连享会 · 文本分析专题

连享会 | 推文 | 公开课

1. 课程概览

A. 基本信息

- **嘉宾**：陈婷 (香港浸会大学)
- **时间：** 2025 年 11.22, 11.29, 12.6 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
- **方式**：线上直播 + 30 天回放
- **课程主页**： https://kc.lianxh.cn/text.html
- **PDF 大纲**： https://kc.lianxh.cn/text.pdf
    - 预读资料和参考文献： 点击查看
- **助教招聘**： https://www.wjx.top/vm/QZGUC9I.aspx#
- **报名链接**： https://www.wjx.top/vm/r9ZLwkR.aspx#

B. 嘉宾简介

陈婷，香港浸会大学经济系副教授，商学院商业分析与数字经济中心副主任。主要研究方向包括政治经济学、经济史与长期经济发展；成果见诸 Quarterly Journal of Economics (**QJE**), Economic Journal (**EJ**), Journal of Politics、Journal of Development Economics (**JDE**), Journal of Econometrics (**JoE**), Journal of Comparative Economics (**JCE**), 以及《经济研究》《经济学(季刊)》等期刊。教学方面，陈婷老师在浸会大学、复旦大学等高校开设机器学习与文本分析相关课程，强调从核心概念到可复现实践的连贯路径，颇受好评。相信不少人对「科举万岁！中国科举制度所产生的深远影响」这篇论文并不陌生：其研究设计与识别策略颇具启发性，也常成为博士生组会讨论的重点范例，而这正是陈婷老师的代表作之一。详见 Google Scholar 。

LONG LIVE *KEJU*! THE PERSISTENT EFFECTS OF CHINA'S CIVIL EXAMINATION SYSTEM*

Ting Chen, James Kai-sing Kung and Chicheng Ma

China's civil examination system (*keju*), an incredibly long-lived institution, has a persistent impact on human capital outcomes today. Using the variation in the density of *jinshi*—the highest qualification—across 278 Chinese prefectures in the Ming-Qing period (c. 1368–1905) to proxy for this effect, we find that a doubling of *jinshi* per 10,000 population leads to an 8.5% increase in years of schooling in 2010. The persistent effect of *keju* can be attributed to a multitude of channels including cultural transmission, educational infrastructure, social capital and, to a lesser extent, political elites.

2. 课程导引

A. 为何要学习文本分析?

在经管研究中，越来越多关键变量开始直接来自「政策文件、年报、舆情与专利」等非结构化文本。为了让这些信息真正服务研究问题，我们需要把文本稳定地转化为可进入计量模型的指标。基于这一目标，文本分析的基本理念是“**文本即数据 (Text as Data)**”：先完成必要的预处理，再选择合适的表示方式，并据此建立模型，将原始文本转换为结构化的特征或指标。

具体来说，我们可以用 **TF-IDF/相似度** 衡量文本差异，以 **主题模型** 概括议题结构，并借助 **词嵌入与上下文嵌入** 获得更稳健的语义表示。随后，这些指标便能自然嵌入分类、聚类或回归等熟悉的计量框架，最终进入回归、面板或 DID 的经验分析流程。需要强调的是，文本分析的目的并非“炫技”，而是**更好地回答研究问题**，即在可复现、可解释的前提下，把文本信息纳入识别与机制检验。

过去，文本分析往往需要扎实的编程与机器学习基础，令不少研究者望而却步；然而，随着「生成式人工智能」的发展以及「多模态方法」的逐步普及，文本、表格、图像与影音等信息可以在同一研究设计中被更全面地利用。进一步地，借助 「GenAI 工具与提示词范式」，我们只要掌握若干关键概念与流程，即可在普通电脑上跑通从清洗、表示、建模，到指标导出与实证整合的全链路，从而显著降低了落地成本。

B. 如何学好这门课?

本课程遵循“**可理解** → **可复现** → **可应用**”的递进路径。首先，我们从整体流程入手，建立连贯的心智图：文本清洗 → 表示 → 建模 → 指标导出 → 与结构化数据合并 → 回归 / DID。通过清晰步骤将环节串联起来，学员能够在每一次推进时明确 “为什么这样做” 以及 “下一步如何衔接”。

