南京工業大學
NANJING TECH UNIVERSITY

用户数据采集与关联分析

（结课作业）

任俊杰

第一讲 课程导言与分词

1. 学习使用在线NLPIR分词系统或微词云分词或清华大学分词演示系统（案例演示截图）；
2. 安装python（anaconda）（编写输出“Hello World. Hello ‘你的姓名’”）；
3. 完成课后作业（**001-004，4**份代码的运行）。

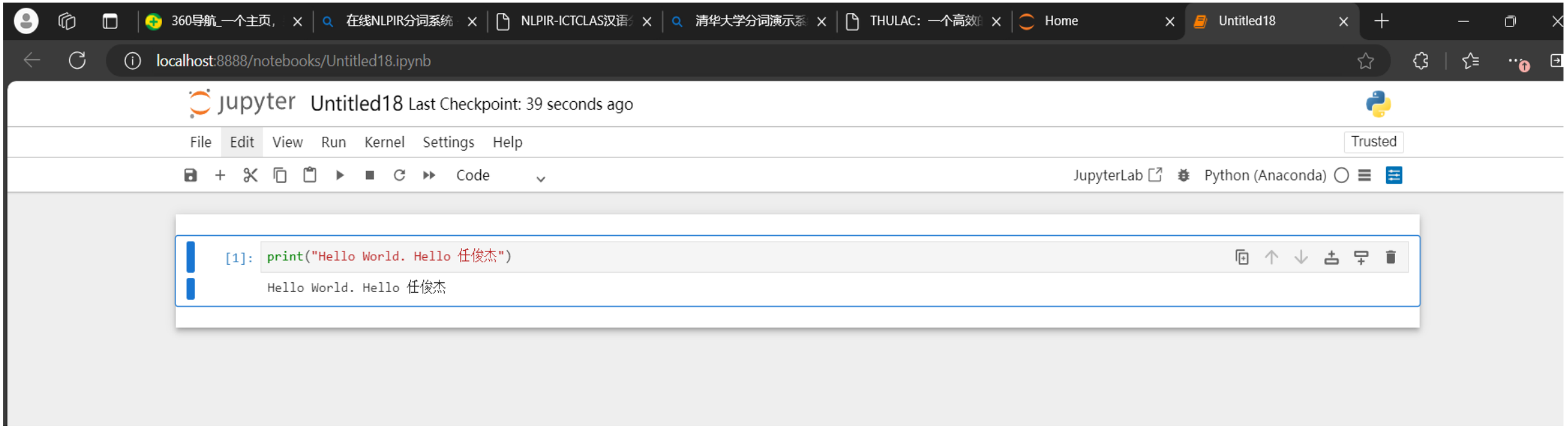
4. 阅读压缩文件中（“实体抽取论文-换成PDF”）中的其中一篇论文，并做阅读总结（1页PPT即可）（仅信管）。
5. 谈一谈在营销学科/领域，文本、文本分词以及实体的内涵。例如：客户关系管理中，文本分析的价值。（仅营销）

