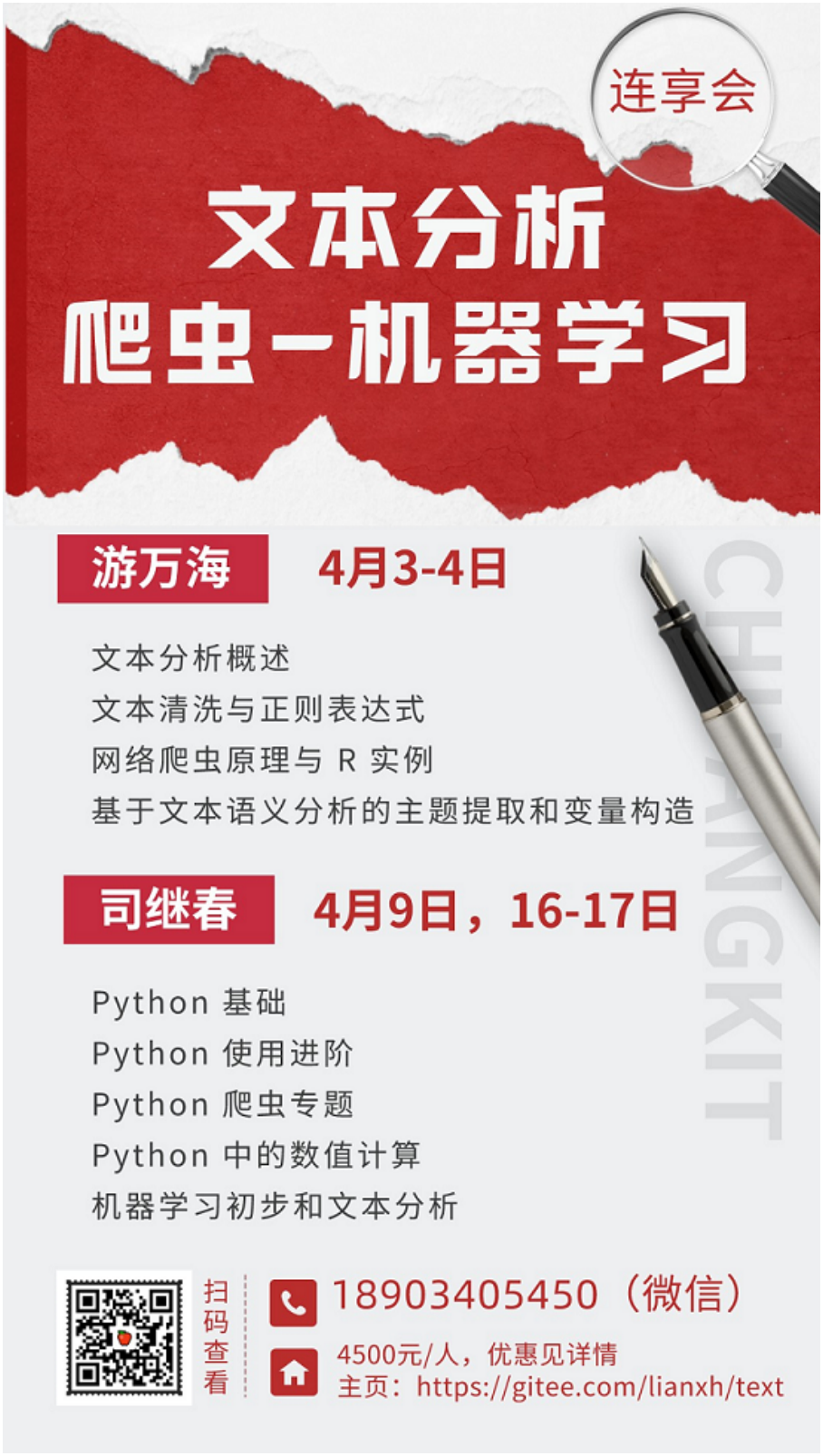
连享会 · **文本分析 | 爬虫 | 机器学习**



目录

1. 课程概览

- **课程主页：** https://gitee.com/lianxh/text
- **课程大纲：** **下载 PDF 版本**
- **授课方式：** 线上直播 (5 天) + 一周回放。
- **授课时间：** 2022.4.3-4; 4.9, 4.16-17 (三个周末)
- **授课嘉宾：**
    - 游万海，福州大学 (第 1-4 讲，2022.4.3-4)
    - 司继春，上海对外经贸大学 (第 5-10 讲，2022.4.9，4.16-17)
- **报名链接：** **连享会-文本分析与爬虫专题**
    - **http://junquan18903405450.mikecrm.com/4EAq2yM**
      或 长按/扫描二维码报名：

