文本分析专题：从文本到论文

> 连享会　主页 || 视频 || 推文 || 知乎 || Bilibili 站

1. 课程概览

- **嘉宾**：王菲菲 (中国人民大学)
- **时间**：2023 年 11 月 11, 18, 25 日 (三个周六)
- **时段**：上午 9:00-12:00，下午 14:30-17:30 (17:30-18:00 答疑)。
- **课件/计量软件：** 提供全套 R 程序、数据和核心论文复现资料 (开课前一周发送)
- **方式**：在线直播 + 20 天回放
- **PDF 课纲**：https://file.lianxh.cn/KC/lianxh_Text.pdf (网页版)
- **参考文献和预读资料**：在线查看，(或) 打包下载
- **报名链接**：https://www.wenjuan.com/s/UZBZJvMpWH/#
- **助教招聘**：https://www.wjx.top/vm/PpL5PH7.aspx#

2. 嘉宾简介

王菲菲, 中国人民大学应用统计科学研究中心研究员, 中国人民大学统计学院副教授, 北京大学光华管理学院统计学博士。研究兴趣包括：文本挖掘及其商业应用、社交网络分析、大数据建模等，成果见诸 Journal of Econometrics, Journal of Business and Econometric Statistics, Journal of Machine Learning Research, 管理学报等。主持国家自然科学基金青年项目和面上项目各 1 项、全国统计科学研究重大项目 1 项。曾获中国人民大学优秀科研成果奖、课外优秀指导教师奖等。

3. 课程简介

3.1 课程导引

随着互联网技术的发展，新闻、网页、日志、博客等文本信息都出现了爆发式增长，对文本数据的分析需求也随之变得越来越迫切。文本挖掘，作为数据挖掘的重要组成部分，已经成为将信息转化为知识的不可或缺的工具，并且在经济、管理等领域有着越来越广泛的应用。文本分析在经管研究中的火爆程度可以从如下论文中窥豹一斑。

例如，在金融和会计领域，Loughran and McDonald (2011) 发表于 **Journal of Finance** 上的有关文本分析技术的综述性文章，短短十余年，Google 引用已 4800 余次。二人于 2016 年发表于 **Journal of Accounting Research** 的另一篇介绍文本分析在会计和金融领域应用的综述性文章目前已被引用 1700 余次。两位学者在近十年中基于文本分析方法发表的文章遍布 JFE，JF, JAR, JFQA 等顶刊，获得了广泛的关注。二者在 2020 年发表于 **Annual Review of Financial Economics** 的综述文章 Textual Analysis in Finance (-PDF-) 对相关文献和方法进行了系统梳理。在最近的研究中，García et al. (2023, JFE) 展现了文本情感分析的重要性。

- García, D., X. Hu, M. Rohrer, **2023**, The Colour of Finance Words, **Journal of Financial Economics**, 147 (3): 525-549. -Link-, -PDF-, Replication, -cited-

那么，文本信息有哪些特点？ 文本挖掘有哪些通用方法和套路？文本分析如何与你目前的研究内容相结合？ 这些恰恰是本次课程尝试帮各位解决的疑问。

我们将细致梳理文本挖掘在经济管理等领域中的应用场景和挑战，力求帮助大家熟悉并掌握文本挖掘的框架和体系，能够在实际场景中使用文本挖掘的各种方法，并对方法背后的原理有清晰、深入的理解。具体内容如下：

- **T1. 文本分析速览**：流程和场景
  - 文本挖掘标准流程
  - 在经管类 Top 期刊的应用
  - 文本预处理方法及 R 代码实现
- **T2. 情感分析**：探究客户的态度
  - 情感分析是什么
  - 常用情感词典
  - 基于机器学习的情感分析方法
  - R 代码实现
- **T3. 主题模型**：文本归类和内容提炼
  - 原理和结果解读
  - 应用实例 + R 代码实现
- **T4. 基于 AI 的文本分析方法**
  - 词嵌入方法
  - 深度学习模型
- **T5 / T6. Top 期刊论文精讲**：没你想象的那么难
  - 精讲两篇 Top 期刊论文，综合训练上述各类分析方法
  - 提供论文的 R 代码复现

3.2 课程特色

> **A. 顶天+立地**

此次课程以 TOP 期刊的论文为指引，通过案例教学的方式帮助学生掌握文本分析的思路、流程和常用方法。各个模型都辅以 R 代码讲解和复现，以便各位将文本分析与自身的研究兴趣相结合，将文本分析方法移植或嫁接到自己的研究中。