第一讲 课程导言与分词

1. 学习使用在线NLPIR分词系统或微词云分词或清华大学分词演示系统（案例演示截图）；

第一讲 课程导言与分词

2. 安装python（anaconda）（编写输出“Hello World. Hello ‘你的姓名’”）；



第一讲 课程导言与分词

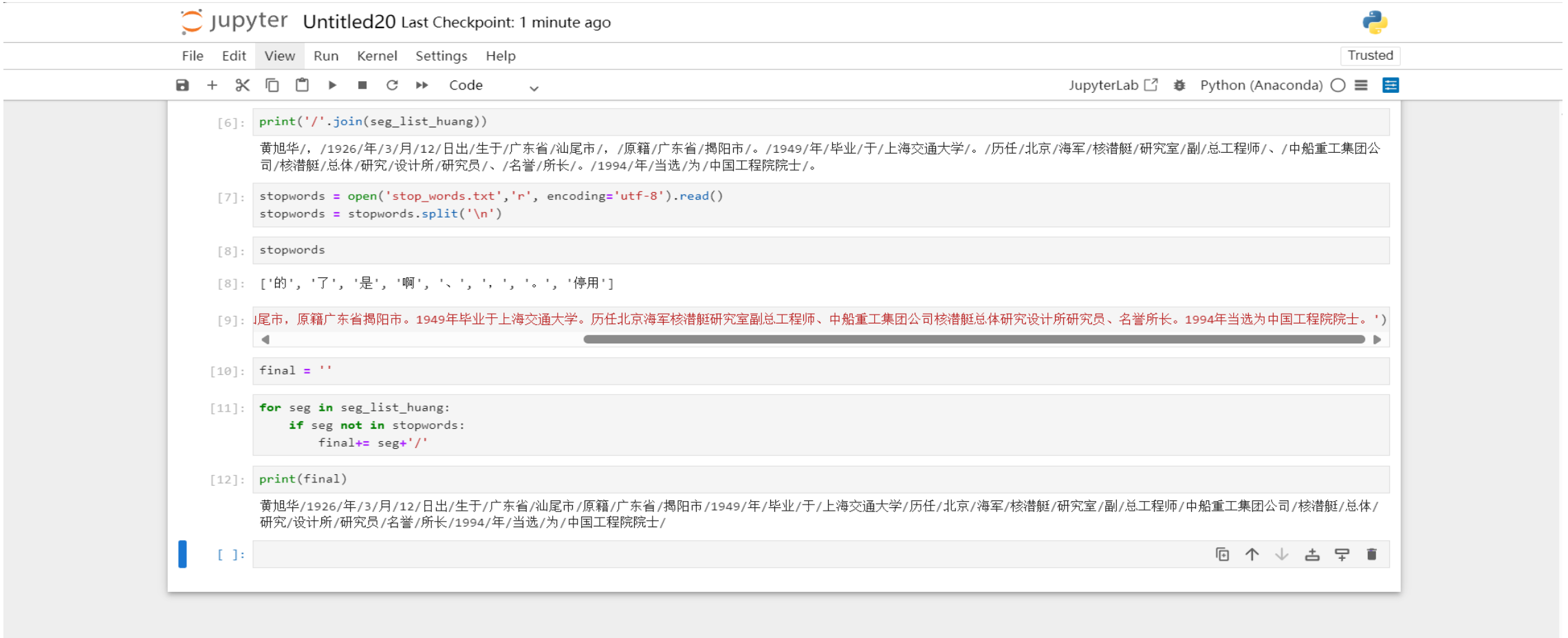
3. 完成课后作业（**001-004，4份代码的运行**） 001；

第一讲 课程导言与分词

3. 完成课后作业（**001-004，4份代码的运行**） 001；

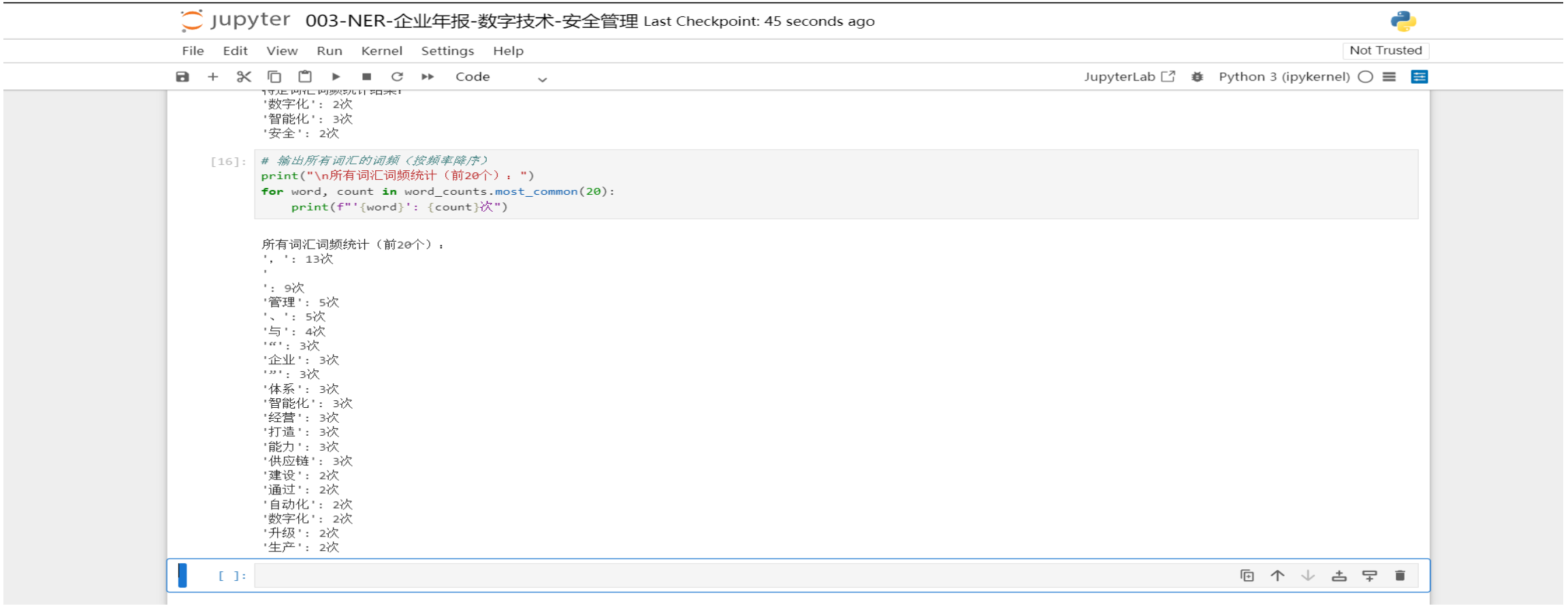
第一讲 课程导言与分词

3. 完成课后作业（**001-004，4份代码的运行**） 002；



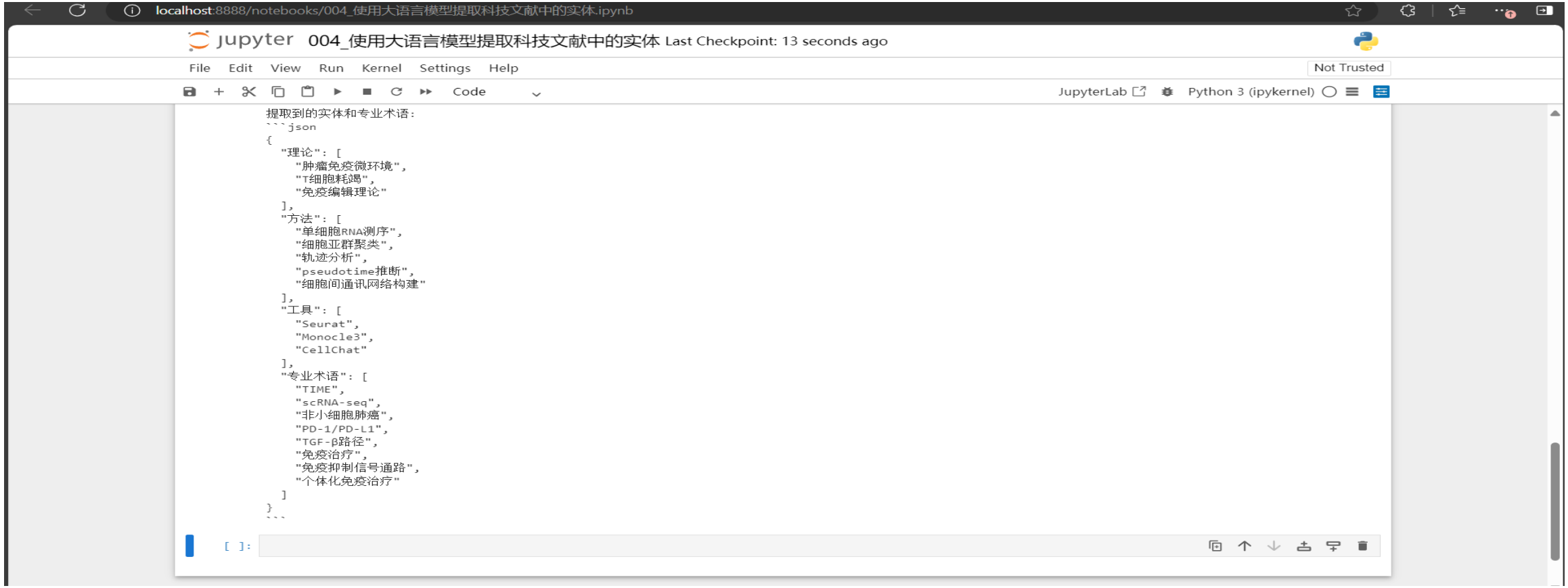
第一讲 课程导言与分词

3. 完成课后作业（**001-004，4份代码的运行**） 003；



第一讲 课程导言与分词

3. 完成课后作业（**001-004，4份代码的运行**） 004；



第一讲 课程导言与分词

4. 阅读压缩文件中（“实体抽取论文-换成PDF”）中的其中一篇论文，并做阅读总结（1页PPT即可）（仅信管）。

阅读总结：

结论：

全文内容显著优于摘要：所有模型在全文数据上的表现均远超仅使用摘要的基线。最好的模型（CC-NB）在全文上的F1值达到0.705，而摘要仅为0.656。这有力证明了全文包含的丰富上下文信息对于精准识别研究方法至关重要。

模型性能差异：在所有模型中，朴素贝叶斯在分类器链（CC-NB）策略下表现最佳。这表明在考虑标签间潜在关联性的同时，NB算法在本任务中具有良好的适应性。相比之下，ML-KNN模型表现一般，暗示研究方法间的强依赖性可能不显著。

类别不平衡与特征表征能力的影响：研究发现，不同研究方法的分类效果差异巨大。样本量大且特征鲜明的“实验法”（F1=0.836）和“计量法”（F1=0.803）分类效果极佳；而“内容分析法”（F1=0.460）和“其他方法”（F1=0.443）则表现较差。这揭示了两个核心挑战：一是训练数据规模不足会导致泛化能力弱；二是不同研究方法本身的语言特征表征能力不同，有些方法（如内容分析）缺乏独特、稳定的关键词汇。

第二讲 词频统计

1. 基于CNKI数据库统计分析2014-2024年（近10年），“信息资源管理”或“网络营销”或其他你感兴趣的主题变化趋势。
2. 完成ppt中的程序运行，包括全文词频统计，指定类型词频统计；
3. 链接功勋科学家：把ppt中的文本换成功勋科学家黄旭华院士的传记序言文本（文件夹中，科学家博物馆-黄旭华传记序言.txt），1）统计全文词频；2）统计指定词频，如“黄旭华”；
4. 阅读论文“2018-Wang 等 - Long live the scientists Tracking the scientific”，并做阅读总结（1页PPT即可）。

第二讲 词频统计

1.基于CNKI数据库统计分析2014-2024年（近10年），“信息资源管理”或“网络营销”或其他你感兴趣的主题变化趋势。

一、从“传统文献管理”向“数字与数据驱动”转型（2014–2017）

此阶段研究仍以图书馆学、情报学和档案学三大支柱为基础，但已明显转向数字化议题。关键词如“数字图书馆”“元数据”“信息组织”“知识服务”高频出现。研究重点在于如何将纸质文献资源转化为结构化数字资产，并构建统一的资源描述与检索体系。同时，“大数据”概念开始渗透，部分学者尝试探讨信息资源在政府开放数据、智慧城市等场景中的整合路径。

二、学科融合加速，“健康信息学”“数据科学”兴起（2018–2020）

随着《“健康中国2030”规划纲要》等政策推动，健康信息学成为信息资源管理的重要分支。CNKI数据显示，关于“医院信息资源”“医疗知识库”“CHDK医药总库”等主题的论文显著增长。与此同时，数据管理与数据治理成为新热点，“数据资产”“数据生命周期”“数据质量”等术语频繁出现。学科边界进一步模糊，信息资源管理开始与公共卫生、生物医学、人工智能等领域深度交叉。