1.1 课程提要

- **课程目标：** 在经济和金融等领域，通过网络爬虫和文本分析提取非结构化信息逐渐成为一种重要的手段。本课程旨在帮助学员掌握爬虫、文本分析和机器学习的核心方法和基本分析流程；掌握 `R` 和 `Python` 软件的基本用法，能够做到“左手用 `R`，右手用 `Python`”。
- **核心内容：** 正则表达式相关语法规则及非结构化数据处理； `R` 和 `Python` 的基础知识，并使用 `R` 和 `Python` 进行数据处理、数值计算、网络爬虫、文本分析等不同任务的处理；介绍机器学习常用算法，如决策树、随机森林、支持向量机等的基本原理，并使用 Python 实现各类算法。
- **主要工具/ 方法：** Scipy, Numpy, Pandas, Matplotlib, Plotly, BeautifulSoup, Request, Selenium, Scikit-learn, TensorFlow, Jieba, NLTK, gensim 等。
- **主要软件\语言：** R, Python
- **课程资料：** 开课前一周，将课程资料包发送至学员邮箱。包括：课程讲义，范例程序和数据，以保证大家能够在自己的电脑上重现课程中讲述的所有内容。

1.2 课程特色

- **深入浅出**：掌握最主流的文本分析和爬虫方法
- **电子板书**：全程电子板书演示，课后分享；听课更专注，复习更高效。
- **电子讲义**：分享全套电子版课件 (数据、程序和论文)，课程中的方法和代码都可以快速移植到自己的论文中。

1.3 听课说明

1.3.1 听课软件

本次课程可以在手机，Ipad 以及 Windows 系统的电脑上听课。

> **特别提示：** 一个账号绑定一个设备，且听课电脑 **需要 Windows 系统** ，请大家提前安排自己的听课设备。

1.3.2 实名制与版权保护条款

- 本次课程实行实名参与，高校老师/同学报名时需要向课程负责人提供真实姓名，并提交教师证/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。
- 报名后默认同意版权保护协议条款 （附一个版权保护协议书链接）
- **连享会：版权保护声明**

2. 嘉宾简介

游万海，福州大学经济与管理学院副教授，主要研究领域为空间计量模型、分位数回归模型及相关实际问题的应用，以主要作者在国内外期刊《World Development》、《Energy Economics》、《Economics Letters》、《Finance Research Letters》、《Journal of Cleaner Product》、《Energy Sources, Part B: Economics, Planning, and Policy 》、《统计研究》发表学术论文 30 余篇，担任《Energy Economics》、《Journal of Cleaner Product》、《Economic Modelling》、《International Review of Economics & Finance》等期刊匿名审稿人。

司继春，上海对外经贸大学统计与信息学院讲师，主要研究领域为微观计量经济学、产业组织理论。在 Journal of Business and Economic Statistics、《财经研究》等学术刊物上发表多篇论文，**-主页-**。 其实，大家更熟悉的是知乎上大名鼎鼎的 **慧航**，拥有 30w+ 个关注者，获得过 16w+ 次赞同，他就是司继春老师 —— **知乎-慧航**。

3. 课程简介

3.1 课程导引

实证研究中，数据的 **获取及处理** 是第一步，也是至关重要的一个环节！

在传统的经济和金融分析中，我们使用的主要是结构化的数据 (多数数据来源于统计年鉴、商业数据库，如 GTA，Wind 等)，而在大数据时代，大量有价值的信息以文本等非结构化、异构型的数据格式存储于互联网网页或者各类文档中。从 Web 上快速、有效地提取这些信息对人文社会科学的深度研究尤为重要。

> 这些文本信息如何 **获取**？答曰：**网络爬虫**。
> 这些文本信息如何 **处理**？答曰：**正则表达式**。
> 这些文本信息如何 **分析**？答曰：**话题分析和情感分析**。
> 如何应对变量过多/高维数据问题？答曰：**机器学习**。

事实上，无论是在管理学、经济学、金融学、会计学，还是人文社会科学的其他领域，基于网络爬虫的文本分析的研究越来越广泛。其火热程度可以从相关论文的 Google 学术引用次数窥豹一斑。

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 10 年时间，Google 引用已 3590 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 1100 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。

