TEXT-文本分析：从文本到论文-王菲菲 Day1 问答 （2023/11/11）

> 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。

请王老师解答的问题

Q1.文本分析在经济管理领域得到广泛的应用，老师也在深入研究的基础上转向其他的方向。老师觉得从方法论的角度，文本分析面临哪些批评，文本分析方法又是如何回应这些批评的？

> **A:** XXX（回答人：王老师；整理人：XXX)

Q2.请问老师，文本分析的适用边界在哪里呢？比如最近有争议的从企业年报中提取“数字化”相关词频当做企业数字化转型的代理变量。

> **A:** XXX（回答人：王老师；整理人：XXX)

Q3.如果想要研究媒体在能源领域对于化石能源和非化石能源的态度是否可行呢？现在媒体的真实态度可能是比较隐晦的，这种情况下态度如何界定，有好的方法方法可以帮助我们获得或者处理得出更接近真实态度的文本数据吗？

> **A:** XXX（回答人：王老师；整理人：XXX)

Q4.不同的文本来源，如何区分其衡量的聚焦点：比如老师在举例时提到，在高管发布的会议纪要中查找“可持续发展”相关的词汇，以衡量可持续发展。但同样的词汇，在年报中或者是在年报的M&D中进行查找，是不是也能衡量企业的可持续发展？各个文本数据来源有何区别？

> **A:** XXX（回答人：王老师；整理人：XXX)

Q5.在中文文本情境下，迷雾指数的衡量有通用的R代码或者文献参考吗

> **A:** XXX（回答人：王老师；整理人：XXX)

Q6.举一个产品描述的文本，我如何提取它的使用场景化个数的呢？比如一个化妆镜，产品描述时说消费者可以出差时使用，可以放在家里的化妆间等，那这里有提到了两个场景。另外一个例子：黄金的转运珠，产品描述中提到可以带在手上，寄在小孩脚上，也可以给妈妈带在脖子上跳广场舞，这段文本中有三个场景。

> **A:** XXX（回答人：王老师；整理人：XXX)

Q7.老师，能否简要介绍下数字经济的行业划分的一些关键点呢？

> **A:** XXX（回答人：王老师；整理人：XXX)

Q8.老师为我们讲了很多文章的主要内容，代码可以自己下来跑。可否找篇中文的，带着我们详细讲内容，然后数据获取、分析呢？整个听下来还是不知道如何用文本分析写作，完全没有概念呀，有种感觉在翻译文章。老师很厉害，可以我们是小白

> **A:** XXX（回答人：王老师；整理人：XXX)

Q9.老师想请教一下App的描述有什么现有词典推荐

> **A:** XXX（回答人：王老师；整理人：XXX)

Q10.选取单词的词频设置有没有什么标准？或者参考依据？

> **A:** XXX（回答人：王老师；整理人：XXX)

Q11.之前没有做文本分析的基础，感觉老师讲的内容很专业，但是有点听讲座的感觉，能否用具体的项目讲解文本分析具体怎么做呢？比如从爬取数据开始，预处理及分析过程，这样是不是比您分享的经验总结给小白更深的感

> **A:** XXX（回答人：王老师；整理人：XXX)

Q12.老师在基于文本分析算指数（在讲刻画经济政策不确定那页PPT中）还是不是很清楚Yit=xit/方差。这个的作用是什么。以及后面的按月份对十种报纸进行平均的意义以及后续的EPU指数这个的代码可以分享吗?

> **A:** 【3】这个是Baker等(2016)构建EPU指数的网站，里面提供了EPU数据以及代码指南 https://www.policyuncertainty.com/index.html （回答人：王老师；整理人：XXX)

Q13.上课提到的，复杂模型做情感分析，需要大量数据，想请问老师，多少数据算大量？是指标记的部分需要大量数据吗?

> **A:** XXX（需要请王老师再帮忙完善一下回答）
> 训练模型的时候需要的是标注的，多少数据算大量有点主观，在一些情况下，几千条精心标记的样本就足够了，而在其他情况下，可能需要数十万甚至数百万条样本。评估你的具体情况和可用资源，以确定所需数据量。（回答人：刘晓飞；整理人：XXX)

Q14.几个英文词典中关于emoji表情的情感分析是否可以应用到中文呢？如果用的话，是否需要对整个英文都进行翻译吗?

