|  |
| --- |
| 2023-2024学年第二学期 |
| 数据科学在数字法治领域中应用现状与趋势 |
|  |

|  |
| --- |
| **学号：2023202128**  **姓名： 沈逸尘** |

# 报告摘要

随着数据科学和大数据技术的迅猛发展，其在各个领域的应用日益广泛，法学也不例外。虽然笔者的专业并非法学方向，但对于这一领域有独特的兴趣，因此选择了法学领域进行分析。本调研报告旨在探讨数据科学在法学领域的现状及其对于其未来的前景做出展望。通过对现有文献数据库（Scopus数据库）的系统检索和分析，试图回答数据科学在法学领域应用的趋势如何，在未来几年内会有怎样的变化，主要的应用领域是什么，局限在哪里。调研发现，数据科学在法学研究文献中的应用总体数量可观，并且随着时间的推移呈现波动上升趋势，未来根据预测会更加频繁，应用领域主要集中在数据安全保护、司法数据库、法学量化研究、数据驱动决策等方面。本报告通过概括性分析当前数据科学在法学中的文献应用，揭示其潜在价值和发展趋势，并提出未来研究方向和应用建议。然而，本次调研也存在一些缺陷和不足，如对于文献数据的检索过程还可以改进等。

# 1.调研工作的设计

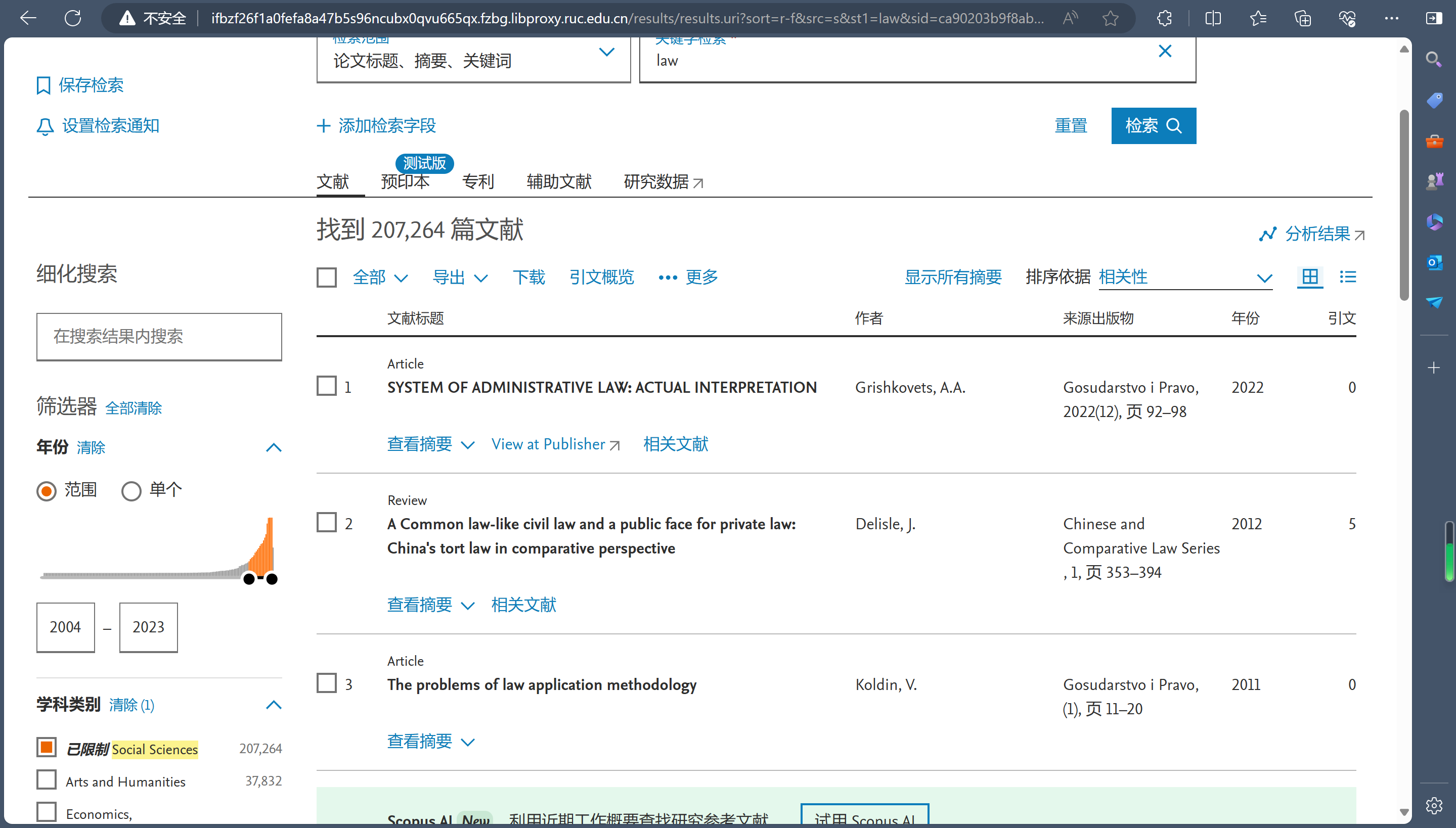
本调研的主要目的是探讨数据科学在法学领域的应用现状，分析其发展脉络和趋势，分析其在法律分析、法学研究、法律数据库、立法等方面的具体应用，并评估其对法学研究的影响和潜在价值。