利用 Google 学术 顺藤摸瓜，可以发现，各个领域文本分析的来源非常丰富，包括：**历史典籍**、报纸新闻、明清房契、**刑科题本**、公司公告、明星微博等等，从而使研究的主题也得到了极大的扩展。

下面是一些我们摘录的一些论文信息：

- 经济学领域：
    - Hoberg, Gerard, and Gordon Phillips. 2016, Text-based network industries and endogenous product differentiation,? **Journal of Political Economy** 124, 1423-1465. Google 引用率超过 500 次。 -PDF1-，-PDF2-
    - Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." **Journal of Economic Literature**, 57 (3): 535-74. -Link-，-PDF-，Google 引用超过 600 次.
- 金融和会计领域:
    - Loughran, Tim, and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks,? **Journal of Finance**?, 66, 35-65. 该文是金融和会计领域颇具影响力的一篇文章，2011 年发表，短短几年的时间，Google 引用率已经超过 1000 次。 -Link-，-PDF1-，-PDF2-
    - Loughran, T., and B. Mcdonald. 2016, Textual analysis in accounting and finance: A survey,? **Journal of Accounting Research**?54, 1187-1230.? -PDF1-，-PDF2-。Google 学术引用率已经飙升到接近 1200 次。
    - Jegadeesh, Narasimhan, and Di Wu. 2013. Word power: A new approach for content analysis,? **Journal of Financial Economics**? 110, 712-729. -PDF1- -PDF2-
    - Neuendorf, Kimberly A, 2017. The content analysis guidebook (Sage). 自 2002 年第一版以来，目前引用率已经达到 15000 余次。
    - Li, K., F. Mai, R. Shen, X. Yan,I. Goldstein, **2021**, Measuring corporate culture using machine learning, **Review of Financial Studies**, 34 (7): 3265-3315. -Link-，-PDF-
    - Li, K., X. Liu, F. Mai, T. Zhang, **2021**, The role of corporate culture in bad times: Evidence from the covid-19 pandemic, **Journal of Financial and Quantitative Analysis**, 56 (7): 2545-2583. -Link-，-PDF-
- 社会学领域：
    - Fairclough, Norman. 2003. Analysing discourse: Textual analysis for social research (Psychology Press). -Link-，被引用次数： 18334
    - Weninger, C., **2020**, Multimodality in critical language textbook analysis, **Language, Culture and Curriculum**, 34 (2): 133-146. -Link-，-PDF-
- 政治学领域：
    - Grimmer, J., B. M. Stewart, **2017**, Text as data: The promise and pitfalls of automatic content analysis methods for political texts, **Political Analysis**, 21 (3): 267-297. -Link-，-PDF1-. Google 引用率超过 2640 次。
    - Grimmer, J., M. E. Roberts, B. M. Stewart, **2021**, Machine learning for social science: An agnostic approach, **Annual Review of Political Science**, 24 (1): 395-419. -Link-，-PDF-

3.2 课程特色

A. 以需求为导向的课程架构

完整的知识架构是你长期成长的动力源泉！ 我们力求此次课程设置成一个比较完整的文本分析体系，以期达成如下目标：

- **其一**，能够实现多数网站数据爬取任务；
- **其二**，能熟练掌握正则表达式对非结构化数据进行清理；
- **其三**，能够建立起文本分析的基本架构，熟知文本分析所涉及的基本概念和分析流程。

B. 为什么要学 Python 和 R?

Python 的易用和流行趋势已经不必多言。Python 在文本分析、爬虫、机器学习等方面有独特优势。作为 Stata 的老用户，连玉君老师建议「左手 Stata，右手 Python」，而李春涛创办的爬虫俱乐部也将其公众号更名为「Stata and Python 数据分析」，足见 Python 的魅力。

R 作为一门免费、开源的语言，已被国外大量学术和科研机构所认可，其被广泛应用于数据挖掘、机器学习、数据可视化、计量经济学和空间统计等领域。正是因为其拥有众多使用者，大量的外部包被开发应用于各个领域 (**19000+** 个，截止 2022.2.12)。刚开始接触 R 语言的学员经常有这样的一个疑问：**为什么 R 体积小，功能却如此多**，这是因为有大量的外部扩展包。

到底学 R 还是 Python? 其实两者是不冲突的！这两门语言存在大量的相似之处，功能都非常强大，其实现都是基于 **基础包(模块) + 外部包(模块)**。 因此，将两者进行对比学习往往能起到事半功倍的效果。