> **A:** XXX（回答人：王老师；整理人：XXX)

Q15.老师之后讲课可以介绍一下sentence-BERT 嵌入模型吗

> **A:** XXX（回答人：王老师；整理人：XXX)

Q16.情感分析中Valence怎么翻译成中文?

> **A:** XXX（回答人：王老师；整理人：XXX)

Q17.王老师，您可以推荐一些能有助于构建中国房地产投资者情绪指数的情感词典吗？我的研究方向是房地产经济学

> **A:** XXX（回答人：王老师；整理人：XXX)

Q18.根据商标Logo分析情感分析可以吗？如果可以有什么包推荐

> **A:** XXX（需要请王老师再帮忙完善一下回答）
> logo如果指的是图片可能深度学习CNN这些基于标注的数据更适合，一般来讲还是根据自己的任务基于CNN构建模型，直接通过一个包识别图片情绪，目前了解比较少；如果logo是文本的话可以试试王老师讲的。
> （回答人：刘晓飞；整理人：XXX）

Q19.情感分析不是根据词频统计进行的情绪计算吗？如果不考虑语句逻辑的话，否定词+积极词汇=中性情绪，否定词+消极词汇=更消极情绪，这样并不合理啊？应该如何处理呢？

> **A:** XXX（需要请王老师再帮忙完善一下回答）
> 个人认为这可能也是根据词频法来识别情绪的缺陷所在，更优的算法可能是深度学习LSTM\RNN\BERT等算法，会考虑到语句逻辑。
> （回答人：刘晓飞；整理人：XXX）

Q20.情感分析如果是在一个否定句中，如何界定情绪？比如否定句中的积极词汇，但其本意是消极的。又如："这部手机没有很差"，单纯词频统计，显然是消极评价，但实际上是中性或偏积极的评价。

> **A:** XXX（需要请王老师再帮忙完善一下回答）
> 文本情感分析不只是根据情绪词，否定词也会影响文本的情感倾向
> 有专门的否定情感词词典，否定词通常在情感词的左边，它改变了情感极性，但偶数个否定词并不改变。
> （回答人：刘晓飞、李增杰；整理人：高文歆）

Q21.假如负向情感词典中有"真不"，正向情感词典中有"不错"，那么情感分析是否会认为这里有负向词和正向词各一，因此最终导致研究者将之判定为两相抵消的中性?如何准确地识别"真不错"，并将之判断为正向呢?

> **A:** XXX（回答人：王老师；整理人：XXX）

助教回答的问题

Q1. 老师的PPT和代码会分享吗?

> **A:** 课程资料里面包含老师的PPT和代码，请注意邮件，及时下载课程资料（回答人：张家星；整理人：张家星）

Q2. 请问老师现在文本分析的外刊论文发表可以用R和python实现吗？我的研究方向是财会方面的文本分析，因为我看很多这个领域的文本分析的论文都是用的stata，所以我有一点疑惑？谢谢老师！

> **A:** 当然可以用。目前主流文本分析主要还是用R和python实现的。stata可以进行简单的文本分析，但是受限比较多，建议还是使用R和python。不过，无论采用哪个软件，仅是软件上的差别，思想内核是一致的（回答人：张家星；整理人：张家星）

Q3. ppt的第16页，是不是加号应该改成减号

> **A:** 是的，王老师上课有提及，应改为：
> vector(“KING”) - vector(“QUEEN”) = vector(“MAN”) - vector(“WOMAN”)（回答人：张紫艺；整理人：张家星）

Q4. 如果向量空间模型，维度1w+，是否可以选取特定的1000个维度构造向量，这个操作可行吗？常见吗?

> **A:** 操作可行，也比较常见。降维能提高计算效率和可解释性，但是也会造成一定程度的信息损失，维度选择需要根据具体问题具体分析（回答人：刘晓飞；整理人：张家星）

Q5. 主题模型的label可以是“unknown”吗？（PPT 29页右表 topic 27）研究者对难以概括的主题可以有哪些技巧处理?