第二讲 词频统计

1.基于CNKI数据库统计分析2014-2024年（近10年），“信息资源管理”或“网络营销”或其他你感兴趣的主题变化趋势。

三、技术赋能深化，“AI+知识管理”成为核心范式（2021–2023）

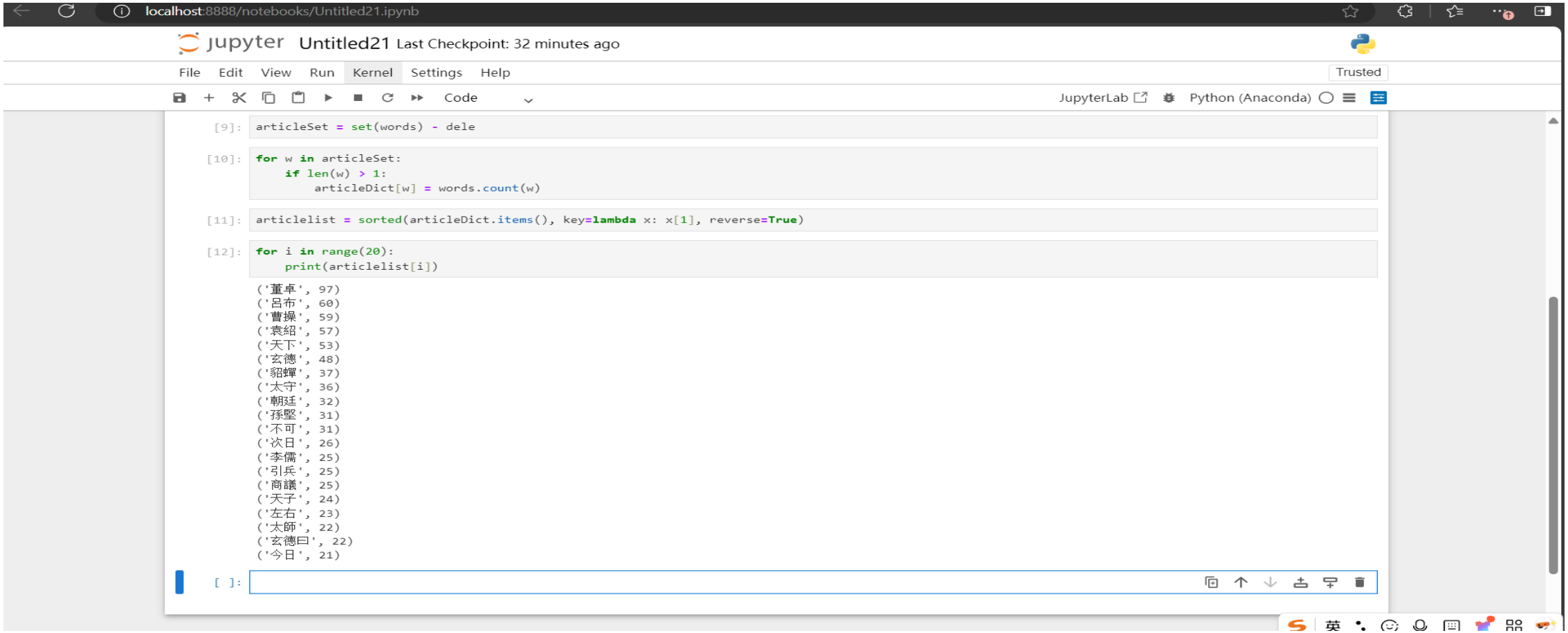
人工智能、自然语言处理和知识图谱技术被广泛应用于信息资源的智能组织与服务。研究焦点从“资源存储”转向“知识发现”与“智能推荐”。例如，利用BERT等模型对学术论文进行研究方法自动分类（如章成志等，2020），或构建中医药方剂知识网络（依托《中国医药卫生知识资源总库》）。此外，“信息素养教育”内涵扩展，新增“数据素养”“健康信息素养”“元素养”等子维度，反映社会对公民信息能力的新要求。

四、战略价值凸显，迈向“国家信息资源体系”建设（2024至今）

近年研究更强调信息资源的国家战略属性与制度化治理。关键词如“国家数据局”“数据要素市场”“信息资源资产化”“公共数据授权运营”集中涌现。2023年国家数据局成立后，学界迅速响应，探讨如何构建覆盖政府、企业、科研机构的全域信息资源协同管理体系。同时，信息资源管理被纳入新文科、交叉学科建设框架，其作为一级学科的地位得到强化，研究视野从技术操作层面上升至制度设计与生态构建层面。