C. 平时用 Stata 较多，听课会有困难吗?

2019 年 6 月 26 日，Stata 官网正式推出 Stata 16。一大亮点在于增加了与 Python 互通的接口：你可以在 Stata 的 Do-file 编辑器中编写 Python 代码，也可以直接从 Stata 界面下直接调用既有的 Python 代码和函数。与此同时，Stata 用户还编写了 `rcall` 等命令，以很方便地在 Stata 环境中执行 R 代码。Stata 在日后的升级版本中很可能会提供 R 接口。

这意味着什么呢? 这意味着我们日后的分析工作都是任务导向的，不再依赖于特定的某种软件或语言。更直接地说，不久之后，每个 PhD 学生都至少要掌握 2-3 种软件和语言，以便应对不同的分析任务。这种改变使得我们得以迈出自己的小圈子，在一个更为广阔的天地中与更多的 **「同行」 分享-被分享** 知识和经验。在 github 浏览一些活跃用户的主页时，你会发现他们通常都会同时使用 2-3 中软件/语言，比如 S. Cunningham，A. Baker，P.H.C. Sant'Anna 都呈现了最新的 DID 估计方法的 R 和 Stata 代码。

为此，本次课程中我们希望能引领各位顺畅地从温馨的 Stata 小屋跨入 Python 和 R 的天地，以便更好地适应未来的发展趋势。

3.3 授课内容

在内容安排上，本着「由浅入深，循序渐进」的原则，在第一部分中 (第 1-4 讲) 我们首先从整体上介绍文本分析的基本框架和分析流程，以便让大家对其分析套路有一个整体上了解。进而从文本处理函数和正则表达式入手，依序介绍爬虫方法、文本信息提取和文本匹配方法。在第二部分中 (第 5-10 讲)，依次介绍 Python 基础，数值运算方法，爬虫进阶 (动态网页的爬取)，最后介绍几种机器学习法算法，并应用于文本情感分析。

各个专题的具体介绍如下：

第 1 讲 介绍文本分析概念及基本流程，并辅以具体实例。从四则运算、基本数据类型和结构、流程控制、函数编写及包的安装与使用等方面介绍 `R` 的基础知识。此外，用一篇具体文章讲解利用 `R` 语言进行数据分析的基本步骤和范式。

第 2 讲 讲解字符处理函数，正则表达式匹配原理和规则，正则表达式的基本语法，包括元字符、序列、数量词、回溯引用、零宽断言、贪婪与懒惰等，真正做到学以致用。

第 3 讲 介绍网页解析中的 XPath 表达式，讲解爬虫的一些基础知识；实现对静态网页和动态网页信息的下载，并利用正则表达式对所爬取的数据进行清理。

第 4 讲 以「深交所问询函」和「上市公司年报」的下载和分析为例，介绍文本文档信息的处理方法；借助正则表达式进行关键词提取，分析文本复杂性、文本相似性；绘制文字云图、提取文本主题等。

第 5-6 讲 介绍 Python 的基本语法，包括基本的数据结构：数字、字符串、元组、列表、词典等，以及常用的控制结构，包括条件、循环、函数等，最后介绍列表推断、Lambda 表达式、类和对象等数据科学中常用的 Python 语法。在此基础上，还会从读取和写入文本文件及 csv 文件、正则表达式在 Python 中的使用以及简单的词频统计等方面介绍 Python 处理字符串的方法。

第 7 讲 为 Python 爬虫专题。首先介绍如何结合 BeautifulSoup、正则表达式等对 HTML 进行解析，提取所需要的有用信息。接着，介绍如何使用 Request、httpx 等工具发送请求，从而爬取静态页面以及简单的动态页面。最后，针对一些比较复杂的动态页面，介绍使用 Selenium 对网页进行动态渲染并进行爬取。此外，对于网站中经常遇到的反爬问题，我们还会介绍一些反爬的思路。最后，我们通过一个实际的综合案例介绍以上所有的技术是如何综合使用并建立一个实时更新的爬虫项目。

第 8 讲 这一讲主要介绍 Python 中的数值计算常用的包，包括使用 Numpy 处理向量、矩阵及其运算；使用 Pandas 整理数据；使用 Matplotlib 以及 Plotly 等工具进行画图等。这部分是进一步学习机器学习、文本分析等内容的重要工具。

第 9 讲 主要包括两部分。第一部分是机器学习的入门，主要介绍监督学习的基础知识、模型评价方法以及常用的避免过拟合的工具等等，在此基础之上，介绍决策树、随机森林、支持向量机、神经网络等常用的监督学习方法，在此基础上，介绍机器学习方法如何与计量经济学相结合，完成回归分析、因果推断等计量经济学的数据分析工作。