> **B. 方法覆盖面广，经典与前沿并重**

课程覆盖了文本分析的几个重要领域，比如「情感分析」、「主题模型」等。课程覆盖的方法既包括经典方法，如「向量空间模型」，「TF-IDF 编码」，「潜在狄利克雷分配模型」(Latent Dirichlet Allocation)等；又结合人工智能领域的新发展，介绍一些主流的 AI 算法和模型，如「词嵌入」(Word Embeddings)、「循环神经网络」(Recurrent Neural Network)，「长短期记忆模型」(Long Short Term Memory) 等。

3.3 开课前的准备

本课程实操部分均采用 R 代码实现。因此，你需要花点时间学一下 R 的基础知识。请相信我，学习 R 没你想象的那么困难，你只需要老老实实地对着 **R4DS** 操作一下就可以很快上手了。

> R 入门和基础

- **R4DS** Hadley W., and G. Grolemund. 2017/2023. **R for Data Science**. O'Reilly Media.
  - 2017 版： 主页 , 在线阅读 ， github ， 习题解答： V1 ， V2
  - 2023 版： R4D-2E-在线阅读
- **R2** Kabacoff, Robert I., **R in Action: Data analysis and graphics with R**, 2011. -PDF- . 备受好评的 R 语言学习书，可以通过该书开启你的 R 编程之路.

> R 文本分析

- Silge, J., D. Robinson. **Text mining with R: A tidy approach**[M]. O'Reilly Media, Inc., 2017. Reilly 在线阅读 , -主页阅读- , 中文版 , Github-Codes . R 文本分析经典书目
- **Text Analysis: R packages list**. University of Pennsylvania. 列出了各类用于文本分析的 R 包. -Link-

> 其他：

- **Roadmap** A roadmap for getting started with R , R 打怪升级指南
- rstudio.com - Finding Your Way To R ， 介绍了入门、中级和高级学习过程中的主要参考资料
- **BBR** Baruffa, Oscar. 2023. Big Book of R. -Link- . R 语言资料清单：包括各类书籍、包等资料的清单。
- 连享会： R 资料 ， Stata v.s R 对照表

4. 课程详情

温馨提示： 课程大纲中涉及的文献和资料，可以 <在线查看>，(或) <打包下载>

T1. 文本挖掘简介 (3 小时)

本节分为两部分。首先介绍文本数据的特点，并概括总结既往的经管类文献中是如何使用文本数据的，为听众进行后续研究打开思路；其次将介绍文本挖掘的方法论，并介绍中英文两种文本分析的预处理方法，包括分词、关键词提取、文本可视化展示等。具体的内容安排如下：

- 介绍文本数据的特点
- 介绍文本挖掘的方法论
- 介绍文本挖掘在经济、金融、营销领域中的应用
- 介绍中文数据的预处理方法
- 介绍英文数据的预处理方法
- 以影评和股评为例，介绍基于 R 代码的实现
- 主要参考文献：
    - 沈艳, 陈赟, 黄卓. (2019). 文本大数据分析在经济学和金融学中的应用：一个文献综述. 经济学 (季刊)(4), 34-51. -PDF-

T2. 文本情感分析 (3 小时)

情感分析 (Sentiment Analysis)是文本挖掘中的经典研究方法，它指对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。随着互联网的快速发展，网络上产生了大量的用户评论信息。这些评论信息表达了人们的各种情感色彩和情感倾向性，如喜、怒、哀、乐和批评、赞扬等。基于此，通过分析网络上评论信息的感情倾向和主观色彩就可以进一步了解大众舆论对于某一事件或某一主题的具体看法。自 2000 年初以来，情感分析已经成长为自然语言处理中最活跃的研究领域之一。

情感分析的方法大体可以分为两类。

- **情感词典**。通过判断文本中包含情感词典中词语的情况来计算文本的情感得分，这种方法在使用时简单、便捷，但会受制于情感词典的完备性以及和研究问题的匹配程度，例如某些特定问题 (如金融科技)在通用情感词典上的结果并不好。
- **机器学习**。考虑一个简单的情感分析问题 (如积极 v.s. 消极)，则可将情感分析转化为一个二分类问题，因此可以应用各种机器学习算法甚至深度学习算法来训练情感分类器。但是这种方法的应用前提是需要有一个足够大的训练数据集 (即已经知道情感分类的文本数据)。

