为每个程度副词分配一个权重值，并在计算情感得分时考虑这些权重。

大致的文本分析逻辑如下（下面的例子经过了简化，具体应用中需要按照你自己的文本语料来进行调整，人工预设词典时最好经过多人复核，以确保词典相对合理科学。）

```
1.定义情感词典和程度副词
sentiment_dict <- data.frame(
  word = c("好", "差"),
  weight = c(1, -1)
)
degree_dict <- data.frame(
  degree_word = c("非常", "挺", "很"),
  weight = c(1.5, 1.2, 1)  # 这里是示例权重，根据实际情况调整
)

2.使用分词工具对文本进行分词：
library(jiebaR)
text <- "非常好的产品，性价比挺高的。"
seg <- worker()$segment(text)

3.遍历分词结果，匹配情感词和程度副词，并根据它们的权重系数计算情感得分：
#定义函数
get_sentiment_score <- function(seg, sentiment_dict, degree_dict) {
  score <- 0
  i <- 1
  while (i <= length(seg)) {
    word <- seg[i]
    if (word %in% sentiment_dict$word) {
      sentiment_weight <- sentiment_dict$weight[which(sentiment_dict$word ==
word)]

      # 检查前一个词是否是程度副词
      if ((i - 1) >= 1) {
        prev_word <- seg[i - 1]
        if (prev_word %in% degree_dict$degree_word) {
          degree_weight <- degree_dict$weight[which(degree_dict$degree_word ==
prev_word)]
          score <- score + sentiment_weight * degree_weight
          i <- i + 1  # 跳过情感词，因为已经处理过了
        } else {
          score <- score + sentiment_weight
        }
      } else {
        score <- score + sentiment_weight
      }
    }
    i <- i + 1
  }
  return(score)
}
```

```
# 调用示例
sentiment_score <- get_sentiment_score(seg, sentiment_dict, degree_dict)
print(sentiment_score)
```

有篇文章和您说的问题高度关联，您可以参考：***Bochkay K, Hales J, Chava S. Hyperbole or reality? Investor response to extreme language in earnings conference calls[J]. The Accounting Review, 2020, 95(2): 31-60.***

（回答人：范思妤；整理人：范思妤）

Q17. 如何把网上下载的pdf研报转换成csv 结构的文档?

请参考助教答疑-问题5。（回复人：范思妤；整理人：范思妤）

Q18. 请问老师和各位助教老师backward_bic函数来自哪个包?

具体哪一个包不知道，但是你library下面的所有包，一定不会报错。

我自己昨天亲测好用，程序不会报错且程序能够正常运行。

```
library(knitr)
library(rmdformats)
library(jiebaRD)
library(jiebaR)
```

（回复人：马洪栋；整理人：范思妤）

Q19. 好评数/总评论数。如何从用户的评论来判断是否是好评?

前面讲过文本情感分析，文本情感分析可以判断是好评还是差评。具体判断方法请见老师发的PPT中的“3. 文本情感分析”，并且老师提供了相应的中英文数据和代码，数据和代码请见“3.情感分析-词典法”下面有英文情感分析示例和中文情感分析示例。（回复人：马洪栋；整理人：范思妤）

Q20. 想问下老师处理同时包括文本数据和企业财务数据时，是都放在一个CSV/EXCEL文档里直接调用分析吗，还是用其他数据库工具，例如SQL，那种比较方便呢

一般来说文本数据会比较大，因此和财务数据不太好放在一起处理。总体来说，需要在软件中现将文本数据处理好（比如提取词频等操作），然后将数据在“企业-年份”层面汇总好，之后再与财务数据匹配起来。（助教：曹昊煜；整理人：范思妤）

Q21.变系数是什么意思？beta会变？如何解决?

变系数是变系数模型。是的，beta会变。变系数模型不仅允许每个案例拥有自己的截距项，还允许每个案例拥有自己的回归方程斜率，旨在进行随机系数模型回归分析。一般的回归模型不允每个变量拥有自己的斜率。变系数模型的STATA命令是

```
xtrc profit income cost, betas
```

代码解析：xtrc命令的意思是进行随机系数模型估计。其中，选择项betas的意思是显示对每组系数的估计。profit是因变量，income和cost是自变量。

（回复人：马洪栋；整理人：范思妤）

Q22. BIC选择如何做?