第 10 讲 主要包括文本分析的初步知识，如中文分词、词向量表达、词嵌入等技术。在此基础上，介绍文本分类、情感分析等分析方法。这一部分主要介绍自然语言处理的一些常用方法，如词嵌入 (embedding) 的常用方法 word2vec、使用词袋、TF-IDF 等方法计算文本相似性以及话题模型等，最后将对最新的基于神经网络的自然语言处理工具做一个简单的介绍。

4. 课程大纲

Part A：游万海老师 (2 天)

> **参考文献下载**

第 1 文本分析概述 (3 小时)

在大数据时代，大量有价值的信息以非结构化文本数据格式存储于互联网网页或者各类 `PDF` 和 `Word` 文档中，有效利用这些信息的关键是寻找一种快速有效的方法对文本数据进行**爬取和清洗**，进而利用相关方法进行文本分析。

为什么选择 `R` ？其作为自由、免费、开源的软件，是用于统计计算和统计制图的一种有力工具。不管是在学术界还是业界，使用群体均非常庞大。根据 `TIOBE` **编程社区指数**，2022 年 2 月份数据显示，`R` 软件排名第 13 位。

`R` 语言从 2000 年 2 月份的 1.0.0 版本发展到 2021 年 11 月份的 4.1.1 版本，目前大小为 86 M。刚开始接触 R 语言的学员经常有这样的一个疑问：**为什么 R 体积小，功能却如此多**？

`R` 语言和 `Python` 相似，用户们开发了大量的外部包 (18933 个，截止 2022.02.12)，被广泛应用于 **网页爬虫**、数据挖掘、机器学习、数据可视化、计量经济学和空间统计等领域。其运行方式是，核心包提供了基础平台，而大量的外部包可「按需索取」，具有极大的灵活性和扩展性。

- 文本分析含义及使用场景
- 文本分析的基本流程
- 当文本分析与 `R` 语言相遇
    - 软件下载、安装和基本设置
    - 包的管理：安装、加载
    - 数据类型和数据结构
    - 文件管理：本地及网络数据读取和处理；
    - 应用-基于 Twitter 的幸福指数提取：Abubakr Naeem 等 (2020, International Review of Finance)、Lee 等 (2020, Tourism Economics)
    - 流程控制函数: `if` 、 `for` 、 `while` 等
    - 常用功能包介绍： **图形绘制**、**计量模型**、**网页爬虫**、**机器学习**、**自然语言处理**、**贝叶斯推断** 等。
- 范例讲解：王分棉等 (2021, **中国工业经济**)

第 2 讲 文本清洗与正则表达式 (3 小时)

与结构化数据不同，文本数据显得很不“友好”，在使用之前需要进行一系列预处理，从复杂文本中抽取有用信息。这其中就包括**文本定位**、**文本匹配**、**文本抽取**等操作。但是，在进行这些操作之前需要定义**具体的规则**，如阿拉伯数字 `0-9` 、小写字母 `a-z` 、小写元音字母 `aeiou` 等如何表示，这就是正则表达式语法。

例如：在 2021 **年第三季度报告（A 股）** 中如何快速抽取**客户存款总额**和**客户贷款总额**？又如，在王分棉等 (2021, **中国工业经济**) 一文中，如何从豆瓣电影中抽取每一部电影的评分？通过此次讲解，力争使学员能够系统掌握正则表达式的语法，并且能够独立进行文本信息的提取。

- 文本文档读取
- 字符串函数：文本信息定位、文本匹配、关键词抽取、文本清洗等
- 正则表达式匹配原理和示例讲解
- 正则表达式语法：元字符、字符类、数量词、贪婪懒惰模式、位置匹配符、回溯引用等
- R 实操：范文 1 篇，王分棉等 (2021, **中国工业经济**)

第 3 讲 网络爬虫原理与 `R` 实例 (3 小时)

网络爬虫是将呈现在网页上以非结构格式存储的数据转化为结构化数据的技术，中间涉及到“抓取”、“解析”、“储存”三个阶段。

网络爬虫可以分为静态网页和动态网页爬虫。静态网页就是当打开一个网页，可以直接看到想抓取的数据，网址固定后，页面显示的数据也是固定的，如果要加载更多的数据，必须跳转到其他网址。动态网页就是想要爬取的内容不在源代码中，网址固定后，页面显示的数据并非固定的，通过“下拉”、点击“下一页”、点击“更多”等，以此加载更多数据。