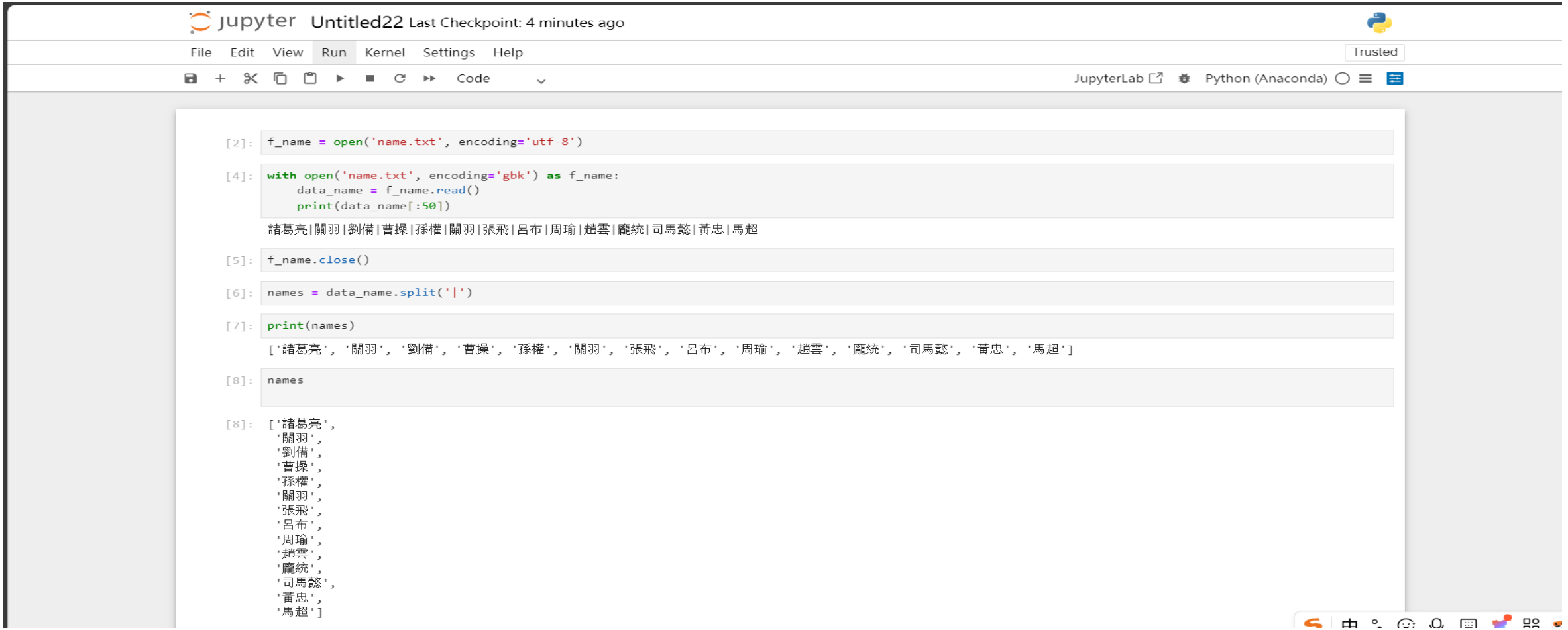
第二讲 词频统计

2.完成ppt中的程序运行，包括全文词频统计，指定类型词频统计*全文词频统计



第二讲 词频统计

2.完成ppt中的程序运行，包括全文词频统计，指定类型词频统计*指定类型词频统计

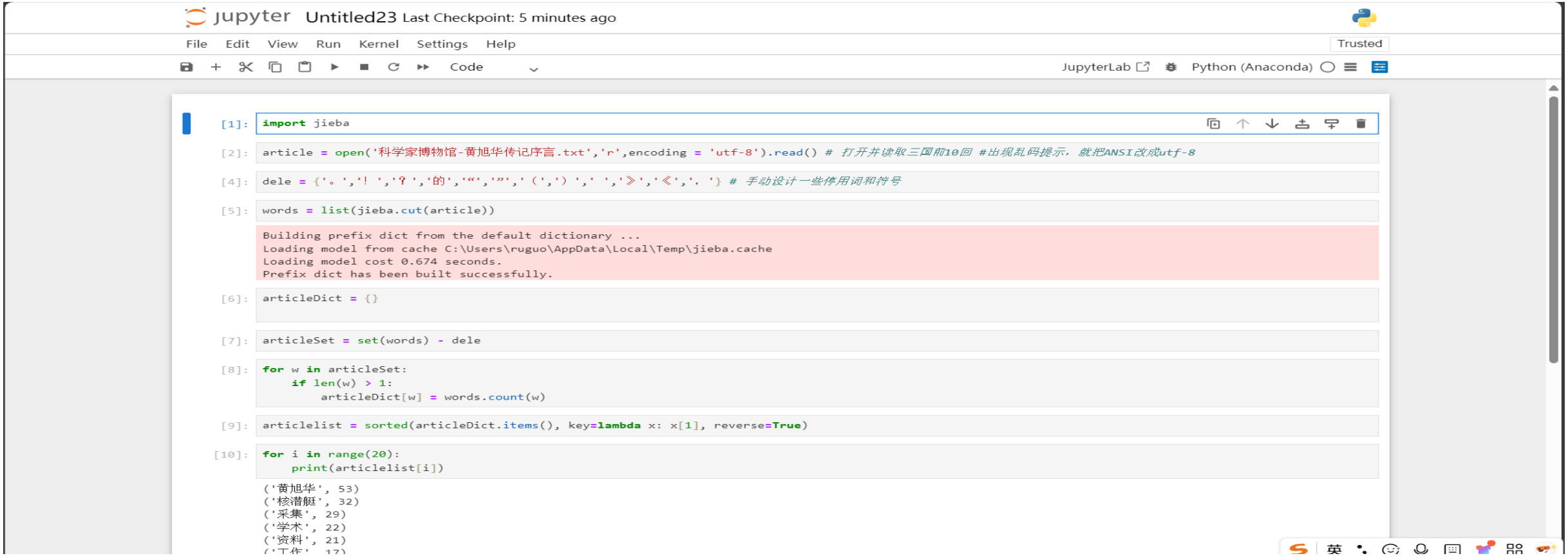


第二讲 词频统计

2.完成ppt中的程序运行，包括全文词频统计，指定类型词频统计*指定类型词频统计

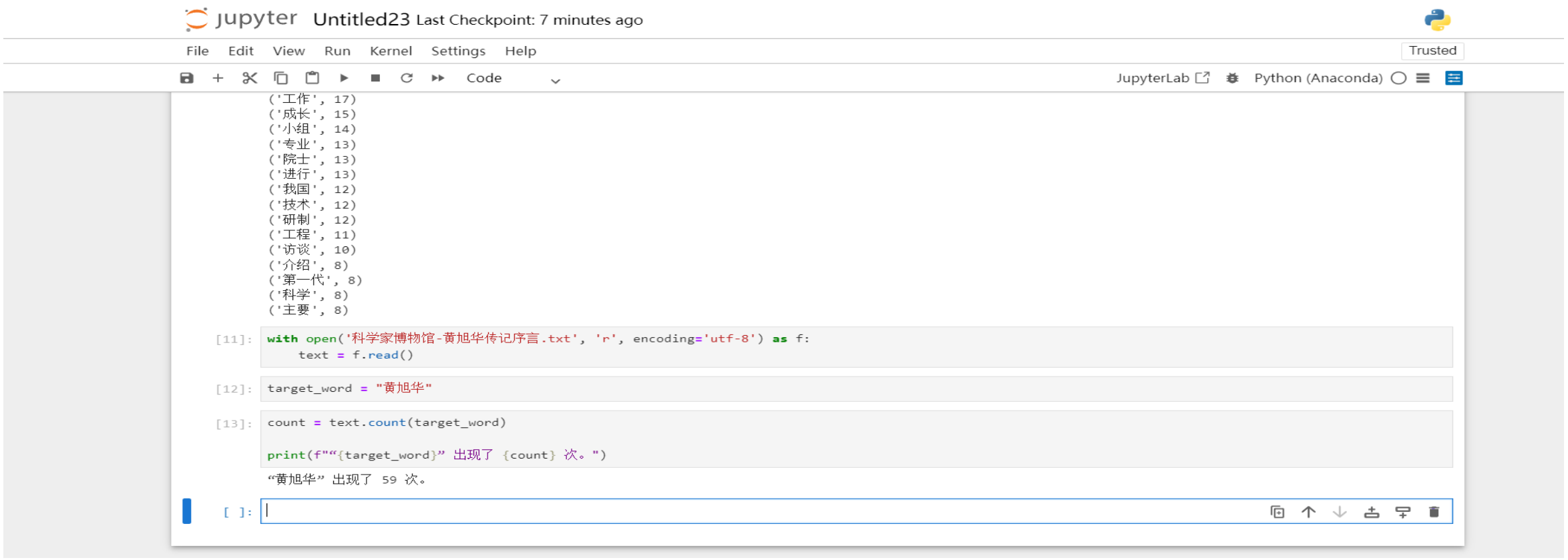
第二讲 词频统计

3.链接功勋科学家：把ppt中的文本换成功勋科学家黄旭华院士的传记序言文本（文件夹中，科学家博物馆-黄旭华传记序言.txt），1）统计全文词频；2）统计指定词频，如“黄旭华”；*（1）



第二讲 词频统计

3.链接功勋科学家：把ppt中的文本换成功勋科学家黄旭华院士的传记序言文本（文件夹中，科学家博物馆-黄旭华传记序言.txt），1）统计全文词频；2）统计指定词频，如“黄旭华”；*（2）



第二讲 词频统计

4.阅读论文“2018-Wang 等 - Long live the scientists Tracking the scientific”，并做阅读总结（1页PPT即可）

阅读总结;

核心研究方法与数据：

研究创新性地结合了两种数据源：一是覆盖3600万册图书的谷歌图书，用于衡量科学家在公共及文化层面的知名度（通过其姓名在书籍中的出现频率）；二是索引9100万篇学术文献的谷歌学术，用于衡量其在学术界的影响力（通过被引次数）。此外，研究还利用谷歌Ngram Viewer按不同语言（如英式英语、美式英语、德语等）进行细分分析，并通过共现分析探究科学家与其标志性成就的关联度。

主要研究发现：

1. “伟人虽逝，声名永存”：研究证实，历史上最伟大的科学家并未被后世遗忘。他们的影响力跨越数百年，持续存在于人类的文化记忆中。例如，牛顿在17-19世纪享有极高声望，而爱因斯坦自20世纪中叶起在全球范围内的提及度已超越牛顿。

2. “同群偏好”效应显著：研究发现了强有力的“同群偏好”（own-group preference）证据。科学家在其本国或使用相同语言的群体中享有更高声誉。最典型的例子是，在英式英语语料库中，牛顿的提及度始终高于爱因斯坦，这与2005年英国皇家学会的民意调查结果一致；而在美式英语、德语等语料库中，爱因斯坦则更受推崇。

3. 声望源于标志性成就：共现分析揭示了科学家声望的具体来源。牛顿的声望主要与其万有引力定律和运动定律相关；而爱因斯坦的声望则高度集中于相对论（尤其是广义相对论）和量子理论。

第三讲 词云与可视化

1. 用任意一款词云工具，制作一个好看的词云（内容合理即可），并对词云图有一段话的解释。
2. 使用Echarts，制作3个以上图，其中一个必须是“关系”，图的概念越明确（可解释，而不是自带的模板）越好。
3. 使用Gehpi、VOSViewer、CiteSpace…其中任意一款工具，绘制任意你感兴趣的图谱1-2张。
4. 采用给的程序，实现一段科学家文本的词云图绘制，越清晰越好（生成的词云图要单独拿出来）。

