文本分析：从文本到论文-王菲菲 Day2 问答（2023/11/18）

> 全文由王菲菲老师解答的问题、助教解答的问题两部分共同构成。

请王老师解答的问题

Q1：老师在讲用户流失预警的时候，研究框架那边想请教一下主题概率向量是怎么算的?

> **A:** 主题概率向量不需要计算，使用主题模型后计算机可以直接得出每个文档在主题上的概率值γ。
> (回答：王菲菲　整理：陈韵竹)

Q2：情感得分，后续要做Panel Data时，如何变成月度数据

> **A:** 如果分析文本带时间戳，需要对每一个文本计算情感得分，月度数据的话对每一个月的文本进行加权取平均即可。
> (回答：王菲菲　整理：陈韵竹)

Q3：请问主题12“教学内容反馈”的主题词语加总概率如果大于1是如何处理的?

> **A:** 、“教学内容反馈”是老师自己设置的主题，不属于真正意义上的主题，所以总概率大于1也没关系，且一般情况下也不会大于1。
> (回答：王菲菲　整理：陈韵竹)

Q4：ppt54页，第12个主题，是单独计算出来的吗？还是原来的文本，重新定义K=12，重新LDA建模得出的?

> **A：** 是单独计算出来的，是老师自己定义了一个主题利用主题模型计算得出。这里的主题并不是严格意义上的主题，是对一堆词语进行了一个总结归纳，只是为了去处理reviewer的建议进行的处理。
> (回答：王菲菲 整理：陈韵竹)

Q5：请问用户流失案例中模型3的回归结果为什么没有汇报主题4、主题10没有了？是因为不显著去掉了吗？还有立方项的含义是什么?

> **A：** 利用BIC做了变量选择，发现不显著所以去掉了。立方项是reviewer给的建议，没有具体含义。
> (回答：王菲菲 整理：陈韵竹)

Q6：有很多文档，想精准到确定每句话都主题，可以用LDA吗，是先把每个句子存成一个文档吗？是否有其他更合适的模型?

> **A：** 可以用LDA，LDA算出来的是每个文档在每个主题上的概率，如果想要知道某句话在每个主题上的分配概率那就需要把每句话当成一个文档来处理。
> (回答：王菲菲 整理：陈韵竹)

Q7：票房预测的项目中，遇到了什么内生性问题?

> **A：** 票房预测中有一些评论相关变量，此外还有前一天的票房变量，前一天的票房这个变量和后一天的票房评论相关，因此存在内生性问题。
> (回答：王菲菲 整理：陈韵竹)

Q8：主题模型是否必须要多份文本，如果是只在一份年报或者是一个企业连续几年的年报里运行，可行么?

> A：运行主题模型必须要有多个文档，如果只在一份年报或者一个企业连续几年的年报里运行也是可以的，前提是要把一份年报或者一个企业连续几年的年报分成多个文档，比如将里面的每段话或者每句话当成一个文档来处理。
> (回答：王菲菲　整理：陈韵竹)

Q9：老师在讲动态主题模型幻灯片70页（动态主题模型）时的推导是否能提供一下?

> **A：** 没办法提供，可以去看原始的文章，后面会发群里。在R里面没法跑，在python里面可以跑(回答：王菲菲 整理：江蓉)

Q10：作者主题模型、动态主题模型、有监督的主题模型这些有命令示范么?

> **A：** 这些在R里面都跑不了，都得用Python, 大家可以自己去搜(回答：王菲菲 整理：江蓉)

Q11：想老师讲解一下主题模型与人工编码混合方法，Academy of management Annnals在2019年Hannigan写到主题模型分析流程，第一步是语料库呈现（选择，修建），第二步是主题呈现（算法应用，适配），第三步是理论产物呈现（创造，构建）。有点不是很明白，想老师讲解一下

> **A：** 这个问题有点抽象。第一步的意思是你要选择什么样的文档去建立主题模型，还有就是在建主题模型的时候自己是可以定义和调整词典，这个问题可能更多的是说你的词典到底是什么样的；第二步主要是说你生成出来的主题是什么，是否有可解释性；第三步，如果是我做主题模型的话，就是我做完主题模型，又加了哪些计量模型，或者是这些主题和管理理论对应的关系等。(回答：王菲菲 整理：江蓉)

Q12：课上神经网络的幻灯片没有出现在邮件课件内,请问老师会提供么?

> **A：** 会提供，已经发送在群里(回答：王菲菲 整理：江蓉)

Q13：激活函数怎么选择?

> **A：** 如果是在输出层，激活函数就是softmax;如果是在中间层，激活函数就是ReLU,这是用的最多的。(回答：王菲菲 整理：江蓉)

Q14：分词中若只用停用词，是不是就把句子中的停用词去掉？并按停用词把句子拆开?