本调研的主要调研对象是学术文献，通过文献数据库检索近年来全球学界在学术期刊、会议论文集和电子图书中发表的法学领域有关数据科学应用的研究论文，下载这些文献的主要信息包括作者、标题、摘要、发表年份、EID等等。

本调研所使用的主要调研方法是数据分析法，即对收集到的文献和调查数据进行统计分析，揭示数据科学在法学中的应用现状、趋势和主要细分领域。针对调研目的，即分析发展脉络和趋势，本调研计划采用根据年份进行分组统计，并进行数据可视化，随后利用机器学习方法对于未来的趋势进行预测。对于研究领域的划分，本调研计划利用自然语言处理对文本进行分类，集中分析在各个领域的文献占比。

# 2.调研工作的开展情况

为了实践以上所述的设计，首先需要完成文献数据的检索和数据收集工作。针对现有的在线文献数据库，我们希望找到一个在全球文献数据库中相对比较完整、权威也易于下载导出文本信息的平台。此处本调研选择了Scopus，该数据库是由爱思唯尔出版社推出的科研管理、学科规划数据库，是全球最大的同行评审期刊文摘和引文数据库，覆盖了自然科学、技术、工程、医学、社会科学、艺术与人文学科，全球重要的大学排名机构都是采用Scopus数据库的数据做学科产出以及学科发展评估。在Scopus数据库中检索law，选择学科为social science并选出近二十年的形式为Article和Review的文献，将他们的基本文献信息和摘要、关键词导出为csv格式文件，如图。





然后，为了检索法学领域与数据科学有关的文献，将检索关键词改为law + data science, law + big data, law + database以及law + artificial intelligence，虽然可能遗漏一些关键词，但是由于数据量较大，少量的遗失对于分析过程和结果的影响基本不大，可以忽略，文献导出的过程和前一部分相同，但在字段选择上，加上了题录信息字段，用于后续的国家分布分析。

经过整理，基于关键词law检索的数据量在20w左右，其他几个关键词组检索的结果在4w左右，总数据量24w左右。

另外，后续还有用到全球各国GDP数据，数据集来自ModelWhale平台。

总体而言，数据本身符合调研在设计时的一些猜测，因此在检索和后续的分析中执行情况都较好，基本无偏差的完成了调研设计的要求。为保证调研项目的可重现性，调研项目的完整数据和数据分析过程均已上传至Github的Datascience\_survey仓库，链接：<https://github.com/YichenShen0103/Datascience_survey>