- **在表示层面**，我们会酌情选用 **TF-IDF** 与 **句向量** 方法：当研究强调关键词稀疏性与可解释性时，TF-IDF 往往更合适；当研究需要捕捉上下文语义与句法结构时，基于 **词嵌入与上下文嵌入** 的句向量更具优势。围绕这些选择，我们通过参数对比与误差来源的拆解来说明“表示如何改变结论”，而不是停留在名词解释。
- **在建模层面**，我们基于「能跑通、能分析、能迁移」这一目标，重点介绍 **监督学习** 与 **无监督学习** 两类方法：前者侧重问题表述、标注策略、类别不平衡与阈值选择，并讨论外推风险；后者则聚焦于主题结构、相似度网络与聚类稳定性，并说明超参数的敏感性。所有模型均配有可直接运行的 `.ipynb` 与成体系的提示词范式，在普通电脑上 5–10 分钟即可得到可复核结果；输出统一为 `score.csv` 、 `topics.csv` 等表格，便于与面板数据或 DID 脚本一键合并。
- **在“原理 v.s 应用”的取舍上**，我们采用“**理论打底、应用牵引**”的双轮驱动：必要理论以直观方式讲清 **词向量、词嵌入、Transformer** 的工作机制与局限，它们既是文本分析的核心概念，也是通向多模态方法的基础；应用环节则对齐 **Top 期刊** 的可复现做法，演示如何将情感分数、主题比例、政策相似度等指标纳入回归与机制检验。通过这种安排，理论提供可解释与可迁移的稳固基座，应用确保产出能够进入标准计量框架并经得起复核。

为确保“可复现”与“可迁移”，每个专题都配套对照的代码与提示词、典型范例数据、统一的指标命名与输出格式，并安排简短的错误分析与稳健性演示。由此，你不仅能看到“模型为何有效”，也能在出现偏误时进行定位与修正：从清洗与表示可能引入的偏差，到样本外验证与灵敏度分析的操作步骤，课程还提供“把课堂脚本改造成论文”的清单式说明，帮助你将示例稳妥移植到自己的研究项目中。

C. 课程目标

我们希望大家完成本课程的学习后，能够达成如下目标：

- 掌握从 **中文文本清洗** → **表示** → **建模** → **指标导出** → **并入回归/DID** 的完整流程。
- 形成一套可直接改用的 **最小工作范式** (统一输入输出、固定随机种子、可重复结果)。
- 能将文本指标用于 **面板回归、异质性分析** 等实证任务，并完成 **稳健性** 与 **可视化**。
- 学会用 **提示词** 与 GenAI 辅助生成、校对与解释代码与图表，提高研究效率与可读性。

D. 课程特色

- **中文场景优先**：分词、繁简转换、PDF/OCR 清洗、数字与单位标准化，覆盖文本到矩阵的关键细节。
- **拿来即用**：每个专题有 5–10 分钟可运行的小例子，统一导出 `clean.csv` 、 `features.csv` 、 `score.csv` 、 `topics.csv` ，便于一键并入回归脚本。
- **期刊导向**：精选 **Top 期刊** 应用作为范式，强调可解释、可复现实证结果与写作呈现。
- **提示词驱动**：配套一组覆盖「清洗–表示–建模–评估–回归」的系统提示词，便于零基础学员用自然语言编程完成作业。

3. 课程详情

本课程围绕 **文本** → **指标** → **实证** 的主线展开：

- **Day 1** (T0-T2)：概览与研究设计；中文文本预处理与表示；相似度与差异度的研究范式。
- **Day 2** (T3-T4)：LDA 主题模型；STM/keyATM 与短文本主题；主题稳定性与可读性。
- **Day 3** (T5-T6)：词嵌入与上下文嵌入、LLM 的教学化用法；文本指标接入回归/DID 与写作呈现。

T0. 概述：我们能用文本做什么?