> **A：** 不是特别理解这个问题(回答：王菲菲 整理：江蓉)

Q15：词怎么变成词向量的？能否举个例子？比如BERT的Python代码中每个词是怎么对应一个数字的?

> **A：** 已经用代码展示举例，需要保存output结果，研究向量结构，把向量提取出来。(回答：王菲菲 整理：江蓉)

Q16：为什么说前馈神经网络输出结果可能是文本所属的类别（这个类别是怎么来确定？）

> **A：** 神经网络是一个有监督学习，要看前馈神经网络的Y是什么。基于文本要做一个正向或负向的情感判定时，Y就是文本所属的类别。如果不做情感判断，只是做前馈神经网络时，Y对应的就是前面给定的文本的下一个词。(回答：王菲菲 整理：江蓉)

Q17：上周的情感分析代码，随机森林没有给出绘制AUC图形的代码（其他几种方法都给出了），可否请老师分享一下

> A： 之后确认一下会做回复（回答：王菲菲 整理：江蓉）

Q18：最近在学习清华大学的大数据和因果推断研讨会的系列课程。其中，老师说R最开始就是一堆统计学家开发的，现在顶级的统计学家这些人也在用，所以R比Python 来说更适合来数据科学。今天王老师提出了相反的观点，让多学习一下Python ，请问如何回应这两种观点?

> A： 助教可以给大家分享一个博客，不代表王老师的意见。链接：https://segmentfault.com/a/1190000021653567
> (回答：陈俊良)

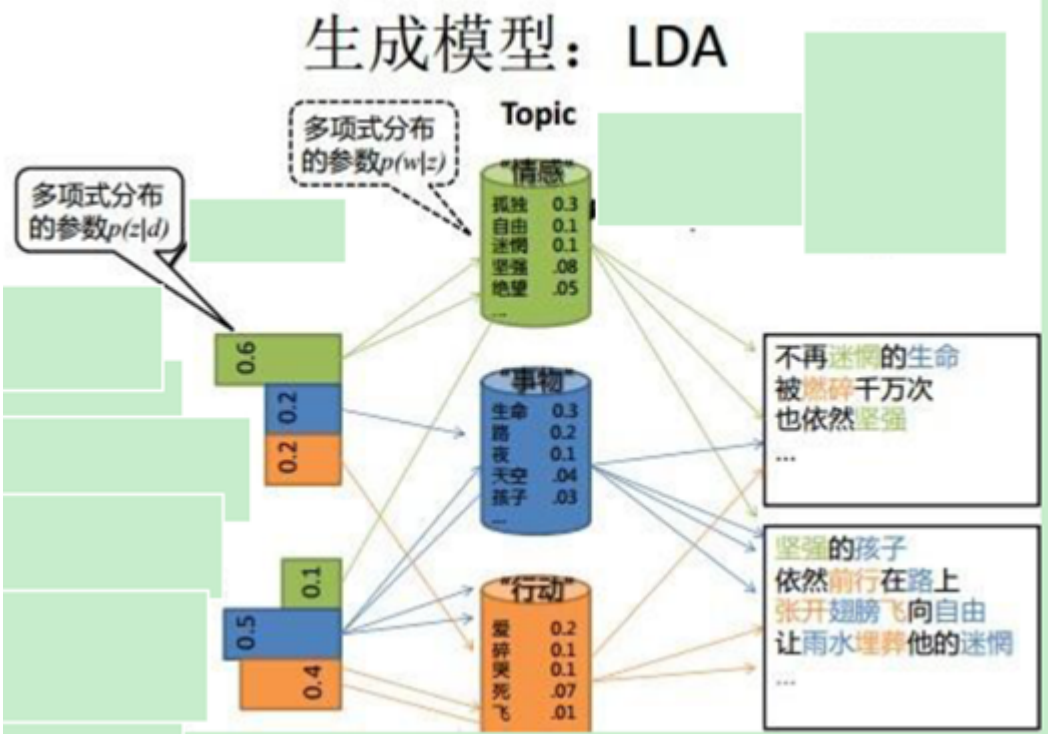
Q19：老师上课的时候说她最开始用Python 来做文本分析，很多时候用Python 也更方便。为什么此次课程宣传的时候，却以R为重点?