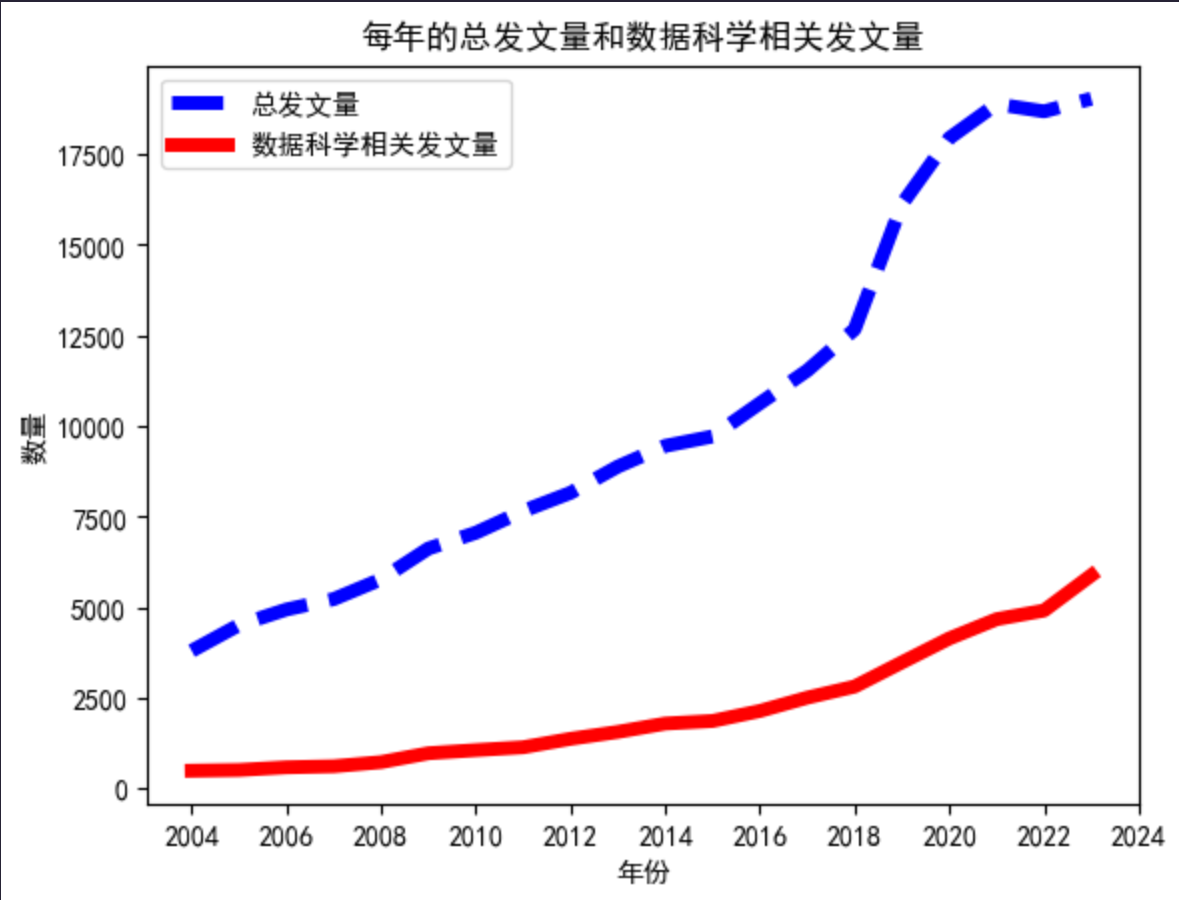
# 3.调研数据的分析与洞察

以下将分为三个主题对于获取到的数据进行分析洞见，分别是法学领域的数据科学应用随时间推移的变化，法学领域数据科学应用的国家分布，法学领域数据科学应用的主要细分领域。数据分析用到的主要工具是Python程序，主要基于anaconda（windows）环境，其中Python版本是3.11.5，主要用到的pandas包版本为2.0.3，matplotlib包版本为3.7.2，sklearn包版本为1.3.0，pyecharts包版本为1.8.1，特此说明。

**（一）随时间变化趋势分析**

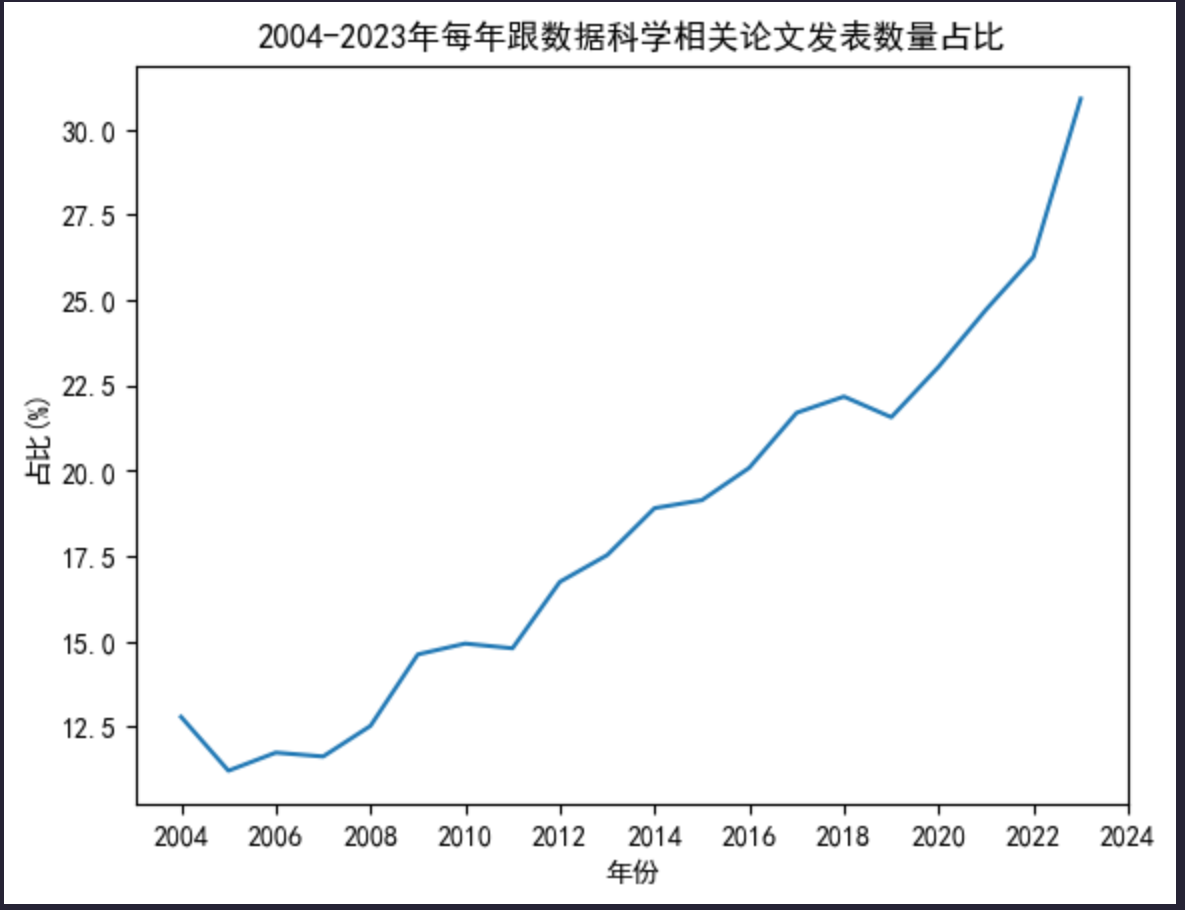
首先利用pandas包读入存有2004-2023年law相关论文和law+数据科学相关关键词相关论文的两个csv文件，以下分别称为表一和表二。读入后为确保数据无误，可以先查看两表的描述性统计数据，如以下两图。