对于静态网页或者动态网页渲染后，可以在源代码中看到数据，一种方法是利用第 2 讲知识，下载页面后利用正则表达式进行清洗，过程较为复杂；另一种方法是利用 `R` 中一些现成的包，如 `rvest` 提供一套较为完整的数据抓取方案，配合 `SelectorGadget` ，可快速实现网页爬虫。

- 网页相关知识介绍
- 网络爬虫技术介绍
- 静态网页爬虫
  - `Xpath` 表达式、 `CSS` 选取器 `SelectorGadget` 、 `rvest` 包
  - R 实操 1：豆瓣电影 **top250** 信息下载 (电影名称、评分、上映年份、主演及导演等)
    - 参考文献：王分棉等 (2021, **中国工业经济**)
  - R 实操 2： **中国银行外汇牌价**
- 动态网页爬虫简介与 `R` 实例

第 4 讲 基于文本语义分析的主题提取和变量构造 (3 小时)

现如今，大量的信息以文本、图片、表格等形式存贮于各类文档 (`txt, word, pdf`) 中。

例如：年报问询函作为监管机构监督上市公司信息披露的重要机制，其是否具有信息含量？ **张俊生等 (2018)** 研究公司是否被交易所发布年报问询函(被问询为 1，否则为 0)对公司股价崩盘风险的影响。 **陈运森等(2018)** 研究问询函特征 (问询函问题数量、问询函的问题内容等、公司是否回函)的短期市场反应。相似的研究还有陈运森等 (2019, **管理世界**)、李晓溪等 (2019, **经济研究**)。

又如，上市公司年报是否具有信息含量？ **李哲 (2018)** 从公司年报和独立社会责任报告中抓取环境战略信息和环境行动信息相关词条，以词条出现次数作为环境信息披露的频数，考察环境信息披露对交易者和监管者行为的影响。李哲等 (2022, **财经研究**)在公司年报中抓取有关环境规划和环境行动的词频数，作为企业环境责任表现“言”和“行”的度量。

- 基于正则表达式的文本信息挖掘：关键词提取、文本复杂性分析、文本相似性计算、文本主题提取等
    - R 实操 1：深交所问询函爬取与分析
        - 参考文献：陈运森等 (2019, **管理世界**)、陈运森等 (2018, **金融研究**)、李晓溪等 (2019, **经济研究**)
    - R 实操 2：上市公司年报下载与分析：上市公司参与环境治理的"言"和"行"
        - 参考文献：李哲等 (2022, **财经研究**)

除了上述形式，一些数据以表格形式存贮在文档中，如何进行有效提取尤为重要。例如，论文附录中的表格数据提取。

- `PDF` 表格数据提取与清洗
- 图片形式数据提取与清洗
- 文本情感分析
- 课程总结

参考文献：

- 陈运森,邓祎璐,李哲.证券交易所一线监管的有效性研究:基于财务报告问询函的证据. **管理世界**, 2019 (3): 169-185. -Link- ， -PDF-
- 陈运森,邓祎璐,李哲. 非处罚性监管具有信息含量吗?——基于问询函的证据. **金融研究**, 2018 (4): 155-171. -Link- ， -PDF-
- 李晓溪, 杨国超, 饶品贵. 交易所问询函有监管作用吗?——基于并购重组报告书的文本分析. **经济研究**. 2019 (5). -Link- ， -PDF-
- 王分棉,任倩宜,周煊. 生态位宽度、观众感知与市场绩效——来自中国电影市场的证据. **中国工业经济**, 2021(11):155-173. -PDF-
- 李哲,王文翰,王遥. 企业环境责任表现与政府补贴获取——基于文本分析的经验证据. **财经研究**, 2022,48(02):78-92+108. -PDF-
- Naeem M A, Mbarki I, Suleman M T, et al. Does Twitter happiness sentiment predict cryptocurrency?. **International Review of Finance**, 2021, 21(4): 1529-1538. -Link- ， -PDF-
- Lee C C, Chen M P, Peng Y T. Tourism development and happiness: International evidence].. **Tourism Economics**, 2021, 27(5): 1101-1136. -Link- ， -PDF-
- You W, Guo Y, Peng C. Twitter's daily happiness sentiment and the predictability of stock returns. **Finance Research Letters**, 2017, 23: 58-64. -Link- ， -PDF-
- Elbahnasawy N G. Can e-government limit the scope of the informal economy?. **World Development**, 2021, 139: 105341.
- Canh P N, Schinckus C, Dinh Thanh S. What are the drivers of shadow economy? A further evidence of economic integration and institutional quality. The Journal of International Trade & Economic Development, 2021, 30(1): 47-67. -Link- ， -PDF-