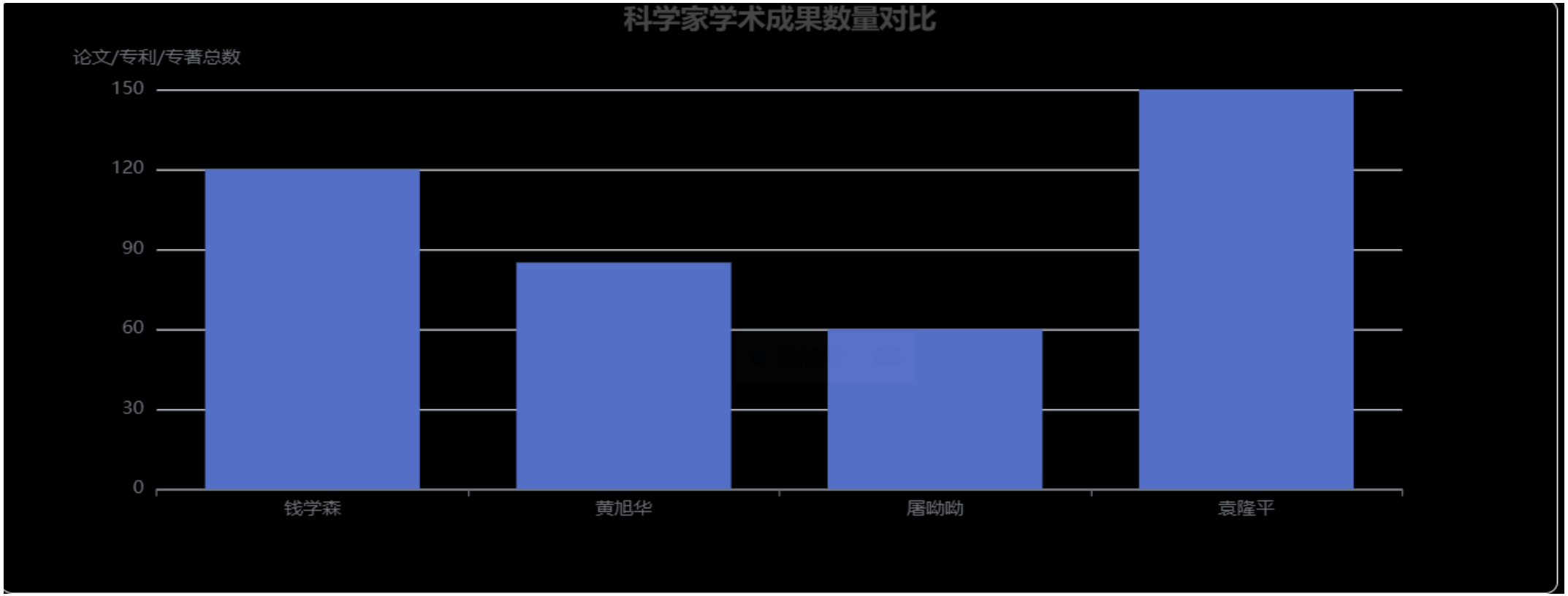
第三讲 词云与可视化

1. 用任意一款词云工具，制作一个好看的词云（内容合理即可），并对词云图有一段话的解释。

一段关于人格分裂的解释的文字生成的词云图中人格占最大最中间，说明这段话中“人格”这个词的占比最大。其他的词例如“患者”“差异”等都是文字中出现词频较少的词

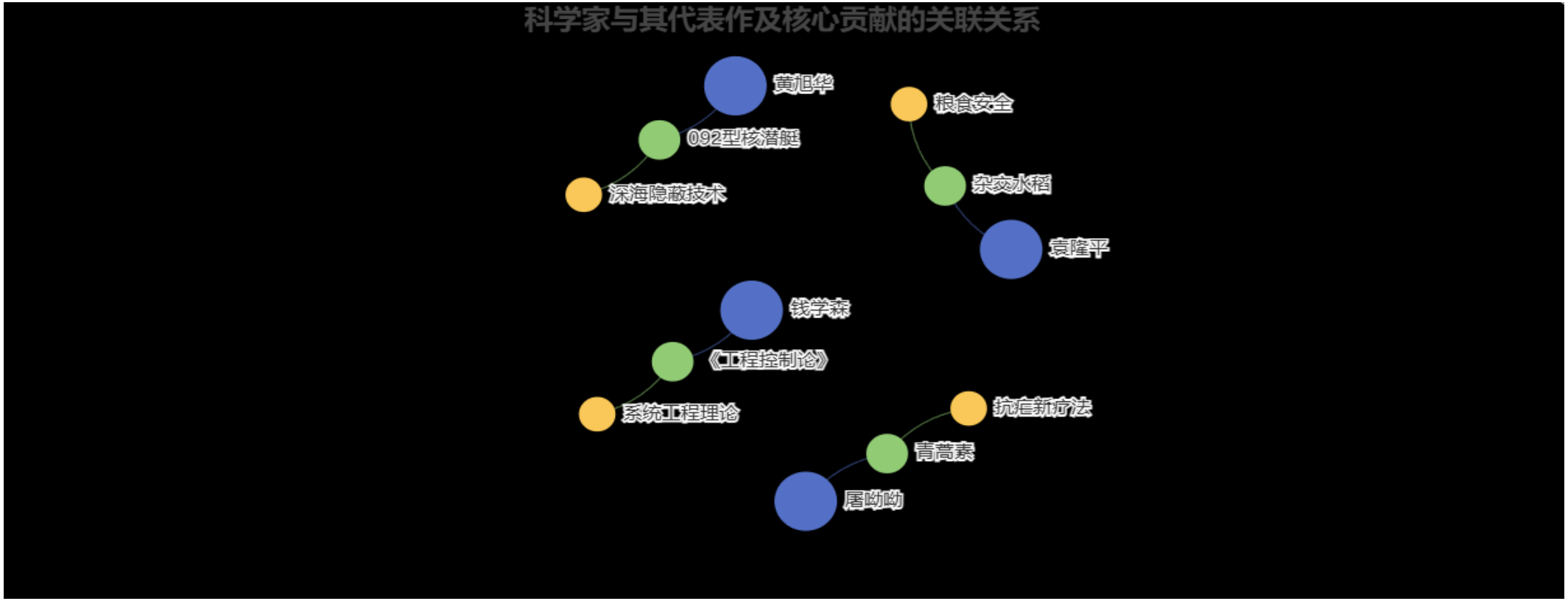
第三讲 词云与可视化

2. 使用Echarts，制作3个以上图，其中一个必须是“关系”，图的概念越明确（可解释，而不是自带的模板）越好。



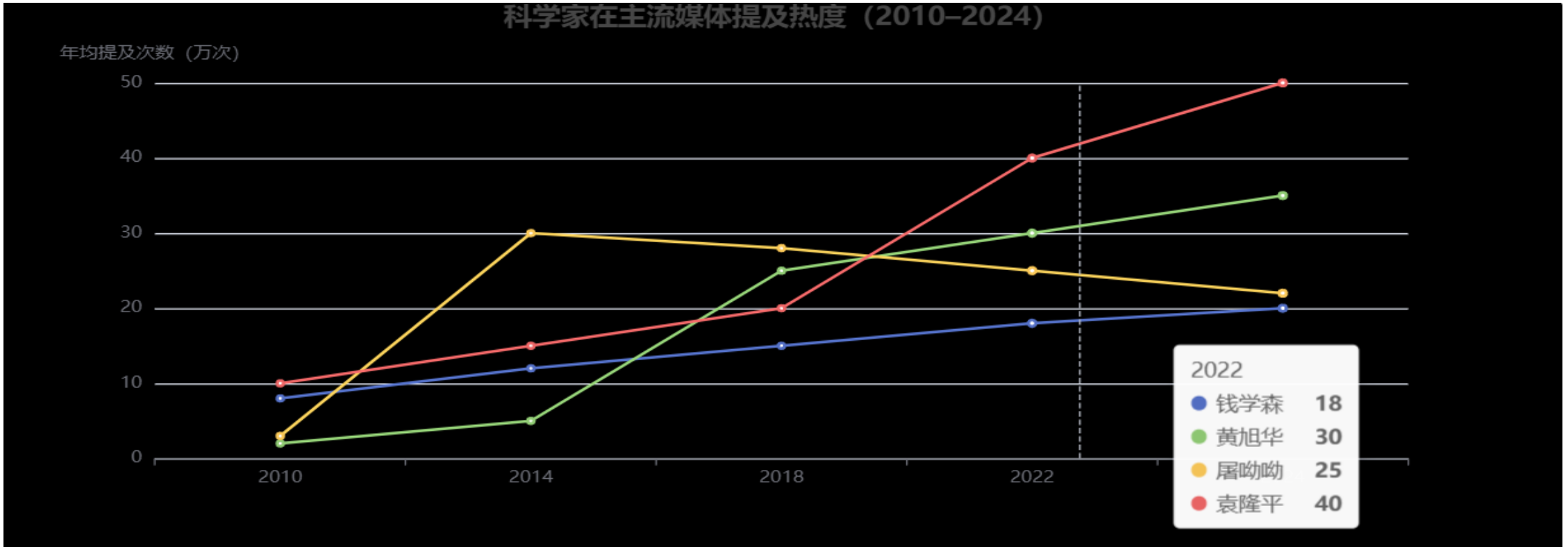
第三讲 词云与可视化

2. 使用Echarts，制作3个以上图，其中一个必须是“关系”，图的概念越明确（可解释，而不是自带的模板）越好。