本讲介绍文本数据在经济学和社会科学中的应用，旨在引导大家了解文本如何作为一种数据形式进入计量分析框架。主要内容包括：

- **文本数据的核心应用**：重点讨论文本数据在经济学中的应用，如从政策文件、年报、专利中提取可量化变量；如何将其转化为回归分析中的控制变量和因果推断中的识别变量。通过经典文献来说明文本任务（如情感分析、主题建模等）如何与经典计量模型（如面板数据分析、DID 等）结合，实现在政策评价和经济决策中的数据驱动。
- **文本分析方法的脉络**：从词袋模型、TF-IDF 到主题模型、词嵌入，再到现代的上下文嵌入和大语言模型，梳理各类方法的背景、原理及其优缺点。重点讨论在实际研究中如何根据研究问题的需要，合理选择方法并解决文本分析中遇到的挑战，如超参数选择、模型解释性等。

亮点

- 通过经典文献案例，展示文本数据在经济学研究中的各类应用，扩展大家的研究和分析思路。
- 学生将能够理解如何将文本数据转化为具有经济学含义的可量化指标，并能够在实际分析中合理选择与任务相关的文本处理技术（如主题模型、情感分析、词嵌入等）。
- 本讲还将强调如何控制模型的可复现性和稳健性，确保分析结果的可靠性。

参考文献

- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, 57(3), 535–574. Link, PDF, Google.
- Ash, E., & Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, 15, 659–688. Link, PDF, Google.
- Wilkerson, J. D., & Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science*, 20, 529–544. Link, PDF, Google.
- Roberts, M. E. (2016). Introduction to the Virtual Issue: Recent Innovations in Text Analysis for Social Science. Political Analysis, 24(V10), 1–5. Link, PDF, Google.

T1. 文本预处理：文档-特征矩阵与文本清洗

本讲介绍文本预处理和表示的基本方法，重点讲解如何将非结构化的文本数据转化为结构化的形式，以便进行后续的分析和建模。主要内容包括：

- **文档-特征矩阵的构建**：首先介绍经典的向量空间模型（Vector Space Model）以及基于词袋模型的文档表示方法，重点讲解如何通过词频、TF-IDF 等方式来为每个文档赋予特征向量。这些特征向量是文本分析中的基础，能够将文本信息转化为可处理的数字形式，为后续分析打下基础。
- **文本预处理的挑战与方法**：讲解文本预处理中的常见问题，如中文分词、中文情感分析等，并结合具体案例分析预处理在文本分析中的作用和潜在问题。特别是讨论在无监督学习中，文本预处理如何可能导致误导性的结果，并为此提供改进建议。

亮点

- 学生将掌握如何将非结构化的文本转化为结构化的数据，理解文档表示的基本方法，如词袋模型和 TF-IDF 模型，并能够应用于文本分类和聚类等任务。
- 本讲通过深入讨论文本预处理的挑战，帮助学生认识到预处理环节对最终分析结果的影响，并学习如何合理地选择预处理方法，避免误导性处理。
- 强调如何将文本数据与传统的统计模型相结合，为后续的因果推断分析奠定数据基础，确保分析的准确性和可靠性。

参考文献

- Manning, C. D., Raghavan, P., & Schütze, H. (2009). Chapter 6: Scoring, Term Weighting, and the Vector Space Model. *Introduction to Information Retrieval* . Cambridge University Press. Link , PDF , Google .
- Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis* , 26(2), 168–189. Link , PDF , Google .

T2. 文本相似度与语义距离：指标与应用

本讲介绍文本相似度与语义距离的基本概念，重点探讨如何通过度量文本之间的相似性来进行内容比较和风格分析。主要内容包括：

- **测量语言独特性与语言风格**：本部分介绍如何通过分析文本的语言特点（如词频、句式结构、用词习惯等）来衡量语言的独特性，尤其在不同领域和作者之间的差异。通过具体案例，阐述如何通过文本的语言风格区分技术创新、政治立场等不同主题的文本特征。
- **语义距离与语义相似度的比较**：介绍不同的度量方法，如余弦相似度、Jaccard 相似度等，用于计算文档之间的语义距离，并讨论其在实证研究中的应用。讲解如何利用这些方法比较和分类不同文档，评估文本内容的相似性，特别是在经济学和社会科学研究中的实际应用，如衡量产品差异化、技术创新或政治观点的相似度。

亮点

- 学生将掌握如何运用文本相似度和语义距离度量方法，在实证分析中对文档进行比较，识别语言风格和主题之间的潜在关系。
- 通过案例分析，学生将学到如何衡量技术创新、政治立场等在文献中的表现，并能够运用这些技术分析文本数据，揭示不同文本之间的语义相似度。
- 本讲还将讨论这些度量方法在实证研究中的局限性和挑战，帮助学生深入理解文本相似度与语义距离的计算原理和实际应用中的注意事项。