因此，本节将对上述两种情感分析方法分别进行详细介绍，具体来说：

- **情感词典法**：归纳总结现在中英文中常用的各种情感词典，对比它们的优缺点和使用范围
- **机器学习法**：介绍情感分析中常用的各种机器学习方法，讲解方法的核心思想和利弊
- **代码实现**：基于股评数据，展示上述两种情感分析方法的 R 代码实现
- 主要参考文献：

- 王靖一、黄益平. 金融科技媒体情绪的刻画与对网贷市场的影响.《经济学(季刊)》, 2018, 17(4): 1623-1650. -PDF-

T3. 主题模型 (3 小时)

以 **LDA** 模型 (Latent Dirichlet Allocation) 为基础的主题模型 (Topic Models)是文本分析的利器，主要用于文本分类。LDA 模型自 2003 年一经提出就引起了学者们的广泛关注。在 LDA 提出之前，常用的文本结构化方法是 one-hot。什么是 one-hot 编码呢？想象你有一个新华字典，里面一共出现了 V 个词，将这个字典中的词从头到尾逐一编号，因此每个词就有了一个编号。假设有一个三个字组成的句子“我 爱 你”，one-hot 编码就是用一个 V 维的 0-1 向量来表示这句话，向量的每个位置会对应一个新华字典中的词，如果这个词出现就标记为 1，反之为 0。这种表示方法会得到一个高维稀疏的向量，因此不利于后续建模。

主题模型可以如何改进呢？它假设所有文本其实表达了 K 个主题，每个文本在 K 个主题上的表达权重是不同的，因此可以用一个 K 维的向量来表示这条文本。通过这种方式，可以将文本表示为一个主题所占权重的 K 维向量，从而实现文本的降维表示 (因为 K 往往小于 V)。与此同时，找到的 K 个主题可以帮助总结概括文本集合的含义，帮助读者更好的理解文本内容。

本节将对主题模型进行详细介绍，具体包括：

- **主题模型简介**：从基础的 LDA 模型开始，介绍主题模型的核心思想、文档生成过程，主题模型输出的结果，以及基于主题模型的各种扩展模型
- **应用实例：** (1) 电影影评内容是否能帮助票房预测？(2) 客户-客服对话文本是否能用于预测客户流失？
- **代码实现**：以影评数据为例介绍主题模型的 R 语言代码实现以及可视化方法
- 主要参考文献：
  - Blei, D. M., A. Y. Ng, M. I. Jordan, **2003**, Latent dirichlet allocation, **Journal of machine Learning research**, 3: 993-1022. -Link-, -PDF-, PDF2.
  - Blei, D. M., **2012**, Probabilistic topic models, **Communications of the ACM**, 55 (4): 77-84. -Link-, -PDF-, PDF2, -cited-
  - 王菲菲, 刘雯珺, 朱立奥, 吕晓玲. **2023**. 基于客户-客服沟通文本信息的客户流失研究. **管理学报**. 即将刊出. -PDF-

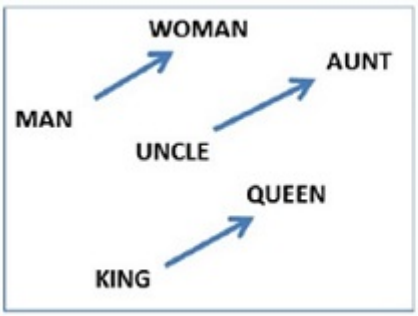
T4. 文本深度学习 (3 小时)

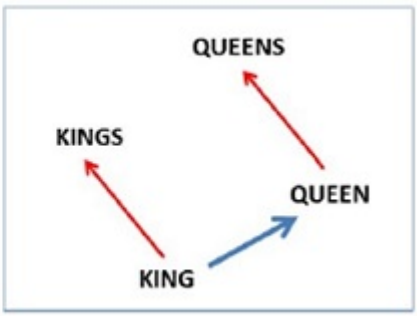
深度学习是一种复杂的机器学习算法，在近年来得到了非常广泛的应用，并显示出强劲的生命力，在很多领域取得了非凡的成果，效果远超先前的相关技术。在文本分析领域，深度学习也取得了很好的效果，例如 ChatGPT 就是集深度学习之大成的典型。本节将介绍深度学习在文本分析领域的一些应用，主要涉及两部分内容：