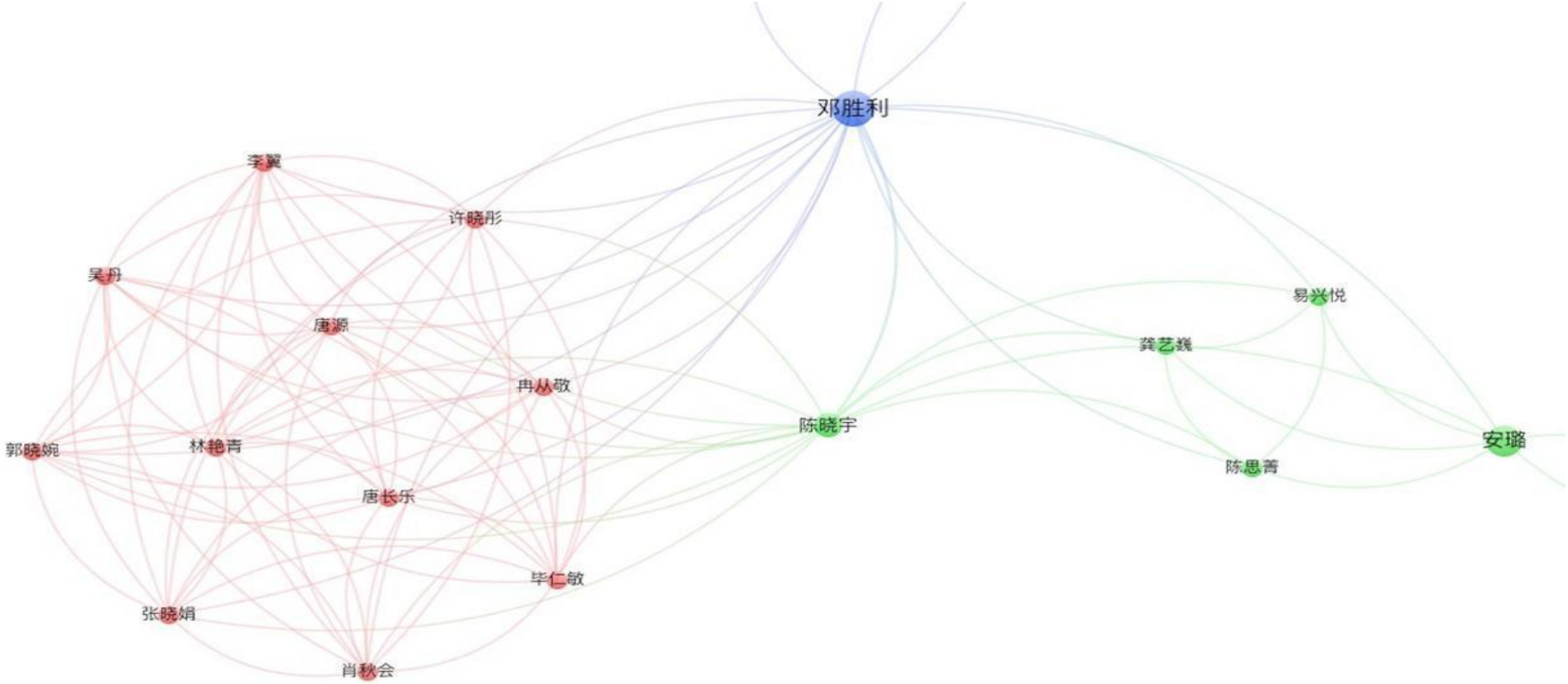
第三讲 词云与可视化

2. 使用Echarts，制作3个以上图，其中一个必须是“关系”，图的概念越明确（可解释，而不是自带的模板）越好。



第三讲 词云与可视化

3.使用Gehpi、VOSViewer、CiteSpace…其中任意一款工具，绘制任意你感兴趣的图谱1-2张。



第三讲 词云与可视化

4. 采用给的程序，实现一段科学家文本的词云图绘制，越清晰越好（生成的词云图要单独拿出来）。

第四讲 情感分析

1. 使用PPT给的情感分析平台（或其它平台），对文本情感进行分析，并截图；
2. 完成sentiment_analysis_1-sentiment_analysis_4，4份代码。做截图，并简要做代码运行总结分析。