确认无误后，对两个表都用年份进行分组计数。为了使得结果更具有可读性，利用matplotlib包以年份为横坐标，发文量（也就是论文EID计数）作为纵坐标，设置好标题，轴坐标，图例等信息后，绘制一幅双折线图，如下图。



可以看到，法学领域和法学的数据科学应用领域都呈现波动上升趋势，且都在2018年左右呈现大幅的提升。诚然我们可以认为，法学领域的数据科学相关发文量是在增加的。然而，仅用这样的指标显然是不够说明数据科学的应用有没有更广泛的，因为既然两者都在增加，我们就难以衡量数据科学在法学领域的发文占比到底有没有变化。

因此，在分组计数后的表二中新增标题为“占比”的索引列，将初始值设定为零。然后再计算每一年的发文量占比存入其中，单位为%。结束后，再以年份为横坐标，占比为纵坐标，设置相关参数后，绘制一幅折线图，如下。

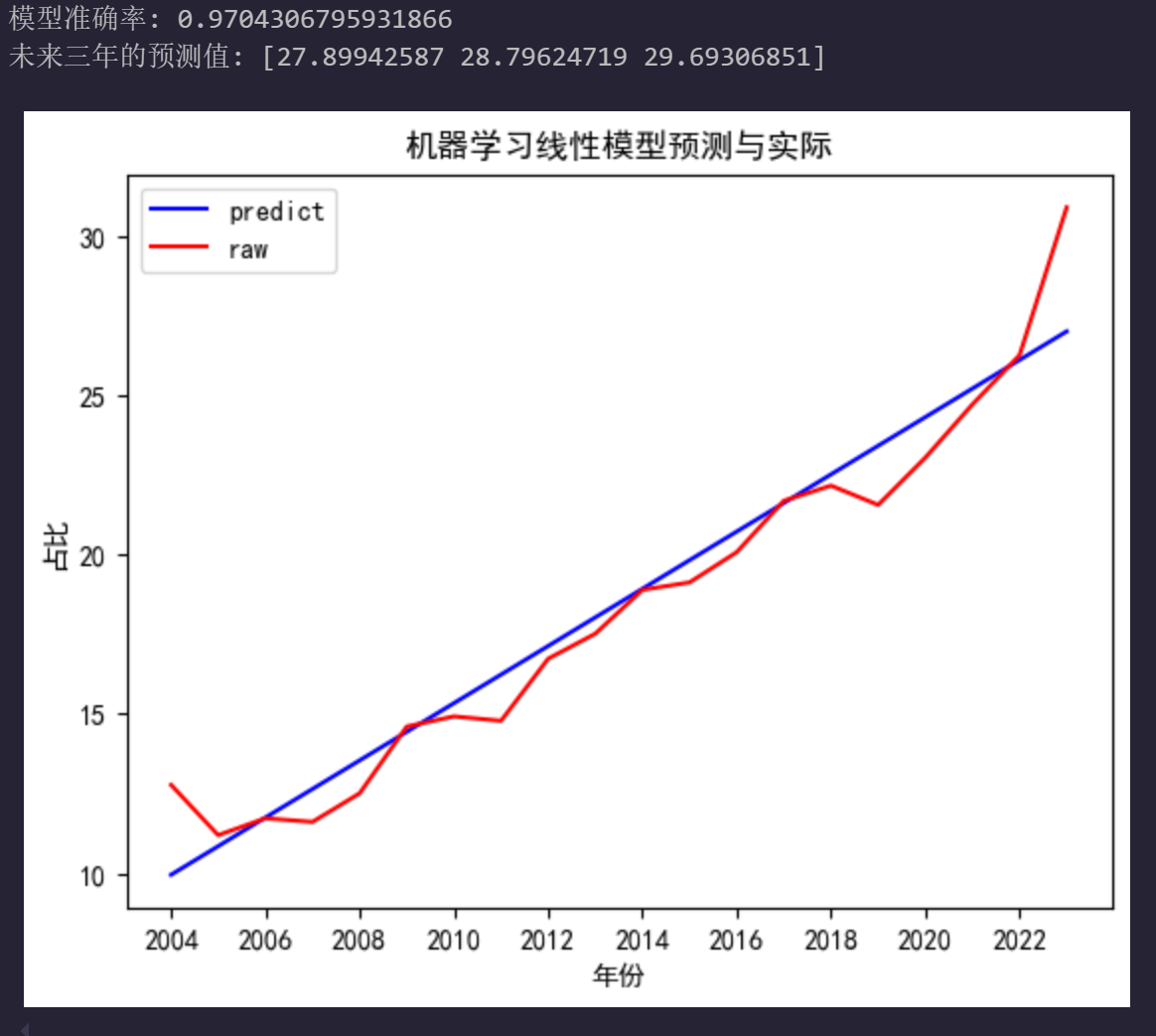


不难看到，法学领域的数据科学发文占比也呈现波动上升的趋势。至此，我们基本可以判断，法学领域的数据科学应用是随着时间推移越来越广泛的。

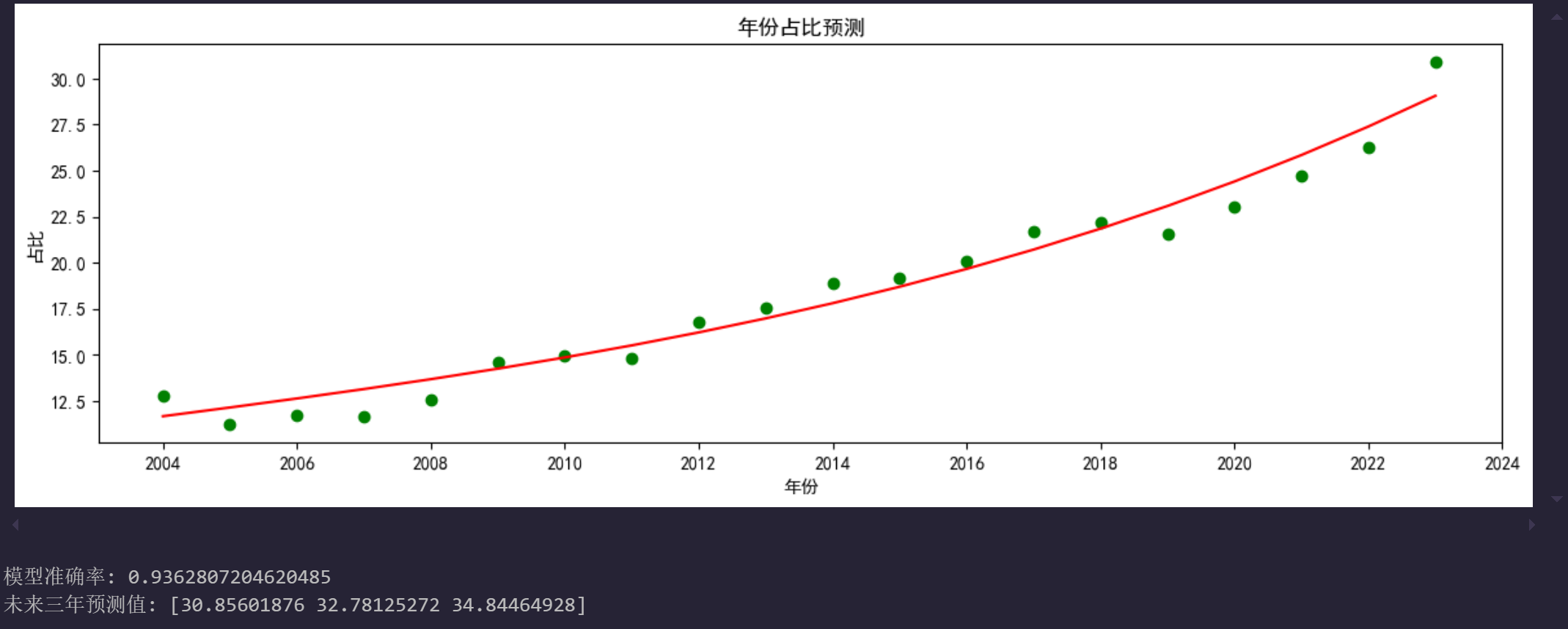
进行到这里，笔者难免好奇，既然我们已经发现了一些规律，我们能否借助这一规律预测未来几年内数据科学再法学领域应用占比变化呢？为了解答这一疑问，我们可以引入机器学习的方法，利用相关模型进行拟合，并作出预测。