4A. 词嵌入

什么是词嵌入呢？前面我们介绍过，最简单的文本结构化表示方法就是 one-hot 编码，也就是用一个 0-1 的高维稀疏向量来表示文本。而词嵌入 (word embedding)方法是将每个词用一个稠密的实值向量表示。如果两个词语的语义比较接近，那么这两个词之间的向量距离很近 (用余弦来衡量)。 词嵌入方法现在已经成为情感分析、文本摘要、语言翻译或其他文本分析任务的基础。

下图给出了理想情况下一些词对应的向量的示意图。MAN 的词向量和 WOMAN 的词向量之间的距离，等于 UNCLE 和 AUNT 两个词对应的向量之间的距离，也等于 KING 和 QUEEN 两个词对应的向量之间的距离，因为上述每对单词之间的差别就在于性别。KINGS 的词向量减去 KING 的词向量加上 QUEEN 的词向量就等于 QUEENS 的词向量，因为 KINGS 和 KING 之间的差异以及 QUEENS 和 QUEEN 之间的差异都是复数和单数之间的差异。因此在本节中我们将首先为大家介绍词向量模型，也就是词向量是怎么来的，以及现在常用的一些词向量调用方法。





- 参考文献：
  - Mikolov, T., K. Chen, G. Corrado, J. Dean, **2013**, Efficient estimation of word representations in vector space, **arXiv preprint**. -Link-, -PDF-, PDF2

4B. 深度学习模型

目前针对文本数据的深度学习模型，包括前馈神经网络 (Feedforward Neural Network， **FNN**)、循环神经网络 (Recurrent Neural Network， **RNN**)、长短期记忆模型 (Long Short Term Memory，LSTM)、卷积神经网络 (Convolutional Neural Network， **CNN**)、基于 Transformer 的双向编码器表示 (Bidirectional Encoder Representations from Transformers， **BERT**)。我们将介绍这些模型的原理及其在在 R 中的实现和解读。

- 参考文献：
  - Mikolov T, M Karafiát, Burget L, et al. Recurrent neural network based language model. **Interspeech, 2010**: 1045-1048. -Link-, -PDF-
  - Hochreiter, S., J. Schmidhuber, **1997**, Long short-term memory, **Neural Comput**, 9 (8): 1735-1780. -Link-, -PDF-, PDF2
  - Devlin, J., M.-W. Chang, K. Lee, K. Toutanova, **2018**, BERT: Pre-training of deep bidirectional transformers for language understanding, **arXiv preprint arXiv:1810.04805**. -PDF-, Slides

T5. 文本分析：论文精讲 I (3 小时)

学习了这么多文本分析方法，能干什么呢？让我们来做一个综合训练。我们的目标文章是：

- Bellstam, G., S. Bhagat, J. A. Cookson, **2021**, A text-based analysis of corporate innovation, **Management Science**, 67 (7): 4004-4031. -Link-, -PDF-, PDF2, Appendix, Replication, -cited-

这篇文章从「文本分析」视角重新测度了企业创新能力。创新在经济发展中具有核心地位，然而对于创新的衡量却缺乏标准化体系，现行方法主要着眼于研发投入和专利数量等特定指标，但是难以衡量更广泛意义上的创新行为。为了解决这一问题，作者以分析师报告为基础，通过对分析师报告的内容建立**主题模型**，挖掘其中和创新有关的主题，然后以此作为企业创新性的测度。基于主题模型的结果，作者还使用回归建模的方法做了很多后续分析，这也是经管文章中应用主题模型的常用套路，即首先得到主题分布，然后使用主题分布进行后续建模。因此本节将首先详细介绍这篇文章的思路、想法、主要结果。

受这篇文章启发，我们将其应用在中国情境下。具体来说，我们针对中国上市公司的企业现状，通过对 2021 至 2022 年间的券商研究报告文本进行主题建模，得到衡量企业创新行为的文本创新得分，并进一步探究了该得分对中国上市公司业绩表现的影响。最后，我们将通过示例数据带领大家复现上述文章。

通过对这篇文章的复现，你会学到什么？

- 只要想法好，不需要模型创新，只是简单应用学过的文本分析方法也能发 TOP
- 掌握应用主题模型分析经管类问题的套路

T6. 文本分析：论文精讲 II (3 小时)

第二篇综合实践的文章选的是：

- Wang, F., J. Liu, H. Wang, **2021**, Sequential text-term selection in vector space models, **Journal of Business & Economic Statistics**, 39 (1): 82-97. -Link- , -PDF- , PDF2