参考文献

- Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring Technological Innovation over the Long Run. *AEA Papers and Proceedings (AER: Insights)* , 3(2), 61–67. Link , PDF , Google . github
- Kostovetsky, L., & Warner, J. B. (2020). Measuring Innovation and Product Differentiation: Evidence from Mutual Fund Prospectuses. *The Journal of Finance* , 75(6), 3255–3303. Link , PDF , Google .
- Bertrand, M., Bombardini, M., Fisman, R., Hackinen, O., & Trebbi, F. (2021). Hall of Mirrors: Corporate Philanthropy and Strategic Advocacy. *The Quarterly Journal of Economics* , 136(4), 2233–2280. Link , PDF , Google .
- Huang, L., Perry, P. O., & Spirling, A. (2020). A General Model of Author “Style” with Application to the UK House of Commons, 1935–2018. *Political Analysis* , 28(3), 412–434. Link , PDF , Google .
- Benoit, K., Munger, K., & Spirling, A. (2019). Measuring and Explaining Political Sophistication through Textual Complexity. *American Journal of Political Science* , 63(2), 491–508. Link , PDF , Google .

T3. 主题模型、LDA 模型

本讲介绍主题模型的基本概念，并详细讲解生成式主题模型（LDA）的工作原理。主要内容包括：

- **主题模型的基本概念**：阐述什么是主题模型，以及它在文本分析中的作用。重点讨论主题模型如何通过无监督学习的方式，从大量文本中自动提取出潜在的主题，并揭示文本的隐藏结构。
- **LDA 模型的工作原理**：讲解潜在狄利克雷分配（LDA）模型的基本假设与算法流程，包括如何根据文本中的词频信息，推测每个文档的潜在主题分布，并从中提取出有意义的主题。通过案例演示，展示如何使用 LDA 从实际数据中提取主题，如何对提取出的主题进行解释，并结合实际应用讨论主题模型的优缺点。

亮点

- 学生将理解 LDA 模型的核心原理，学会如何在实际研究中应用 LDA 模型从文本中提取主题。
- 通过具体的案例分析，学生将学会如何使用 LDA 模型处理实际数据，提取与研究问题相关的潜在主题，并对提取出的主题进行有效解读。
- 本讲还将深入讨论主题模型在实际应用中的局限性，如主题解释的难度、模型的选择与调优，并介绍如何评估主题模型的效果，帮助学生在实际操作中避免常见错误。

参考文献

- Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and Deliberation within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics* , 133(2), 801–870. Link , PDF , -PDF2- , Google .
- Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. *Management Science* , 64(6), 2833–2855. Link , PDF , Google .

T4. 主题模型进阶：引入协变量，提高可解释性

本讲介绍主题模型进阶方法，重点讲解如何引入协变量提升主题模型的可解释性，帮助研究者从文本数据中提取出更有意义的主题信息。主要内容包括：

- **引入协变量的动机与方法**：讲解如何在传统的 LDA 模型中引入协变量，从而提升主题模型的可解释性。例如，如何将文本的元数据（如时间、作者信息、文本来源等）作为协变量，帮助分析文本主题背后的潜在因素。通过结合协变量，模型能够识别更具实用性和可解释性的主题结构，增强文本分析的实际应用价值。
- **结构主题模型（STM）与其他进阶方法**：介绍结构主题模型（STM）的工作原理，探讨如何通过协变量和元数据调整模型，使得主题分析更加精准和具有理论指导意义。讲解其他先进的主题建模方法，如 `Top2Vec` 和 `BERTopic` ，并分析它们在不同数据集和应用场景中的优势与局限性。

亮点

- 学生将学习如何将协变量与 LDA 模型结合，提升主题模型的可解释性，并通过实践案例掌握如何在不同研究情境中合理选择协变量。
- 本讲通过介绍结构主题模型（STM）及其他先进方法，帮助学生深入理解主题模型在社会科学中的应用，尤其是在需要提高模型可解释性时如何进行调整。
- 通过对比不同的主题建模技术，学生将能够理解每种方法的优缺点，并能根据实际数据和研究目标选择最合适的建模技术，提升分析的深度和广度。