> A： R和Python都是在文本分析和文本挖掘领域广泛使用的编程语言。Python作为一种全方位的编程语音，在机器学习和深度学习中有更强大的库支持，使我们可以进行更复杂的文本分析。但是R作为专业的统计分析软件，在统计分析领域拥有更大的用户基数。对于大部分文本分析方法，R都可以提供丰富的库和函数。此外，针对经管专业的学员，R比Python的用户基数更大，R中有大量的针对经管领域的包。当然，如果有时间和精力，学习Python可以帮助我们进行更加复杂和深度的文本分析。
> (回答：孟佳音；整理：黄思佳)

Q20：假设对于一个要分析的项目，设定了主题数K=10。有两个问题：

（1）这个时候并不知道具体10个主题是什么，那模型内部用的是什么方法将项目涉及到的词分为10个主题的，是根据词向量在空间中的距离吗？还是其他什么机制？换句话说，K=10的情况下，词语是怎么自动聚集成10个队伍的？

（2）在同一个项目中，topic中词语出现的频率，对每个文档都一样吗？如下图，最右边的两个文档“不要迷惘的生命”和“坚强的孩子”，都适用同样的topic词语概率吗？



> **A:** （1）每一个主题下面都包含的是同样的词语，只是每个主题下面每个单词的概率不一样而已。我们在归纳某一主题具体是什么内容是，通常都是将计算频率最高的词语（比如列出该主题下频率最高的20个词语）列出来进行主题归纳的。
> （2）参见（1），topic中词语出现的频率，对每个文档都一样。
> (回答：陈韵竹；整理：黄思佳)

Q21：老师说用FastKNN和FastText与传统情感分析做对比，可以展开说一下，是说把正向词和负向词进行完善?

> **A:** 不是做对比，FastKNN和FastText都是分类模型，如果在做文本分析的时候用的是机器学习方法那些传统的分类器，比如支持向量机，决策树，随机森林等，这些是可以和深度学习方法做对比，这与词典法是不一样的。
> (回答：王菲菲)

Q22：长期记忆模型中的状态量怎么理解?

> **A:** 1.隐藏状态（Hidden State）：
> a.定义：隐藏状态是LSTM在时间序列的每个步骤中传递的一种信息形式。它包含了当前时间步骤的信息，以及之前时间步骤的相关信息。
> b.作用：隐藏状态可以被看作是LSTM的短期记忆部分。它被用来做出当前步骤的预测，并且随着每一个时间步骤被更新。
> 2.单元状态（Cell State）：
> a.定义：单元状态是LSTM的核心，是一种在整个时间序列中传递的长期信息。它的设计使得LSTM能够更好地处理长期依赖问题。
> b.作用：单元状态可以被看作是LSTM的长期记忆部分。它通过网络的特殊结构（遗忘门、输入门、输出门）进行维护，允许信息在需要时被保留，并在不再需要时被遗忘。
> 这两种状态量的互动使得LSTM能够在序列数据处理中更有效地学习和保留长期和短期的信息。例如，在文本处理中，隐藏状态可能捕捉到当前句子的语境，而单元状态则可能保存更长期的信息，如文章的主题或整体情感。
> (回答：吴浩然；整理：黄思佳)

助教回答的问题

Q1: 幻灯片第五页的文本示例文本很短,请问在做主题分析时文本数据是否可以使用上万字文本么?主题模型处理文本是否存在最适用(优化)文本词汇数?

Q2: 主题模型对每个文档的字数有什么限制么？文档字数最短可以是多少？有什么标准没?

> **A:** 进行主题分析时，使用上万字的文本是可行的 （可能会增加模型处理时间）。主题模型并没有严格的“最优”词汇数量。但是，模型的性能会受到文本中独特词汇量的影响。太少的词汇可能无法充分表达文本的多样性，而太多的词汇可能包含许多不相关或噪声词汇，这会影响模型的准确性和稳定性。
> （回答： 吴浩然；整理：黄思佳）

> **A:** 在理论上对文档的字数没有硬性限制。然而，在实际应用中，文档的长度会对模型的效果产生影响。如果文档的字数太少，它可能不包含足够的信息来准确地推断出主题分布。如果文档过长，它可能包含多个不同的主题，这可能使得模型难以区分和准确地分配主题权重。在实践中，理想的文档长度应该足够长，以包含足够的词汇来表示一个或多个主题，但又不至于太长以致混淆主题。关于文档的最短长度，并没有一个普遍接受的标准。它很大程度上取决于具体的应用场景和文档集的特性。在一些情况下，即使是短文本（如推文）也可以用于主题建模，尽管这可能需要特别的处理方法和调整。
> (回答：孟佳音；整理：黄思佳)

Q3: LDA的最佳主题个数除了困惑度和主题一致性之外还有没有其他的方法？如果是在论文中审稿人会对主题个数的选取提出质疑么？我们如何去回应质疑?