3. 谈一谈情感分析在营销学科/领域的应用以及价值；并且分析大语言模型（LLM）在该领域可能带来的新应用与新改变（仅营销）。

第四讲 情感分析

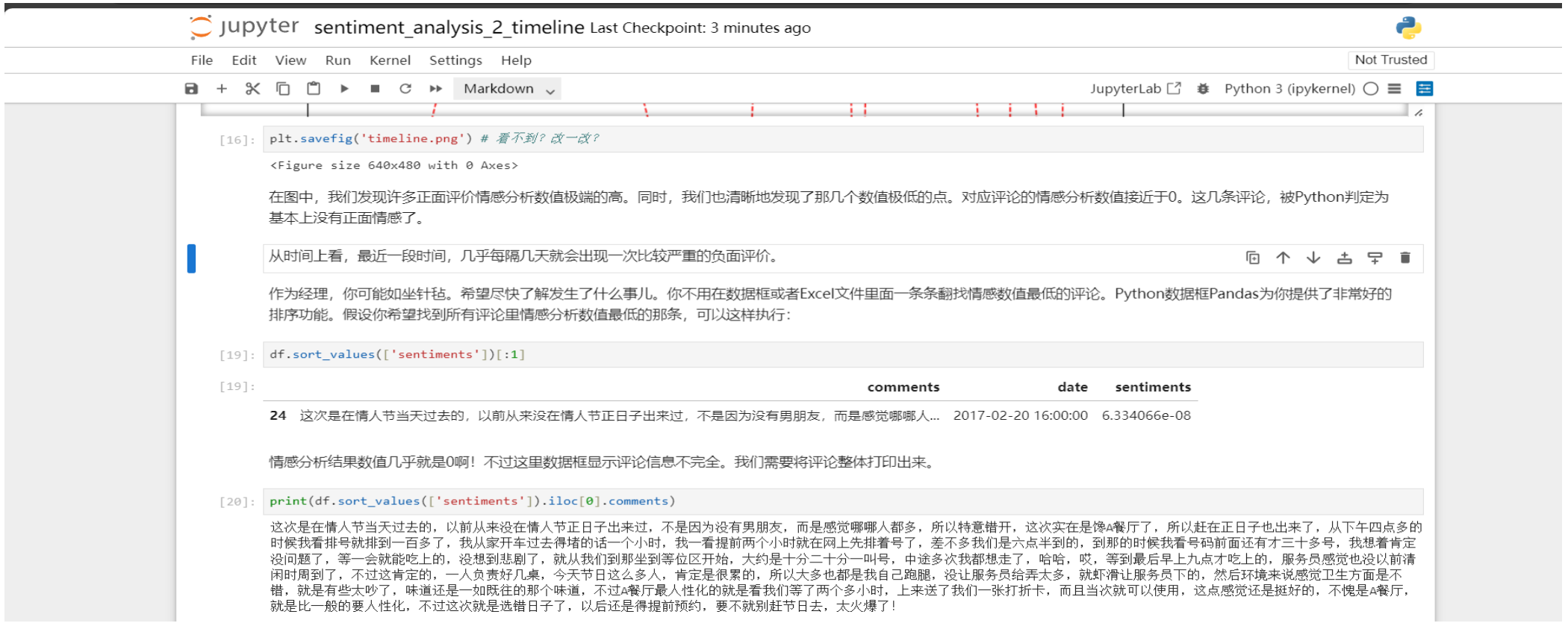
1.使用PPT给的情感分析平台（或其它平台），对文本情感进行分析，并截图；

第四讲 情感分析

2.完成sentiment_analysis_1-sentiment_analysis_4，4份代码。做截图，并简要做代码运行总结分析。1

第四讲 情感分析

2.完成sentiment_analysis_1-sentiment_analysis_4，4份代码。做截图，并简要做代码运行总结分析。2

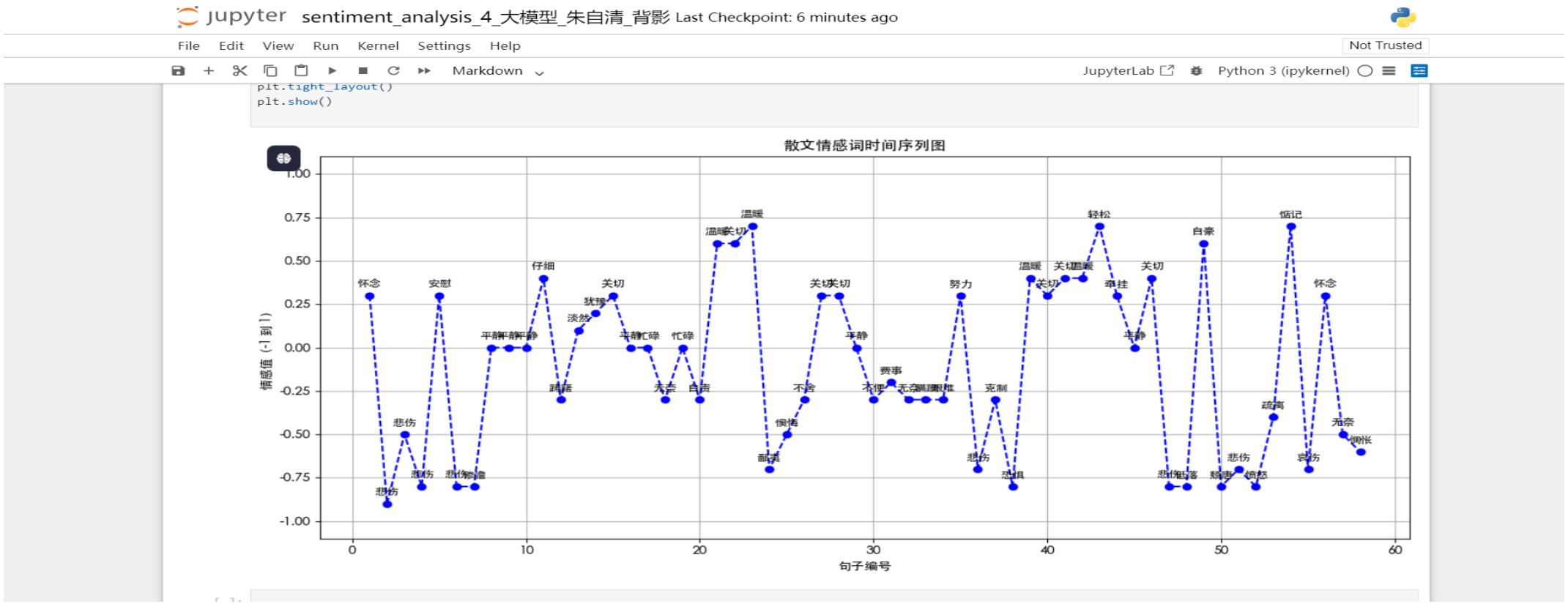


第四讲 情感分析

2.完成sentiment_analysis_1-sentiment_analysis_4，4份代码。做截图，并简要做代码运行总结分析。3

第四讲 情感分析

2.完成sentiment_analysis_1-sentiment_analysis_4，4份代码。做截图，并简要做代码运行总结分析。4



第六讲 知识图谱理念

1. 实际产业案例分析：使用3-5页PPT对“阿里商品大脑”、“美团大脑”、“丁香医生知识图谱”、“领英知识图谱”…其中任意一家机构/公司最新的知识图谱生态构建，进行简要介绍与分析。要求：需要是最新进展（不能复制课程PPT中的内容）；可以是一个简单的案例；有自己的评价。自由发挥。（营销、信管，都可以结合自己的专业兴趣，自由选择分析对象）

美团大脑简介

美团大脑是什么？

美团大脑是美团点评基于其庞大的业务体系（包括外卖、酒店预订、电影票务等）构建的一个AI系统，旨在通过深度学习、自然语言处理(NLP)、知识图谱等多种技术手段提升用户体验和服务效率。

核心功能：(1)用户行为预测(2)智能推荐(3)商家管理优化

最新进展: 美团大脑在2025年引入了增强现实(AR)和虚拟现实(VR)技术，用于改进用户界面体验，并增强了其知识图谱的深度和广度。

第六讲 知识图谱理念

详细介绍与分析：
美团大脑是美团构建的超大规模生活服务领域知识图谱系统，旨在支撑其“Food + Platform”战略。截至2025年，该图谱已覆盖超6亿实体（包括用户、商户、商品、地址、菜品等）和千亿级关系，核心创新在于深度融合多源异构数据——不仅整合交易日志、用户评论、菜单图片，还通过OCR、NLP和多模态模型从非结构化内容中抽取结构化知识（如“宫保鸡丁含花生”“某餐厅适合亲子聚餐”）。
其典型应用包括：智能推荐（基于“用户-场景-供给”三元组）、搜索意图理解（如识别“带娃吃饭”隐含的儿童座椅需求）、以及动态定价与调度优化。尤其在“到店+到家”融合场景中，知识图谱打通了线上行为与线下服务，实现跨业态关联推理。
值得肯定的是，美团大脑将知识图谱从“静态本体”升级为“动态认知引擎”，具备实时更新与因果推断能力。但挑战仍存：一是长尾商户信息稀疏导致冷启动问题；二是用户隐私与数据利用的边界需更透明。总体而言，美团大脑代表了知识图谱在本地生活服务领域的深度落地，其“场景驱动、闭环反馈”的构建范式具有行业示范意义。

第六讲 知识图谱理念

评价：
美团大脑这个知识图谱，说白了就是美团给自己建的一个“超级大脑”，用来搞懂用户到底想要啥、商家能提供啥、以及怎么把这两头高效地连起来。它不光记下你点了什么外卖、看了哪家酒店，还能从评论里扒出“这家店有宝宝椅”“那道菜特别辣”，甚至看菜单图片识别出菜品配料。这样一来，当你搜“带孩子吃饭”，它就能精准推有儿童设施的餐厅，而不是随便列一堆馆子。这玩意儿确实挺聪明，也真有用。比如送餐时间估得更准了，推荐的店也更合口味，背后都是知识图谱在默默算。而且它不是死的，今天新开一家店、明天某路堵车，它都能快速更新，反应很快。但也有让人嘀咕的地方。比如它知道得太多了——你搜过一次减肥餐，接下来一周都在推轻食，有点吓人；还有些小餐馆没数据，就很难被推荐，可能永远没机会冒头。总的来说，美团大脑技术很牛，让生活方便了不少，但怎么用好这份“聪明”，别让用户觉得被盯着，也别让小商家被落下，是它下一步得想明白的事。