> **A:** 在LDA中，并没有预先定义的标签，相反，模型会学习到文档中常出现的单词组合，使用这些组合本身作为"主题"。如果你在训练后，想要给这些学习出的主题加上标签，这完全取决于您的具体应用。例如，如果一个主题包括单词 "cat", "dog" 和 "pet", 你可能想要把它标记为 "Animals"。如果一个主题不能被很好地解释，"unknown" 可以是一个可选的标签，但这并不是必须的。（回答人：姜昊；整理人：张家星）

Q6. 请问百度指数的数据如何下载？是付费吗？还是自己爬取数据?

> **A:** 百度指数应该是可以免费查询的，但如果想要批量下载数据，是需要自己爬取的，这可能会涉及到API调用,API调用一般每天是有限额的,想要突破限额可能需要付费（回答人：张家星；整理人：张家星）

Q7.R包qdap下载不了

> **A:** 可以将报错发到群中，助教们一起帮您看（回答人：张家星；整理人：张家星）

Q8.可读性方面，老师有中文复杂词汇的词表可供分享吗?

> **A:** 有的，参考《现代汉语常用词表》《现代汉语复杂词表》直接网络搜索下载（回答人：李瑶瑶；整理人：张家星）

Q9.请问老师，相关的文献以及讲解的文献主要偏经济类，可否增加一些管理类的文献呢？第二个问题是，在讲解的这些文献中，除了最后两篇文章给出了代码。是否可以提供更多的其它文献代码？因为对这篇文章不感兴趣。也或者说，可以允许将学员自己感兴趣的一篇文章，指导一下文章的作者是具体是如何获得数据到最后得到结果，让学员能够重现作者文章的分析过程呢？第三个问题，中文和英文在代码分析中，会不会有差距。比如，英文的文章代码跑通，是否可以适应英文，相反对应的是否可以相通?

> **A:** 【1】关于复现代码。①有些期刊是强制公开数据的，对于这些文章可以直接去期刊官网下载；②推荐一个可以查文献代码的网站：https://ejd.econ.mathematik.uni-ulm.de/ 通过检索限制“R"可以得到很多基于R写的代码，是一个很好的学习网站
> 【2】考虑到学科背景的多样化，老师尽可能地选择文本分析中比较重要且经典的文献进行讲解。如果您有感兴趣的文章可以发到群内，感兴趣的小伙伴可以一起探讨，这样也能在实践中更好学习老师所讲的理论知识。
> 【3】中英文是存在一点差距的，比如在分词等，但方法的内核是一致的。具体可以参考我们课程资料里”2. 文本预处理“提供的中文与英文的代码（回答人：张家星；整理人：张家星）

Q10.老师，请问对于专利进行文本分析有什么好的英文或中文文献吗？以及对于这种比较大的数据集，如何处理会更高效呢？老师有什么建议吗？

> **A:** 请参考Mann K.,Püttmann L.,2021,Benign Effects of Automation:New Evidence from Patent Texts [J],Review of Economics and Statistics,doi:https://doi.org/10.1162/rest_a_01083. （回答人：姜昊；整理人：张家星）

Q11. 文献中按报社取评论，按句子取平均等取均值操作的目的是什么？

> **A:** 以PPT24页中“刻画经济政策不确定性”为例，在“按月份对十种报纸进行平均”这一步中，其实比较重要的是将根据10种主流报纸计算的Y汇总得到Z，这一步平均的含义在于可以直观的比较各报纸和平均值之间的差异。在下一步中，归一化是为了使数据被限定在一定的范围内,从而消除奇异样本数据导致的不良影响。（回答人：张紫艺；整理人：张家星）

Q12.老师能否适当加一些对文本数据清洗流程的概括性描述，比如：第一步如何，应该用到哪些包，这些包主要功能是什么；第二部如何.....诸如此类，

> **A:** 关于文本数据清洗的概述和具体流程，可以参考课前发的学习资料《2022R语言初级》，里面有数据处理，文本数据清洗的内容。如果没有找到这份资料可以联系【王建秀】或【李晓燕】老师（回答人：张家星；整理人：张家星）

Q13. 请问如果评论内容为一些如“666”这样的内容，或者一些表情包，该怎样分析评论的情感呢?