> **A:** 最常见的指标就是困惑度和主题一致性，一般而言，如果在文章中详细地说明了K值的确定方法和过程，审稿人大概率不会对这两种方法产生质疑。如果被质疑了或者担心被质疑，可以同时采用困惑度和主题一致性这两种指标进行稳健性检验，还可以使用交叉验证方法，选择最优的K值。
> (回答：黄思佳；整理：黄思佳)

Q4:LDA中，千面骰子的总词数是怎么确定的？是像词典一样，提前预设好的吗

> **A：** "骰子有多少面"指的是词汇表的大小。词汇表的大小是提前确定的，基于预处理后的文档集。在创建词汇表之前，文档通常会经过一系列预处理步骤，包括去除停用词、数字、标点符号，以及进行词干提取或词形还原等。这些步骤会显著影响词汇表的最终大小。有时，为了控制词汇表的大小，会设置最小和最大频率阈值。例如，可以排除在文档集中只出现一次的词汇（低频词），或者在几乎所有文档中都出现的词汇（高频词，如某些常见的停用词）。根据应用的领域和目标，可能会选择包含或排除特定的词汇。例如，在某些技术或科学领域，专业术语可能非常重要，即使它们的出现频率不高。
> (回答：孟佳音；整理：黄思佳)

Q5:计算困惑度时，常常遇到困惑都持续波动下降，一直见不到最低点。这时应该如何处理?

> **A：** 数据集不均衡：如果数据集中某个类别的样本数量远大于其他类别，可能会导致困惑度波动下降。这种情况下，可以尝试对数据集进行均衡化处理，例如采用随机抽样或过采样等方法来平衡不同类别的样本数量。
> 模型选择不当：如果使用的模型不适合当前任务，可能会导致困惑度波动下降。这时可以尝试更换不同的模型，例如使用更复杂的模型或调整模型的超参数来提高性能。
> 训练不充分：如果训练时间不够长或训练数据不足，可能会导致困惑度波动下降。可以尝试增加训练时间和数据量，或者使用更复杂的模型来提高性能。
> 过拟合：如果模型过度拟合了训练数据，可能会导致困惑度波动下降。可以尝试使用正则化、增加数据量、减少模型复杂度等方法来减轻过拟合问题。
> (回答：罗银燕；整理：黄思佳)

Q6：比较可靠的计算困惑度的工具是什么，有人说 gensim 的困惑度计算功能有错误。

> **A：** 如果gensim的困惑度计算功能确实存在问题，可以考虑使用其他的自然语言处理工具，例如NLTK、spaCy、TextBlob等，它们也提供了困惑度的计算功能。
> (回答：罗银燕；整理：黄思佳)

Q7：假设我已经使用 LDA 将10000篇文章分为 20 类，此时又来了 500篇文章，我如何将这新增的 500 篇重新分类到已有的 20 类中。

> **A：** 首先，对这500篇新文章进行与之前相同的预处理，包括词汇的清洗、标准化、分词等。其次，使用训练原始LDA模型时创建的词典（Dictionary）将新文章转换为词袋（bag-of-words）格式。将这个词袋模型的新文章输入到已有的LDA模型中，以获取该文章的主题分布。
> 不过，在处理新增文章的时候，如果新的500篇文章在内容或风格上与原有10000篇文章显著不同，可能需要重新训练模型，如果新增文章引入了完全不同的主题，可能需要增加主题数量。
> (回答：黄思佳；整理：黄思佳)

Q8：使用 LDA 分类时，如果不控制随机数种子，那么每次重新运算的结果差异很大，如何解决这种稳健性问题。此外，Seed值会影响主题内容和高频词么?

> **A：** 建议在初始化LDA模型时设置一个固定的随机数种子。这可以确保每次模型运行的初始化状态一致，从而保证结果的可重复性。是的seed值的设定会影响主题的内容和主题内高频词的选择。为了获得稳定和可重复的结果，一个常见的做法是设置一个固定的随机数种子。这样，每次运行模型时都会以相同的方式初始化并进行迭代，从而确保结果的一致性。如果模型对随机数种子非常敏感，这通常表明数据集可能需要更精细的预处理，或者模型的参数需要进一步调整。
> (回答：孟佳音；整理：黄思佳)

Q9：基于问题3主题数质疑,现实操作过程中我们是否要手动尝试K?有什么标准可以证明我们找到了理想的K值?

> **A：** 如果数据量不是很大的话，可以尝试手动判断K值，人为查看每个主题的代表性词汇和它们的可解释性。但是当数据量特别大的时候，人为判断就不现实了。确定最佳主题数K常用的方法是困惑度和主题一致性，一般而言，困惑度越低，或者主题一致性得分越高，模型的性能越好。
> (回答：黄思佳；整理：黄思佳)

Q10：Gibbs抽样是否是说X和Y因果循环?

> **A：** 不是。Gibbs抽样是为了从联合概率分布中高效地采样，变量的更新是基于它们之间的条件依赖关系，而不是因果关系。在Gibbs抽样中，按照一定的顺序逐个更新每个变量，每次更新时都基于其他变量的当前值。这种更新是基于条件概率，反映了变量间的统计关联，而不是因果关系。
> (回答：孟佳音；整理：黄思佳)

Q11：最开始的时候，怎么知道文档中，所有的词分别属于哪个主题?