第六讲（2） 知识图谱工具

1. 使用PPT中知识图谱链接平台，检索、截图（大词林等，可用的）；
2. 使用白板建模绘制一个你感兴趣的“知识图谱”，可以是人物关系，也可以是事物关系，或者概念之间的关系等等，并解释你绘制的图谱；
3. 使用echarts中的关系图，绘制作业2）中的“知识图谱”。
4. 使用Neo4j（可在线版本），编程绘制一款（简单）知识图谱（内容不限）（仅信管）。

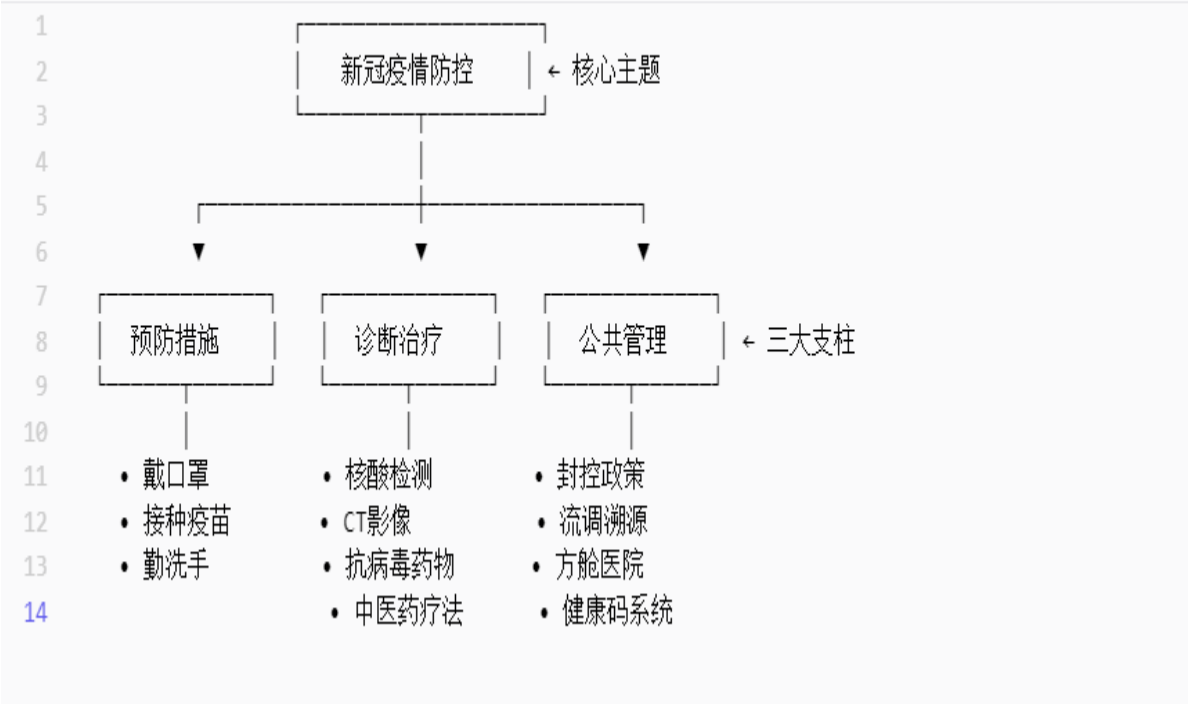
第六讲（2） 知识图谱工具

1.使用PPT中知识图谱链接平台，检索、截图（大词林等，可用的）；

第六讲（2） 知识图谱工具

2.使用白板建模绘制一个你感兴趣的“知识图谱”，可以是人物关系，也可以是事物关系，或者概念之间的关系等等，并解释你绘制的图谱；

text



这张知识图谱以 **“新冠疫情防控”** 为核心主题，构建了一个三层逻辑结构，清晰展现抗疫工作的系统性与协同性。

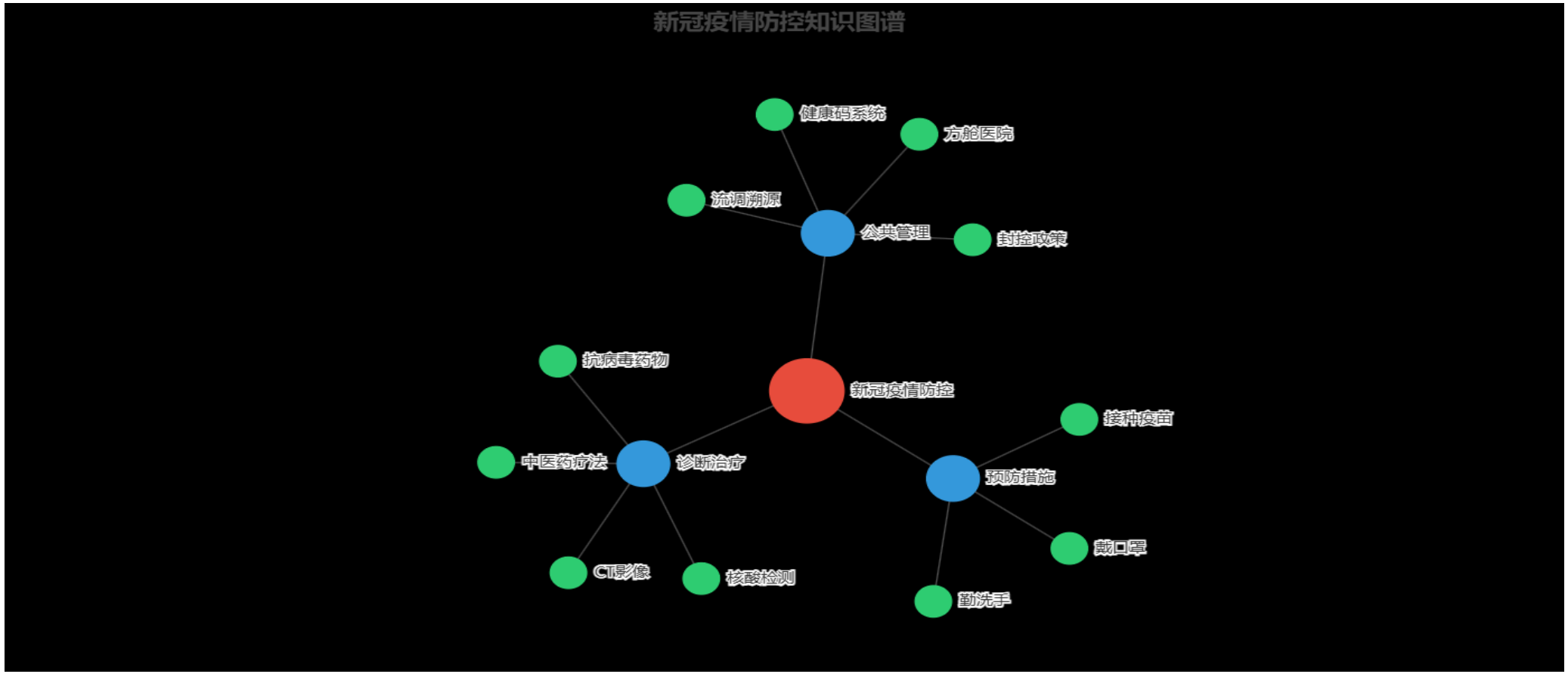
第一层（中心） 是总目标——有效控制疫情传播、保障人民生命健康。

第二层 分为三大关键维度：**预防、诊疗、管理**，分别对应“防得住、治得好、控得稳”的策略逻辑。

预防措施:聚焦个体行为与免疫屏障（如疫苗），是从源头降低感染风险；**诊断治疗**:关注临床路径，涵盖从检测到用药的全链条医疗响应；**公共管理**:则体现国家治理能力，通过流调、封控、信息系统实现精准防控。**第三层** 列出具体手段，每个都是真实落地的抗疫实践。例如，“健康码”连接了个人行动与政府监管，“方舱医院”解决了轻症收治难题。

第六讲（2） 知识图谱工具

3.使用echarts中的关系图，绘制作业2）中的“知识图谱”。



第六讲（2） 知识图谱工具

4.使用Neo4j（可在线版本），编程绘制一款（简单）知识图谱（内容不限）