参考文献

- Cong, L. W., Liang, T., & Zhang, X. (2018/2024). Textual Factors: A Scalable, Interpretable, and Data-Driven Approach to Analyzing Unstructured Information. *(SSRN Working Paper; NBER Working Paper No. 33168)* . Link , PDF , Google .
- Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2020). The Structure of Economic News. *NBER Working Paper* No. 26648. Link , PDF , Google .
- Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. (2013). The Structural Topic Model and Applied Social Science. *NIPS Workshop Paper* . PDF , Google . Slides

- Eshima, S., Imai, K., & Sasaki, T. (2020). Keyword Assisted Topic Models. *arXiv preprint* arXiv:2004.05964. Link, PDF, Google.
- Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12), 2928–2941. Link, PDF, Google.
- Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. *arXiv preprint* arXiv:2008.09470. Link, PDF, Google.
- Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7, 886498. Link, PDF, Google.

T5. 词义的量化：词嵌入, Word2vec, GloVe

本讲介绍词义的量化方法，重点讲解如何使用词嵌入技术（如 `Word2Vec` 和 `GloVe`）对文本中的词语进行向量化，从而为文本分析提供更丰富的语义信息。主要内容包括：

- **词嵌入的基本概念**：解释词嵌入的基本原理和方法，阐述如何通过将词语映射到向量空间来捕捉词语之间的语义关系，进而解决传统文本分析中“词与词之间没有直接联系”的问题。
- **深度学习框架与词嵌入算法**：介绍 `Word2Vec` 和 `GloVe` 两种常见的词嵌入算法，详细讲解它们的工作原理、优缺点以及在大规模文本数据中的应用。同时，探讨深度学习框架（如神经网络）在词嵌入中的作用和应用，帮助学生理解现代文本表示方法。
- **词嵌入在社会科学中的应用**：讨论词嵌入如何用于社会科学中的研究，如分析性别态度、政治概念演变、官僚声誉等，通过具体案例展示如何将词嵌入技术应用于实际研究中，从而揭示文本背后的深层含义。
- **文档嵌入与复杂数据结构嵌入**：介绍如何将词嵌入扩展到文档级别，探讨文档嵌入的方法和应用，以及如何处理更复杂的数据结构（如句子、段落等），以便在更高层次上进行文本分析。

亮点

- 学生将掌握词嵌入技术的基本原理，并能够在文本分析中应用这些技术，进行词语语义的量化和向量表示。
- 通过深入讲解 `Word2Vec` 和 `GloVe` 算法，学生将学会如何使用这些算法将文本数据转化为数字化的表示形式，为后续的文本分析提供支持。
- 本讲通过具体案例，展示词嵌入在社会科学中的实际应用，帮助学生理解如何通过词嵌入技术进行语义分析，揭示文本的深层信息，并能够运用这些技术进行政治、社会及经济研究中的实证分析。

参考文献

- Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing Meaning through Word Embeddings. *American Sociological Review*, 84(5), 905–949. Link, PDF, Google.
- Ash, E., Chen, D. L., & Ornaghi, A. (2024). Gender Attitudes in the Judiciary: Evidence from US Circuit Courts. *American Economic Journal: Applied Economics*, 16(1), 314–350. Link, PDF, Google.
- Rodriguez, P. L., & Spirling, A. (2022). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *Journal of Politics*, 84(1), 101–115. Link, PDF, Google.
- Bellodi, L. (2022). A Dynamic Measure of Bureaucratic Reputation: New Data for New Theory. *American Journal of Political Science*. Forthcoming. Link, PDF, Google.
- Rodman, E. (2020). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. Political Analysis, 28(1), 87–111. Link, PDF, Google.
- Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2023). Embedding Regression: Models for Context-Specific Description and Inference. *American Political Science Review*, 117(4), 1255–1274. Link, PDF, Google.