> **A：** LDA是一种无监督学习，模型用于从文档集合中发现主题。在开始时，模型并没有关于词与主题关系的先验知识，并不知道文档中的所有词分别属于哪个主题。经过迭代逐步发现较为稳定的分配。
> (回答：孟佳音；整理：黄思佳)

Q12：LDA中估计不收敛怎么办？是什么原因导致的?

> **A：** 原因：数据集太小、主题数量设置不合适、迭代次数太少、主题之间相似度过高、模型的参数不合适
> 可以针对上述问题进行调整，解决不收敛的问题
> (回答：罗银燕；整理：黄思佳)

Q13：LDA中topic的数量是自己选的吗？PPT中选了3个（情感，事物，行动），是不是也可以多选出几个?

> **A：** LDA中最佳主题数量K常用困惑度和主题一致性来确定的，K值不是固定的，PPT中K=3只是举一个例子，实际研究中不同的数据不同的情境下，K值可能都不一样。
> (回答：黄思佳；整理：黄思佳)

Q14：LDA和Citespace有什么区别，更推荐使用哪个?

> **A:**
> LDA在文献综述中的应用主要是通过对文献的词频和共现关系进行分析，挖掘出文献中隐含的主题信息，从而对文献进行分类和归纳。相比之下，Citespace在文献综述中的应用主要是通过可视化的方式将文献之间的引用关系、共线关系等展现出来，帮助研究者更好地理解文献之间的联系和演变。
> 因此，选择哪个工具主要取决于研究者的需求。如果需要从大量文献中提取主题信息，并对其进行分类和归纳，则LDA可能更合适。如果需要更好地理解文献之间的联系和演变，则Citespace可能更合适。
> 总的来说，两个工具都有其独特的优点和适用范围，选择哪个工具取决于研究者的具体需求和偏好。
> (回答：罗银燕；整理：黄思佳)

Q15：在做综述使用LDA中，只知道归纳出的主题，不清楚这些主题来源于哪几篇文献。是否可以找出归纳出主题的文献

> **A:** LDA是一种无监督的学习方法，它并不直接提供关于主题来源的文献信息。
> 尝试找出某主题的文献的方法有：
> 1、根据主题的关键词手动搜索相关文献
> 2、尝试使用引文分析工具，例如Citespace或Google Scholar等，通过分析引文网络来找出与主题相关的文献。
> (回答：罗银燕；整理：黄思佳)

Q16：王老师是否能推荐几篇主题模型的经典文献和近年的优秀文献?

> **A:** 经典文献： Blei, David M. 《Latent Dirichlet Allocation》，
> 近年优秀的顶刊文献（会计学）：
> [1] Brown, N. C., R. M. Crowley和W. B. Elliott. What Are You Saying? Using Topic to Detect Financial Misreporting[J]. Journal of Accounting Research, 2020, 58(1):237-291.
> [2] Chychyla, R., A. J. Leone和M. Minutti-Meza. Complexity of Financial Reporting Standards and Accounting Expertise[J]. Journal of

Accounting and Economics, 2019, 67(1):226-253.
[3] Rawson, C., B. J. Twedt和J. C. Watkins. Managers' Strategic Use of Concurrent Disclosure: Evidence from 8-K Filings and Press Releases[J]. The Accounting Review, 2023, 98(4):345-371.
(回答：罗银燕；整理：黄思佳)

Q17：number of topics的Log-likelihood曲线，可以给一下代码吗?

A: 可参考https://cloud.tencent.com/developer/article/1436373
(回答：罗银燕；整理：黄思佳)

Q18：产品的文本介绍可以提取情感嘛？有意义吗?

A: 可以先人为判断一下产品介绍里面是否具有某种情感倾向，如果有，可以用文本分析方法进行分析。至于是否有意义，取决于你的研究问题，以及你能否从数据当中挖掘一些合适的指标来衡量变量，从而支撑你的研究问题。可以参考一下这篇文章， 作者利用文本分析方法来分析公司招聘广告中年龄歧视的情感倾向，及其与实际招聘中年龄歧视之间的关系。
Burn, Ian & Button, Patrick & Munguia Corella, Luis & Neumark, David. (2022). Does Ageist Language in Job Ads Predict Age Discrimination in Hiring?. Journal of Labor Economics. 40(3): 613–667.
(回答：黄思佳；整理：黄思佳)

Q19：entropy的中文是什么意思?

A: "Entropy" 在中文中的意思是"熵"。课程ppt中，Entropy衡量了截止到t时刻，所有reviews在K个主题上概率取值的离散程度，反映的是主题分布的多样性。当在所有K个主题上的讨论概率值都相等时，entropy最大，表明文档集中的主题分布是完全均匀的，每个主题都有同等的重要性。当在某个主题上概率取值为1，其他主题上概率取值为0时，entropy最小，表示文档集中在某个特定主题上非常集中，主题分布极不均匀。
(回答：黄思佳；整理：黄思佳)

Q20：请问老师中文算readability有现成的包吗?