首先进行数据准备，利用numpy的to\_numpy方法将表二的年份转化成numpy数组对象，利用ravel函数将表二的占比转化成一维数组；利用sklearn中的函数将两组数据划分为训练集和测试集。

从sklearn包中取出LinearRegression对象及其相关方法，利用线性回归拟合准备好的数据，在图3-4基础上绘制拟合直线并输出模型准确率和未来三年的预测值，如图3-5。



模型准确率较高，但对于未来的预测却有些偏颇。这从本质上是线性拟合并不适合本组数据，可以看到，发文占比的上升速度是越来越快的，而不是线性。因此我们需要更强的模型来拟合。控制其他方法不变，将一维的年份数据利用sklearn自带的对象分割成高维，笔者又试验了8个更强的回归模型（即多元线性回归、决策树回归、K临近回归、随机森林回归、AdaBoost回归、GBR回归、Bagging回归以及ExtraTree回归）。经过调参，当数据被分割为7维左右的时候，模型拟合最准确。为了更直观的感受区别，笔者绘制了多元线性回归的折线图与普通线性回归进行对比。



碍于篇幅，此处不将所有图像进行展示，但是其他模型图像与上图基本类似。可以看到，在未来，数据科学在法学领域的应用据预测还能有一定程度上的提升。至此，根据时间尺度分析数据科学在法学领域应用的变化趋势的任务基本得到完成。

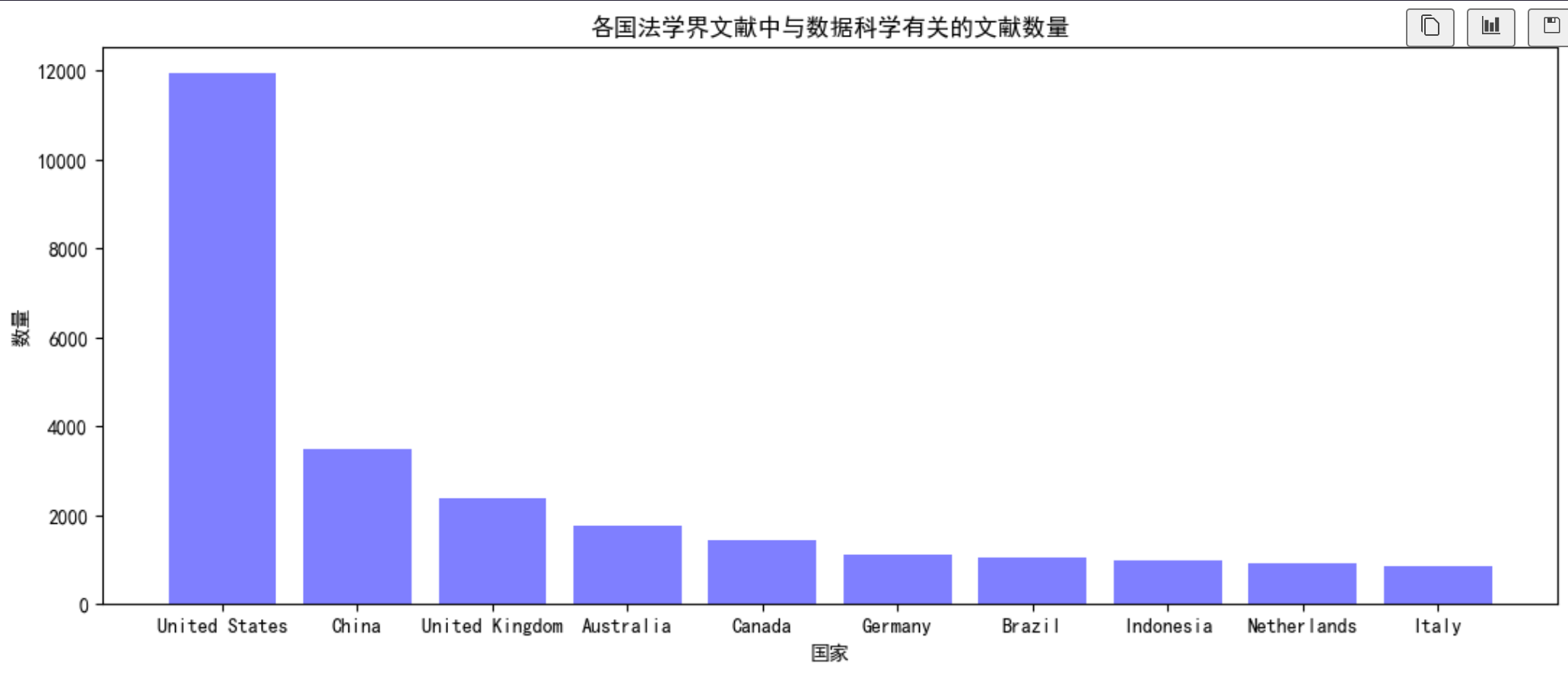
**（二）全球各国数据科学应用发文量分析**

对于这一主题的分析，只需读入表二，同样进行数据确认。确认无误后，进行数据预处理，即要抽取出数据中的“国家”元素。然而在下载到的文献数据中是没有直接的国家索引列的，只有“归属机构”列。以下需要从这一列数据中抽出国家元素。