T6. 序列模型：上下文嵌入和大语言模型

本讲简要介绍序列模型的核心概念和算法原理，重点讲解如何通过上下文嵌入和大语言模型（如 GPT-3）来处理和生成自然语言文本。主要内容包括：

- **N-gram 模型与神经语言模型**：首先介绍传统的 N-gram 模型，并逐步过渡到基于神经网络的语言模型。通过比较不同的语言建模方法，阐述神经网络如何提高语言模型的表现，尤其是在处理长期依赖和上下文信息时的优势。
- **深度学习架构：CNN、RNN、LSTM、BERT 和 GPT-3**：深入讲解卷积神经网络（CNN）、递归神经网络（RNN）、长短期记忆网络（LSTM）、BERT 和 GPT-3 等模型，展示它们在序列数据处理中的应用及其改进的原理。特别是讨论 GPT-3 和 BERT 如何通过自注意力机制和大规模预训练，提升文本生成和理解的能力。
- **Transformer 和大语言模型的革命性影响**：介绍 Transformer 架构的原理及其在 NLP 任务中的应用，重点讲解 Transformer 如何为 GPT-3 等大语言模型提供高效的训练框架，进而推动自然语言处理的重大进步。
- 介绍 Dell (2025) 列举的一些经济学中大语言模型的应用案例，如文本生成、翻译、问答等，讲解这些应用的 Python/R 实现方法，以及可能的扩展方向。

亮点

- 本讲通过深度学习架构的介绍，帮助学生理解 CNN、RNN、LSTM、BERT 和 GPT-3 的工作原理，并展示它们在文本生成、翻译、问答等任务中的应用。这有助于大家持续追踪自然语言处理领域的最新进展，并将这些技术应用于实际研究中。
- 通过讲解 Transformer 和大语言模型，学生将了解当前自然语言处理技术的最前沿，掌握如何利用这些技术进行高效的文本生成和理解。课程中将通过 1-2 个实例，展示实操中的完整流程和常见问题，以便大家能够将这些方法快速应用到自己的研究中。

参考文献

- Dell, M. (2025). Deep Learning for Economists. *Journal of Economic Literature* , 63(1), 5–58. Link , PDF , Google . Slides
- Manning, C. D., Raghavan, P., & Schütze, H. (2009). (Deep learning materials referenced as background in IR book; chapter reference as above). *Introduction to Information Retrieval* . Cambridge University Press. PDF , Google .
- Ash, E., Durante, R., Grebenshchikova, M., & Schwarz, C. (2022). Visual Representation and Stereotypes in News Media. *CEPR Discussion Paper DP16624 / CESifo Working Paper 9686* . PDF , Google .
- Alammar, J. (2018). Visualizing a Neural Machine Translation Model (Mechanics of Seq2seq Models with Attention). *Blog post* . Link , 中文 .
- Alammar, J. (2018). The Illustrated Transformer. *Blog post* . Link .
    - **Update:** This post has now become a book! Check out LLM-book.com which contains (Chapter 3) an updated and expanded version of this post.

附：预读资料

> 开课前，大家可以根据自己的基础阅读一些入门材料，帮助理解课程内容。课后也可以参考这些文献，深入学习相关方法和应用。

> **Textbooks**

- Grimmer, J., Westwood, S. J., & Messing, S. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences* . Princeton University Press. -Link- , Google

- Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed., draft). -Link-and-Slides , Google , -PDF- , github
- Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media. -Link- , Google , -PDF-2E , GitHub
-

> **R 和 Python 补充资料**

- Kohei Watanabe & Stefan Müller. Quanteda Tutorials. (R), -Link- , google
- (Optimal) Cornelius Puschmann. 2019. "Advancing Text Mining with R and quanteda." -Link- , Slides
- (Optimal) Open Source Tools for Text-as-Data/NLP in R -Link-
- (Optimal) Open Source Tools for Text-as-Data/NLP in Python -Link- , google
- (Optimal) Mango Solutions. 2018. "Python for R Users Workshop." -Link-

4. 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.wjx.top/vm/eUttMRE.aspx# ，或 长按/扫描二维码报名：

> **缴费方式 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。

> **缴费方式 2：扫码支付**

> **温馨提示：** 扫码支付时，请务必在「添加备注」栏填写「**汇款人姓名-单位**」信息。

5. 听课指南

> **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。

> **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** 」。

6. 诚聘助教

- **名额：** 10 名
- **任务：** 详情参见 连享会助教工作指南
  - **A. 课前准备**：完成 2 篇推文，风格参见连享会主页 www.lianxh.cn，选题参见 这里；
  - **B. 开课前答疑**：协助学员安装软件和使用课件，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，熟悉常用的 AI 工具，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以直接联系连老师直录。
- **截止时间：** 2025 年 10 月 27 日 (将于 10 月 29 日公布遴选结果于 课程主页，及连享会主页 lianxh.cn)

> **申请链接：** https://www.wjx.top/vm/QZGUC9I.aspx#，或扫码填写：