Part B：司继春老师 (3 天)

> **参考文献下载**

第 5 讲 Python 基础 (3 小时)

这一讲主要目标是在短时间内对 Python 的使用有一个初步的了解。Python 与 R、Stata 相比最大的区别在于，Python 是一门通用编程语言而非统计语言，所以在实际编程中很多时候需要以计算机语言的思想进行，因而不管从语法层面还是编程技巧、思想、习惯方面，与 Stata、R 等有较大区别。

在这一讲，我们先从最基础的数据结构入手，介绍 Python 的基础语法，同时包括文本文件、csv 文件等的读取操作等。

- Python 数据类型和基本运算：数值、字符串、列表、元组、字典等
- Python 控制结构：条件、循环
- Python 中的文件管理：文本文件、csv 文件、二进制文件等
- Python 语法补充：列表推断

第 6 讲 Python 使用进阶 (3 小时)

作为一门编程语言，“抽象”是最强大也是最难理解的概念，可以说现代编程语言都是建立在一定程度上的“抽象”的基础上的，没有“抽象”，编程将会异常复杂且难以调试。

在这一讲，我们主要介绍 Python 的两种抽象方式：函数和对象。前者将特定的功能模块抽象化，后者将具有属性、行为的组件抽象化，这些都是接下来使用 Python 进行爬虫、数据分析的基础。

- Python 中的函数和 Lambda 表达式
- Python 中的面向对象编程：类和对象
- Python 中的字符串函数与正则表达式
- 一种特殊的数据结构：队列 Queue
- 案例：使用 Python 进行简单的中文分词和词频统计

第 7 讲 Python 爬虫专题 (3 小时)

由于 Python 是一门通用编程语言，同时是一门胶水语言，其巨大的灵活性使其能够处理各种爬虫上遇到的疑难问题，然而也意味着使用 Python 进行爬虫程序的编写也尤其独特的困难。

在前序课程对 HTML、HTTP 等有了初步了解后，本节主要介绍完成一个爬虫的标准流程：使用 request/httpx 发送 http 请求、使用 BeautifulSoup 或者正则表达式解析 HTML 并获得需要的关键信息，最终进行存储。此外，还将介绍 Selenium 爬取动态页面的方法，以及常见的反爬方法。

最后，作为综合实例，我们介绍两个爬虫的实际应用：财经专业词汇词典爬取、上市公司公告爬取。

- 使用 BeautifulSoup 和正则表达式解析 HTML
- 使用 request/httpx 发送 Http 请求
- 使用 Selenium 爬取动态网页
- 反爬常用方法
- 案例
    - 财经专业词汇词典爬取
    - 上市公司公告爬取
- 参考文献：

    - 山立威, 甘犁, 郑涛. "公司捐款与经济动机——基于汶川地震后中国上市公司捐款的实证研究?". 经济研究 8 (2008): 38-42. -PDF-
    - 罗进辉. "媒体报道与高管薪酬契约有效性." 金融研究 453.3 (2018): 190-206. -PDF-

第 8 讲 Python 中的数值计算 (3 小时)

由于 Python 不是专门的统计软件，所以没有原生的数据存储、处理的工具，也没有进行科学计算的原生工具，为此，Numpy/Scipy/Pandas 等工具应运而生，解决了 Python 进行数据分析的基础设施建设问题。

这一讲在 Python 的基础语法基础上，继续介绍如何在 Python 中进行简单的科学计算和数据处理，并简要介绍 Python 中图表的制作。这一节的内容是后续机器学习、文本分析的基础。

- Numpy 基础：向量与矩阵的操作
- Numpy 进阶：向量计算与切片
- Matplotlib 简介：直方图、散点图、线图等图形的绘制
- Pandas 基础：序列与数据框
- Pandas 进阶：切片、数据操作和处理
- 案例
    - 文本数据的词频统计、模式匹配统计
- 参考文献
    - 王雄元, 高曦. "年报风险披露与权益资本成本." 金融研究 451.1 (2018): 174-190. -PDF-

第 9 讲 机器学习初步 (3 小时)