> **A:** 对于网络词语，可以自建新的正负情感词表，合并到原有词表使用。
> 目前有些关于表情包emoji等情感分析的文章，用的是图像分析的方法。
> 比如DOI:10.2139/ssrn.4110191，GIF Sentiment and Stock Returns（回答人：张紫艺；整理人：张家星）

Q14. 老师请教一下做文献综述LDA，是用摘要去做还是用文章全文去做

> **A:** 这个可以都做一下，作为文章的稳健性分析，也是一个不错的思路。（回答人：姜昊；整理人：张家星）

Q15. 老师在情感分析中，一个月有好多条评论，这个可以变成月度情感吗？（因为想做面板数据），如果变成月度情感的话，是只有正向，负向这样。还是说正向情感占比是个数值

> **A:** 可以用每个月的评论做，构造月度面板数据，情感分析结果代表这个文档的情感色彩是正向还是负向，结果是一个数值。（回答人：姜昊；整理人：张家星）

Q16. 想请教一个，LDA后的主题，一个文本还能算与其主题之间的相似度吗?

> **A:** 是的，你可以计算lda主题模型的主题之间的相似度。一种常用的方法是使用余弦相似性。在LDA模型中，每个主题可以被看作是一个在词汇空间中的向量，其中每个分量表示一个特定词语的重要性或权重。由于这个原因，你可以使用诸如余弦相似

> 性这样的标准度量来计算两个向量（也就是两个主题）之间的相似性。（回答人：姜昊；整理人：张家星）

Q17. 想请教老师如果两段话求文本相似度，但是这两段话里有些词语是一样的，但是这词语在哪段话里的语意是完全不一样。但是求相似度的时候就算上去了，想问一下如何解决这种问题?

> **A:** 后续课程会讲解如何联合上下文语意计算相似度：Word Embeddings: 使用词嵌入模型（例如Word2Vec或GloVe）可以帮助解析词语的上下文含义。这种方法可以捕捉到词语在特定语境中的含义，并将词语映射到多维空间，这样，语义相似的词语会在这个空间中更接近。BERT：BERT是一个预训练的深度学习模型，为词语的语义理解带来了全新的方法。它可以将词语放入上下文中理解，捕捉到其深层次的含义。用BERT生成的嵌入向量可以用来计算相似性，这可能会给出更符合实际的结果。（回答人：姜昊；整理人：张家星）

Q18. 老师您在讲UGC部分的时候，识别有用的内容，将一小部分句子标记为：含有用信息或者不含有用信息。（这是指自己去打标签吗？）。然后再用标记训练卷积神经网络进行批量运算?

> **A:** 是的，人为处理，打标。（回答人：张紫艺；整理人：张家星）

Q19.请问无监督学习里提到的聚类分析、主成分分析等无监督方法和传统的统计学中提到的这两类方法有什么区别

A: 没有本质区别，只是因为这两种方法在训练模型的时候，训练集不需要标签而被归到了机器学习中的无监督学习，方法本质和传统统计学中是一样。

> （回答人：李增杰；整理人：高文歆）

Q20.在中文文本分析中，仍然会有英文文本（这些英文需要翻译成中文吗）。有没有批量处理语言翻译的API

> **A:** 这个问题需要结合具体的应用场景来考虑，如果不需要中文文本中的英文单词，那么可以不做处理；如果仅仅考虑出现在中文文本中英文单词的意思，则可以在预处理阶段进行翻译；如果研究需要考虑出现在中文文本中的特殊英文单词，那么可以结合正则表达式对英文文本进行识别，或者考虑扩充用于分析的词库。百度翻译api、有道翻译api都可以考虑使用。
> 这个网址python可以批量处理英文转中文https://blog.csdn.net/qq_45339371/article/details/119537301
> （回答人：李增杰、刘晓飞；整理人：高文歆）

Q21. 老师想请教一下，自己建立词库有没有相对便捷的方法

> **A:** 建立自己的词库是为了防止通用词典不包括部分行业专业词汇，造成分词不准确。自己在建立词库时最好可以借鉴相关研究，这样不仅节省时间，而且更有说服力。此外，搜狗输入法中也提供分话题的中文词库，也可以借鉴。
> 除此之外，jieba分词可拆分词性。另外可以运用文本挖掘工具（TF-IDF）、爬虫检索相关词汇；在有大量标注数据的情况下，也可以运用机器学习模型来识别分类。
> （回答人：闫钊鹏、高文歆；整理人：高文歆）