> **A：** cntext也是一种现成的python包
> (回答：董如玉；整理：黄思佳)

Q21：超参数的设定依据是什么?

> **A：** 一般而言，超参数的设定要考虑数据特性、模型性能评估、计算资源和时间限制、以及具体研究问题等。
> 以LDA为例，主要的超参数包括主题数 K 以及两个Dirichlet分布的参数：文档-主题分布的先验参数 α 和主题-词分布的先验参数 β。主题数 K 常用困惑度和主题一致性来确定，α和 β 的确定可以通过调参，观察模型性能的变化来选择最佳值。可以通过网格搜索、随机搜索或贝叶斯优化等方法系统地寻找最优的超参数组合，使用交叉验证方法来评估不同超参数设置下的模型性能。
> (回答：黄思佳；整理：黄思佳)

Q22：有监督主题模型与动态主题模型的区别?

> **A：** 有监督主题模型和动态主题模型都是主题模型的变体，但它们的关注点和应用场景有所不同。
> 有监督主题模型，例如Labeled LDA，是一种主题模型，主要用于对有标签的文档进行建模。这种模型的特点是，它使用带标签的数据集训练算法，以达到准确分类数据或预测结果的目的。例如，在文章的Abstract后面会跟上Key Words，其中Key Words决定了这篇论文(或摘要)要讨论的话题，因此在建模时，可以使用Key Words来约束摘要的话题。
> 而动态主题模型，例如DTM（Dynamic Topic Model），则是一种主题模型，可以用于对文本数据进行建模和分析，同时考虑到时间序列的变化。动态主题模型引入了时间动态的概念，后一时刻的主题从前一时刻演化而来，可以很好的建模主题变化。这种模型的特点是，它能够显示主题在特定时间间隔（如每一年）内的变迁。
> 总的来说，有监督主题模型更关注如何利用已有的标签信息来提升主题模型的性能，而动态主题模型则更关注如何捕捉和建模主题随时间的变化。这两种模型各有其优点，适用于不同的应用场景。
> (回答：陈俊良；整理：黄思佳)

Q23：机器学习前标记的用途是什么?

> **A：** 最主要的目的是要让数据变得更加容易被接受和理解。
> 例如，在监督学习中，提供给算法的训练数据包含所需解决方案的标签。这些标签可以帮助算法更好地理解数据，并从中学习。例如，如果你想让一个机器学习算法学习图像分类，可以为每张图像添加标签，表示图像中的物体类型和各种属性。这样，算法就可以通过标签来学习图像中的物体是什么。
> (回答：陈俊良；整理：黄思佳)

Q24：为什么困惑度越低越好？困惑度是怎么测算的?

> **A：** 困惑度（Perplexity）是一种用来评价语言模型好坏的指标。在自然语言处理中，困惑度用来度量一个概率分布或概率模型预测样本的好坏程度。直观上理解，当我们给定一段非常标准的，高质量的，符合人类自然语言习惯的文档作为测试集时，模型生成这段文本的概率越高，就认为模型的困惑度越小，模型也就越好。
> 困惑度的计算方法与所使用的语言模型有关。对于N-gram模型（如uni-gram，bi-gram，tri-gram等），困惑度的计算通常涉及到句子概率的计算，这通常需要对句子进行分解，然后计算各个部分的概率。对于神经网络模型（如RNN，LSTM，GRU等），困惑度的计算则可以直接使用句子的概率。
> 总的来说，困惑度越低，说明模型对句子的预测能力越强，语言模型的效果越好。因此，在文本分析中，我们通常希望困惑度尽可能地低。这意味着我们的模型能够更准确地预测接下来的词语，从而更好地理解和生成文本。
>
> 具体计算公式： $$perplexity(S) = p(w_1, w_2, w_3, ..., w_m)^{-1/m} = \sqrt[m]{\prod_{i=1}^{m} \frac{1}{p(w_i|w_1, w_2, ..., w_{i-1})}}$$
>
> (回答：陈俊良；整理：黄思佳)

Q25：做LDA可以用一张Excel表里不同列（变量去做）？因为内容全部都在一张表中，这样不是没有多个文档了?

> **A：** 可以用python的pandas包提取excel中的文本进行分析，照样可以做LDA。
> (回答：黄思佳；整理：黄思佳)

Q26：课后想自学推导LDA模型,老师对统计小白有什么建议?