随着大数据的发展，机器学习的应用已经拓展到了社会经济的方方面面，而在学术领域，机器学习方法也被逐步应用到经济学、社会学等的研究中。

这一节主要从机器学习的基础入手，讲解机器学习的目标以及所面临的关键问题，并区分监督学习和无监督学习对机器学习做个全面的介绍，并展示如何使用 scikit-learn 进行机器学习模型的训练。此外，我们还将简要介绍机器学习与计量经济学方法的结合，如因果树、Double machine learning 以及工具变量中 Lasso 等方法的应用。

- 机器学习概述：评价模型的标准
- 机器模型概述：交叉验证和正则化
- 无监督学习：聚类和嵌入 (embedding)
- 监督学习：线性回归、Logistic 回归、决策树、随机森林
- 机器学习与因果推断的结合简介
    - 因果树（森林）
    - Double machine learning
    - Lasso 在工具变量估计中的应用
- 案例
    - Lasso 与回归控制法的应用
- 参考文献：

- Athey, Susan, and Stefan Wager. "Estimating treatment effects with causal forests: An application." **Observational Studies** 5.2 (2019): 37-51. -PDF-
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21: C1-C68. -PDF- , -PDF2-
- Knaus, Michael C. "A double machine learning approach to estimate the effects of musical practice on student's skills." **Journal of the Royal Statistical Society: Series A** (Statistics in Society) 184.1 (2021): 282-300. -PDF-
- Gilchrist, Duncan Sheppard, and Emily Glassberg Sands. "Something to talk about: Social spillovers in movie consumption." **Journal of Political Economy** 124.5 (2016): 1339-1382. -PDF-

第 10 讲 文本分析 (3 小时)

得益于丰富的软件包和灵活的语法，Python 对于文本的处理是非常高效且灵活的。

本节将介绍常见的文本分析方法在 Python 中的实现。结合 Python 强大的文本分析工具，如 NLTK、Jieba 等工具，结合 Pandas、Nump 以及 Scikit-learn 等工具完成文本词嵌入、文本相似度计算、文本分类、情感分析等不同的文本分析任务。

- 词袋模型、TF-IDF 的原理和实现
- 词嵌入模型的使用
- 文本相似性的计算
- 情感分析：情感词典方法
- 情感分析：监督学习法
- 案例
  - 情感词典、监督学习与文本分类
- 参考文献：
  - Piotroski, Joseph D., T. J. Wong, and Tianyu Zhang. "Political bias in corporate news: The role of conglomeration reform in China." The Journal of Law and Economics 60.1 (2017): 173-207. -PDF-
  - 汝毅, 薛健, 张乾. "媒体新闻报道的声誉溢出效应." 金融研究 470.8 (2019): 189-206. -PDF-
  - 阮睿, 孙宇辰, 唐悦, 聂辉华. "资本市场开放能否提高企业信息披露质量?——基于 “沪港通” 和年报文本挖掘的分析." 金融研究 488.2 (2021): 188-206. -PDF-

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)：4500 元/人 (全价)
- **团报 (三人及以上)**：9 折，4050 元/人
- **老学员：** 现场班及专题课学员：9 折，4050 元/人
- **学生**：9 折，4050 元/人，提供学生证照片/学生卡照片
- **会员**：8.5 折，3825 元/人 (可申请成为会员)
- **Note：** 以上各项优惠不能叠加使用。
- **联系方式：**

- 邮箱：wjx004@sina.com
- 王老师：18903405450 (微信同号)
- 李老师：18636102467 (微信同号)

报名链接

> **报名链接：** http://junquan18903405450.mikecrm.com/4EAq2yM
> 或 长按/扫描二维码报名：

扫码报名
连享会：文本分析-爬虫-机器学习

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. 助教招聘

说明和要求

- **名额：** 8 名
- **任务：**
  - **A. 课前准备**：协助完成 3 篇推文，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)；参见 **往期答疑**。
  - 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 R 或 Python 的基本语法和常用命令，能对常见问题进行解答和记录。优先考虑熟悉 R 或 Python 的申请人。
- **特别说明：** 往期按期完成任务的助教自动获得本期助教资格，不必填写申请资料，直接联系连老师即可。
- **截止时间：** 2022 年 3 月 12 日 (将于 3 月 14 日公布遴选结果于连享会主页： lianxh.cn)

申请链接： https://www.wjx.top/jq/45760506.aspx

扫码填写助教申请资料：

课程主页： https://gitee.com/arlionn/text

连 享 会

www.lianxh.cn

中山大学连玉君老师团队，为您分享 Stata, Python, R 应用经验。

专题夯实基础，推文紧追前沿。

信心是通过一个个小进步积累起来的！

搜索公众号ID：lianxh_cn

或长按/识别二维码关注