考察这一列数据可以发现，数据中的格式一般是“第一作者的机构名称，地址，国家；第二作者的机构名称，地址，国家；……”。由于作者数量不等，机构的地址中又含有各自不等数量的“，”，难以使用pandas自带的str.split方法进行分割。因此我们需要编写如下算法，抽取每一个作者的国家归属，并进行存储。

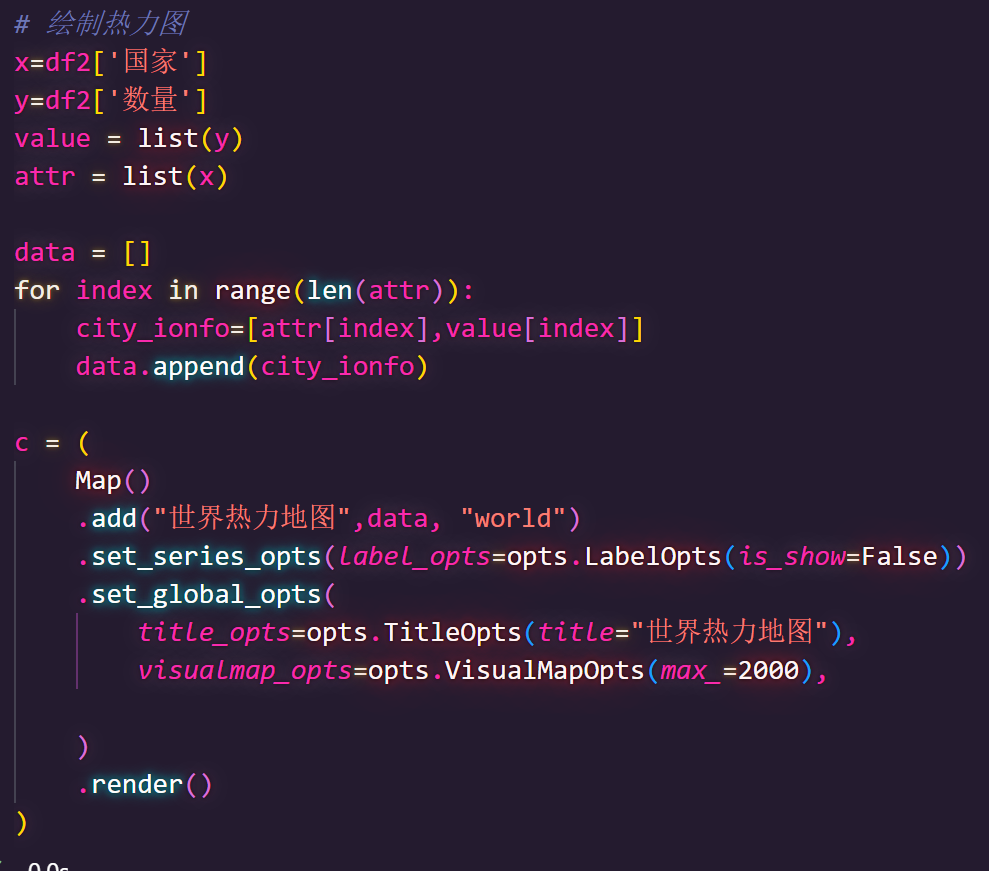


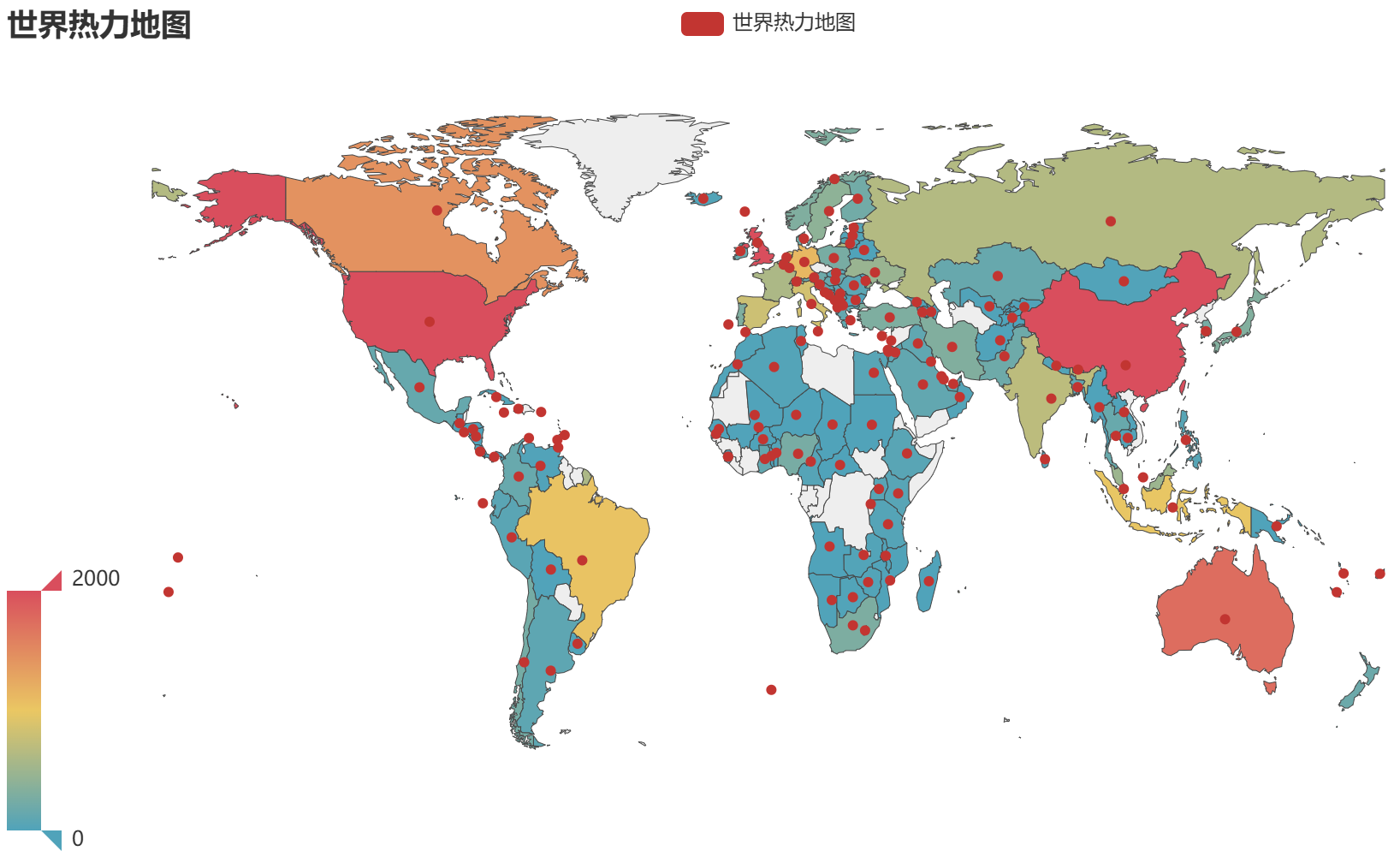
为了直观展现不同国家的发文量区别，首先以发文量排名前十的国家为x轴，发文量为y轴，绘制一幅条形图，如下。



这幅图从数量对比的角度很好的反映了各国法学界文献中与数据科学有关的文献数量。