> **A：** 可以参考以下文章：
>
> 1. http://t.csdnimg.cn/mviRe
> 2. https://blog.csdn.net/qq_39422642/article/details/78730662?spm=1001.2014.3001.5506
> 3. http://t.csdnimg.cn/lKTbI
>
> (回答：黄思佳；整理：黄思佳)

Q27：请问在文本预处理的时候是否有别的更方便的方法？就是在上节课所介绍的删除停用词等之后，如果所提取出的文本仍然需要进一步分类，是否有除了人工阅读之外的更高效的方法?

> **A：** 文本预处理包括文本清洗、分词、去除停用词、词干提取或词形还原、去除低频词等。若在文本预处理后需要对文本进行进一步的分类或分析，还以机器学习、主题建模、聚类分析、关键词提取等方法对文本进行进一步的分析。
> (回答：孟佳音；整理：黄思佳)

Q28：在讲Skip- gram模型 $f = \sum_{j=0, j\neq m}^{2m} H(\hat{y}, y_{c-m+j})$是什么意思?

> **A：** 其中，f 表示的是从中心词 c的视角，考虑前后 $2m$ 个词作为上下文，计算模型预测的概率分布$\hat{y}$与实际的分布 y_{c-m+j}之间的交叉熵损失，然后对所有这些损失求和。
> $H(\hat{y}, y)$ 是交叉熵函数，它衡量了两个概率分布之间的差异。在Skip-gram模型中，$\hat{y}$ 是模型预测的概率分布，而 y 是真实的概率分布，通常使用one-hot编码表示实际的上下文词。
> 模型的目标是最小化交叉熵损失，以便更准确地预测给定中心词的上下文。模型通过这样的训练学会产生能够捕捉语义和语法属性的词嵌入。
> (回答：黄思佳；整理：黄思佳)

Q29：分词分出来数量可以控制吗？觉得分出来词有点少

> **A：** 在进行文本分词时，控制分出来的词的数量是可以操作的，但这主要取决于你使用的分词工具和方法，以及你的文本数据本身。调整分词的数量可以考虑如下方法：调整分词算法设置；修改预处理的步骤（减少去停用词程度或者减少去除低频词的程度）；使用不同的分词工具；考虑N-gram的方法（可以产生更多的词汇组合）等。如果文本本身较短或者内容比较单一，自然会导致分出的词较少。在这种情况下，可能需要更多或更丰富的文本数据。对于关键文档，可以进行人工检查和微调分词结果，确保关键信息没有被遗漏。
> (回答：孟佳音；整理：黄思佳)

Q30： 请问批量化的为不同企业的年报文本，假设1000个企业做主题模型，请问具体的代码是什么？在老师提供的代码中应该如何改动?

> **A：** 1、仿照示例数据的格式，将您收集到的年报文本规范为那种格式。
> 2、根据您的需要修改参数设置
> 3、在实操的过程中不断调整，趋向更好的结果
> (回答：罗银燕；整理：黄思佳)

Q31：请问主题聚类背后的原理是什么，想请老师通俗的解释一下。以“用户流失预警”为例，为什么爱心、打扰、回复等一系列的词会归结到一个主题中，是因为这些词有什么关联么，比如，通俗地说这几个经常一块出现。

> **A:** LDA模型会把每段文本中的所有单词都看作是从某个主题中随机抽样得到的。这里的“主题”可以理解为一组单词的组合，这些单词在大量文本中经常一起出现。
> 那么，为什么有些单词会被分到同一个主题下呢？这是因为这些单词在很多段文本中经常一起出现。比如说，在很多关于用户流失预警的文本中，“爱心”、“打扰”、“回复”等单词经常会一起出现。LDA模型就会把这些单词归为一个主题。
> (回答：罗银燕；整理：黄思佳)

Q32：老师300多篇文章做LDA，老师您觉得用摘要做LDA可以吗?

> **A:** 可以，摘要是文章核心内容的提炼，想要了解哪些文本的主题主要看您的需求。
> (回答：罗银燕；整理：黄思佳)

Q33：投影层向量还是不理解，计算这个向量是用来做什么呢

> **A:** LDA的投影层向量用于将文本中的词或句子从高维空间映射到低维空间，从而简化数据的复杂性并提取出主要特征。
> 具体来说，LDA模型中的投影层向量可以用于降维和特征提取。通过将词或句子投影到低维空间，可以保留最重要的特征，同时去除噪声和冗余信息。这有助于提高模型的泛化能力和效率
> (回答：罗银燕；整理：黄思佳)

Q34：LDA在经管类研究中的应用方式有哪些？比如通过主题提取理论化，还看到有利用LDA来刻画一些概念如产品独特性，想请教下老师还有没有其他应用方式，可以用来刻画哪些类型的变量?