这篇文章提出了一个新方法，具有广泛的应用场景。简单来说，这篇文章给出了一个方法论，用于探究【文本数据】是否会影响某个【连续型因变量】。例如：

- 连续型因变量是：商品评分；文本数据是：用户评论；那么该文章可以分析用户的评论内容是否会影响商品评分？到底是哪些内容有影响？影响的方向是什么？
- 连续型因变量是：新闻阅读量；文本数据是：新闻标题；那么该文章可以分析新闻标题是否会影响新闻的阅读量？到底是哪些词语有影响？影响的方向是什么？
- 连续型因变量是：商品销量；文本数据是：广告文案；那么该文章可以分析广告的文案内容是否会影响商品销量？到底是哪些内容有影响？影响的方向是什么？

在课程中，我们将具体展示如何基于这篇文章的方法探究产品好评率的影响因素。我们将给出该问题的具体分析过程，复现文章中的方法。与此同时，我们也会给出简化版的代码，方便大家在其他研究问题上快速应用该方法进行研究。

5. 报名和缴费信息

- **主办方：** 太原君泉教育咨询有限公司
- **标准费用** (含报名费、材料费)：3300 元/人 (全价)
- **优惠方案**：
  - 三人及以上团购/专题课老学员：9 折，2970 元/人
  - 学生 (需提供学生证/卡照片)：9 折，2970 元/人
  - 连享会会员：85 折 2805 元/人
  - **温馨提示：** 以上各项优惠不能叠加使用。
- **联系方式：**
  - 邮箱： wjx004@sina.com
  - 电话 (微信同号)： 王老师 18903405450 ; 李老师 18636102467

> **报名链接：** https://www.wenjuan.com/s/UZBZJvMpWH/#
>
> 或 长按/扫描二维码报名：

缴费方式

> **方式 1：对公转账**

- 户名：太原君泉教育咨询有限公司
- 账号：35117530000023891 (晋商银行股份有限公司太原南中环支行)
- **温馨提示：** 对公转账时，请务必提供「**汇款人姓名-单位**」信息，以便确认。

> **方式 2：微信扫码支付**

> **温馨提示：** 微信转账时，请务必在「添加备注」栏填写「**汇款人姓名-单位**」信息。

6. 听课指南

6.1 软件和课件

听课软件：**支持** 手机，ipad，平板以及 windows/Mac 系统的笔记本，**但不支持台式机**

> **特别提示：**

- 为保护讲师的知识产权和您的账户安全，系统会自动在您观看的视频中嵌入您的「用户名」信息。
- 一个账号绑定一个设备，且听课电脑不能外接显示屏，请大家提前准备好自己的听课设备。
- 本课程为虚拟产品，**一经报名，不得退换。**
- 为保护知识产权，课程不允许以任何形式录屏及传播。

6.2 实名制报名

本次课程实行实名参与，具体要求如下：

- 高校老师/同学报名时需要向连享会课程负责人 **提供真实姓名，并附教师证/学生证图片**；
- 研究所及其他单位报名需提供 **能够证明姓名以及工作单位的证明**；
- 报名即默认同意「 **连享会版权保护协议条款** 」。

7. 诚聘助教

- **名额：** 10 名
- **任务：**
  - **A. 课前准备**：协助完成 3 篇介绍 Stata 或 Python 或 R 语言 和计量经济学基础知识的文档，风格类似于 lianxh.cn ；
  - **B. 开课前答疑**：协助学员安装课件和软件，在微信群中回答一些常见问题；
  - **C. 上课期间答疑**：针对前一天学习的内容，在微信群中答疑 (8:00-9:00，19:00-22:00)；
  - Note: 下午 5:30-6:00 的课后答疑由主讲教师负责。
- **要求：** 热心、尽职，熟悉 Stata 或者 Python 或者 R 的基本语法和常用命令，能对常见问题进行解答和记录。
- **特别说明：** 往期按期完成任务的助教可联系连老师直录，优先考虑熟悉 Python 和 R 的申请者。
- **截止时间：** 2023 年 10 月 25 日 (将于 10 月 27 日公布遴选结果于连享会主页 lianxh.cn)。

> **申请链接：** https://www.wjx.top/vm/PpL5PH7.aspx#

> 或扫码填写助教申请资料：

> **课程主页：** https://www.lianxh.cn/