Q22. 老师期刊editor会来质疑分词的准确性吗？因为这个直接影响了文本之间的相似度

> **A:** 根据我个人投稿经验，如果使用常见的jieba，Glove等软件包不会被质疑，除非特别明显的词汇被分开，这个主要涉及到自己的词典问题
> （回答人：刘晓飞；整理人：高文歆）

Q23. Excel表里面的内容可以读取画词云图吗？？

> **A:** 只要excel表里面的内容包含词和对应的词频就可以画词云图
> 另外还可以用R【readxl】去读取Excel数据，jieba库分词处理，统计词频，然后用R的【wordcloud2】生成词云图
> （回答人：李增杰、高文歆；整理人：高文歆）

Q24. 文本分析中词典选取的标准

> **A:** 1）任务相关性：词典应该与你的分析任务高度相关。例如，如果你正在进行情感分析，你应该选择一个包含情感词汇的词典。2）领域特定性：如果你的文本数据来自于特定领域（如医疗、法律或技术），使用专门针对该领域的词典会更加有效。3）语言兼容性：确保词典与你分析的文本的语言相匹配。对于非英语文本，可能需要找到或创建适合该特定语言的词典。4）覆盖范围和深度：一个好的词典应该涵盖广泛的词汇，并且在所涉及的主题或概念上具有足够的深度。5）词典的更新和维护情况：语言是不断发展的，特别是在一些快速发展的领域。选择一个定期更新和维护的词典很重要。6）词典的来源和可靠性：使用来自可信来源的词典，特别是那些经过学术验证或在行业内广泛使用的词典。7）词性标记：对于某些分析任务，选择一个包含词性标记的词典可能很有用，这样可以更准确地理解和分析文本。8）开放许可和可访问性：根据你的项目需求，可能还需要考虑词典的许可和可访问性，特别是如果你的项目需要公开发布或商业使用。（回答人：刘晓飞；整理人：高文歆）

Q25.请问如果是多个文档的tf-idf值计算，是需要逐个文档计算还是有可以同时计算多个文档的代码?

> **A:** 在R或Python中通常可以使用现有的库或工具来进行批量计算，例如在Python中可以利用scikit-learn当中的TfidfVectorizer函数来进行计算
> 是需要多个文档一起计算的，因为在计算某一特定词语的IDF时，是需要由总文件数目除以包含该词语的文件的数目得到，所以应考虑全部文档集，得到所有词的TF-IDF值
> （回答人：马英杰、李增杰；整理人：高文歆）

Q26.老师我看到UTD中MIS Q中，文章将用户评论具体分成了Joy，Love，Anticipation，surprise，Anger，Anxiety，Disgust，Sadness这8个类别。想请教一下？因为课上只说了positive and negative这两类

> **A:** 情感分类有很多种，二分类、三分类、六分类、八分类等主要取决于你的数据标注情况，本次课程主要集中在二分类积极、消极
> （回答人：刘晓飞；整理人：高文歆）

Q27.有些文字会用Latent Emotion Topics这是怎么操作?

> **A:** 假设包含文本的向量，然后创建DTM，应用LDA模型并对情感词汇进行标注，进行主题分布和情感倾向分析，可以用R【topicmodels】尝试实现
> （回答人：高文歆；整理人：高文歆）

Q28. 为什么中文情感分析词典法要把三个词典结合起来呢?

> **A:** 这是为了保证词典中的情感词能够匹配文本数据中更多的内容，合并之后unique一下可以去重
> （回答人：李增杰；整理人：高文歆）

Q29.直接分析现成的文本数据，比如一本小说，一本报告？而不是表格格式，怎么操作？必须结构化吗?

> **A:** 对于非结构化文本，提取词干后可以转换成DTM实现。此外，NLP不需要转化结构化数据，是直接对原始文本进行分析的。
> （回答人：高文歆；整理人：高文歆）

Q30.决策树的生长中，每次划分的目标都是使得同一个分支里的样本有尽可能一致的类别标签。此处的类别标签是什么？课程中似乎没有提及。是研究者自行分类吗?

> **A:** 类别标签主要是数据集中的协变量
> （回答人：马英杰；整理人：高文歆）