> **A：** LDA在经管中的经典应用包括王老师上课讲的内容集中在市场研究和消费者行为分析（利用LDA分析消费者评论、论坛讨论等，以识别消费者对产品或服务的主要关注点，揭示消费者需求和偏好；还可以通过对社交媒体数据的分析了解市场趋势和消费者情绪）。其他还可以应用在金融市场分析（刻画投资者情绪）；企业组织行为研究（利用LDA探索管理实践、创新策略和组织变革的主题）；政策分析和公共管理（分析政策文档、法规、公共讨论等，以提取政策焦点和公共关注点）等。
> LDA可以刻画的变量有很多，比如：情感和观点；概念和主题；市场细分（识别不同市场细分和消费者群体的特征）；风险和机遇等。
> （回答：孟佳音；整理：黄思佳）

Q35：词嵌入和词向量研究在金融领域的应用是什么样子?

> **A：** 在金融领域的应用十分丰富，可以简单举几个例子：利用情感分析来评估市场新闻或社交媒体中的情绪倾向以预测股价走势；在风险管理中，使用词嵌入技术来分析财务报告、法律文件或新闻文章，以识别可能的风险因素。银行和金融服务公司可以利用词嵌入技术分析客户的交易记录、查询记录等，以更好地理解客户需求，并提供个性化的金融产品推荐等。
> (回答：孟佳音；整理：黄思佳)

Q36：老师, BERT处理好的只有一个？还是说根据研究问题去选择与自己研究问题最接近的BERT训练好的包

> **A：** BERT有一些新的研究进展。可以根据研究问题选择与自己研究问题最接近的包
> 例如，FinBert是针对金融文本的BERT模型
> (回答：罗银燕；整理：黄思佳)

Q37：词怎么变成词向量？能否举个例子？比如BERT的Python代码中每个词对应一个数字

> **A：** 推荐阅读：https://www.zhihu.com/question/56121488
> (回答：罗银燕；整理：黄思佳)

Q38：老师您好，请问mac m1电脑装不上tensorflow怎么办？有没有替代的包或者替代方法？Bert模型必须要装tensorflow吗？感谢！

> **A：** 如果在Mac M1电脑上无法安装TensorFlow，可以尝试以下替代方法：
> 1、安装TensorFlow的替代包：可以选择安装其他与TensorFlow功能相似的包，如Keras或PyTorch。这些包在Mac M1电脑上通常能够轻松安装，并且可以用于构建和训练神经网络模型。
> 2、使用虚拟环境：可以尝试在Mac M1上使用虚拟环境来安装TensorFlow。通过虚拟环境，您可以为TensorFlow创建一个隔离的环境，避免与其他Python库的冲突。
> 3、使用ARM版本的TensorFlow：TensorFlow针对不同的硬件架构提供了不同版本的安装包。您可以尝试安装针对ARM架构的TensorFlow版本，这可能与Mac M1电脑兼容。
> BERT模型需要使用TensorFlow作为后端支持来构建和训练模型。但是，如果希望在Mac M1电脑上运行BERT模型，可以考虑使用预训练的BERT模型，这些模型通常已经针对特定任务进行了预训练，可以直接使用而不需要重新训练。
> 另外，也可以考虑使用其他的自然语言处理工具包，如Spacy、NLTK等，这些工具包提供了其他替代方案来处理自然语言任务，不需要依赖TensorFlow。
> (回答：罗银燕；整理：黄思佳)

Q39：老师说CNN用于图像处理，可以展开说一下图像处理的哪些方面（图片饱和度这些？）

> **A：** CNN（卷积神经网络）是一种特别适合处理图像数据的神经网络架构。在图像处理中，CNN可以应用于多种任务，包括但不限于：
> 物体定位：CNN可以用于预测图像中包含物体的位置，帮助识别系统找到需要识别的物体在图像中的位置。
> 物体识别：CNN可以用于识别图像中的物体，通过对图像进行分类或者对图像中的物体进行标记来识别物体。
> 目标分割：CNN可以用于将图像中的特定目标分割出来，例如将图像中的汽车、人等物体从背景中分离出来。

关键点检测：CNN可以用于检测图像中物体的关键点，例如人脸的眼部、鼻部、嘴巴等关键点，帮助计算机更好地理解图像中的内容。
图像复原：CNN可以用于图像复原，通过对图像进行去噪、修复等方式来提高图像的质量。
图像超分辨率：CNN可以用于提高图像的分辨率，将低分辨率的图像转换为高分辨率的图像。
风格迁移：CNN可以用于将一种风格的图像应用到另一种风格的图像上，例如将梵高的绘画风格应用到一张照片上。
图像压缩：CNN可以用于图像压缩，通过对图像进行编码和解码的方式来减少图像的存储空间。
除了以上这些任务，CNN还可以应用于图像的色彩空间转换、图像增强、图像金字塔生成等任务。总之，CNN在图像处理中的应用非常广泛，可以解决许多传统的图像处理方法难以解决的问题。
(回答：罗银燕；整理：黄思佳)