> 先回答什么是BIC。BIC(Bayesian Information Criterions)，贝叶斯信息准则。BIC和AIC准则的第一项是一样的，但BIC对复杂模型的惩罚力度比AIC大，考虑了样本数量，更倾向于选择简单的模型。样本数量过多时，可有效防止模型精度过高造成的模型复杂度过高。
>
> 再回答如何做。如果是回归模型，可以使用老师给的代码，具体请见"8.综合实践2-JBES复现"中的代码，具体是在"3 建模，在某个q下进行筛选"。如果是二分类问题，我自己平时使用的代码如下所示，该代码亲测好用，不过要替换成自己的数据。

```
library(e1071)
library(pROC)
library(caret)
set.seed(1)  #随机数种子是1，目的是可重复实现
data<-read.csv("D:/2020.12.31.德国数据集原始数据预处理结果.csv",header = T)
#读入数据含表头， header=T的含义是保留表头
n =1000 #读入样本个数
train_obs = sample(nrow(data), round(nrow(data)*0.8))
# row的意思是一行，nrow的含义是行数，后面的nrow(data)的含义是data里面的行，
round(nrow(data)*0.8)的含义是返回data数据集0.8比例的行的近似整数（四舍五入）。
#train_obs的含义是把上面的近似数的行赋给变量train_obs。数理统计里的"obs"是observation的缩写。
#observation指的是观测值或实测值，与其对应的是统计模型（例如线性模型）的预测值(predicted value)
datatr = data[train_obs, ]  #datatr的含义是训练集，英文datatrain，把train_obs赋给变量datatr
datate = data[-train_obs, ]  #datate的含义是测试集，英文datatest，负号的含义是把data中把train_obs剩余的行赋给变量datate

logit = glm(loan_status~., data = datatr, family = binomial)
#glm是广义线性模型
#loan_status~.的含义是loan_status是因变量，~是间隔，.是省略自变量。
#family：每一种响应分布（指数分布族）允许各种关联函数将均值和线性预测器关联起来。如
binomal(link='logit') --响应变量服从二项分布，连接函数为logit，即logistic回归。binomal的英文的二项式，二项分布
#binomal(link='probit') --响应变量服从二项分布，连接函数为probit。
#poisson(link='identity') --响应变量服从泊松分布，即泊松回归
BIC = step(logit, direction = "both", k = log(n), trace = 0)   #BIC逐步回归
# R语言数据挖掘与机器学习P85写道，AIC值的第二项是+2P，P是参数的个数，这里K=2的含义是对模型参数个数（模型复杂度）的惩罚，AIC最小值的含义是KL距离相对熵最小。
summary(BIC)
```

> （回复人：马洪栋；整理人：范思妤）

Q23. 请问如何去掉零长度变量名？今天下午的示例代码出现这个报错。

> 可能是运行了 RMD 文档中的 ```，不要运行这个代码块符号可以解决这个问题。
>
> （回复人：曹昊煜；整理人：范思妤）

Q24. 请问安装 `dplry` 出现报错如何处理?

> 你试下 `library("dplyr")`？看看这个程序有没有安装上？
>
> 似乎没有什么大问题，如果一定要解决这个问题，可以首先使用 `.libPath()` 来得到包的安装路径，在里面删掉 `dplyr` 之后重新安装。
>
> （回复人：曹昊煜；整理人：范思妤）

Q25. 请问下面代码中tag是什么意思？tag用来做什么？为什么要分词两次？

```
cutter = worker('tag',bylines = TRUE,stop_word=stoppath,user = dictpath)
res = cutter[comments]
cutter2 = worker('tag',bylines = TRUE,stop_word-stoppath,user = dictpath,symbol=T)
res2=cutter2[comments]
```

> tag的意思是词性标注，这是参数"type"里面的取值，type的作用是分词引擎类型，这里包括mix,mp,hmm,full,query,tag,simhash,keyword，分别指混合模型，支持最大概率，隐式马尔可夫模型，全模式，索引模型，词性标注，文本simhash相似度比较，关键字提取。其中，tag的意思是词性标注。
>
> 为什么要分两次？答：第一次和第二次的区别仅仅是后者的symbol=T，symbol默认是F，symbol的意思是输出是否保留符号，第二次是让分词包含了符号。经过核实，后面只用到了第一个res，没有用到第二个res2。后面仅将res的每行的分词结果进行筛选。因此，不用管这个。
>
> 第二个分词只是让分词包含了符号，res2 在后面似乎完全没有用到，供参考。
>
> （回复人：马洪栋；整理人：范思妤）

Q26. 上午课程券商研报我在运行提取主题词的代码时候为什么每次提取的主题词都不一样？请问是否是我的运行出错？如果无错的话，如何确定哪次运行的结果最好？

> 可能是模型的随机性导致的，可以尝试在代码中设置种子值，保证结果一致
>
> ```
> lda <- LDA(dtm, k = 8,control =list(seed =1234))
> ```
>
> 也可以在library包的后面直接写：
>
> ```
> set.seed(1)  #随机数种子是1，目的是可重复实现
> ```
>
> 这样能保证你能以后每次程序结果均保持一致。
>
> （回复人：曹昊煜；整理人：范思妤）

Q27. 老师数据集那边缺失值，直接缺失值处理还是统一为0？

> 缺失值的解决方法是对于定性变量，一般情况下用众数填补；对于定量变量，一般情况下用均值或者中位数填补。用这些方法，审稿人不会质疑你。
>
> 如果想要更高级的缺失值填补方法，可以使用缺失森林，R语言的包是missForest。如果不是专门研究缺失值处理，不建议使用缺失森林，这个程序运行速度相当相当慢。我之前用缺失森林方法填补2000条缺失比例为30%的数据，耗时大约1天。
>
> （回复人：马洪栋；整理人：马洪栋）

Q28. 案例分析中代码打开来文字注释显示乱码？怎么转化下？

> 是 Rstudio的编码问题导致的，可以使用以下的步骤解决：
>
> 1. RStudio菜单栏的Tools -> Global Options
> 2. Code -> Saving -> Default text encoding
> 3. 修改为 UTF-8 即可
>
> （回复人：曹昊煜；整理人：范思妤）