利用pyecharts包中的Map对象可以创建一幅地图，导入国家和对应的发文量数据后，适当调整参数后，可以绘制一幅热力图，代码实现和效果图如下。





热力图比之条形图，在数值的反映上稍有逊色但是对于地理分布的反应更加直观，二者相互映衬，是为了呈现更好的可视化效果。

不难发现，在世界各国中，美国法学研究运用数据科学最多，且遥遥领先，前十名国家主要集中在东亚、欧洲、澳大利亚的经济发达地区，非洲，南美洲除巴西以外，亚洲特别是中亚西亚地区较少，个别地区无相关文献。此处笔者猜测发文量与当地GDP有关联，因此读入全球各国GDP信息，进行简单数据准备后运用pysci包中的函数计算p值考察相关性，以下是代码实现和输出结果。

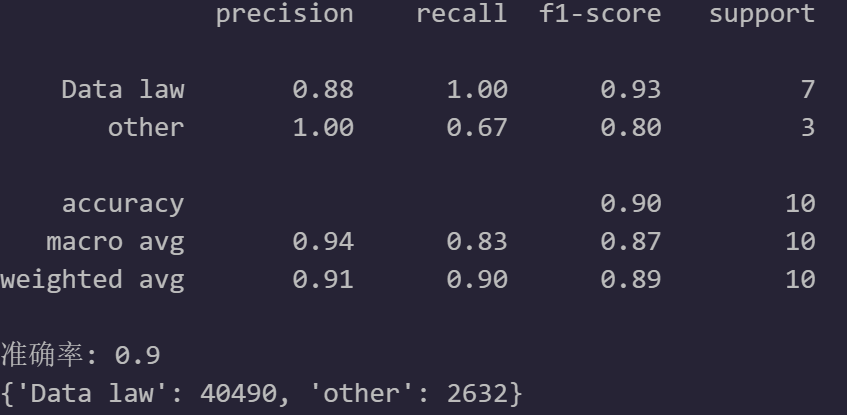


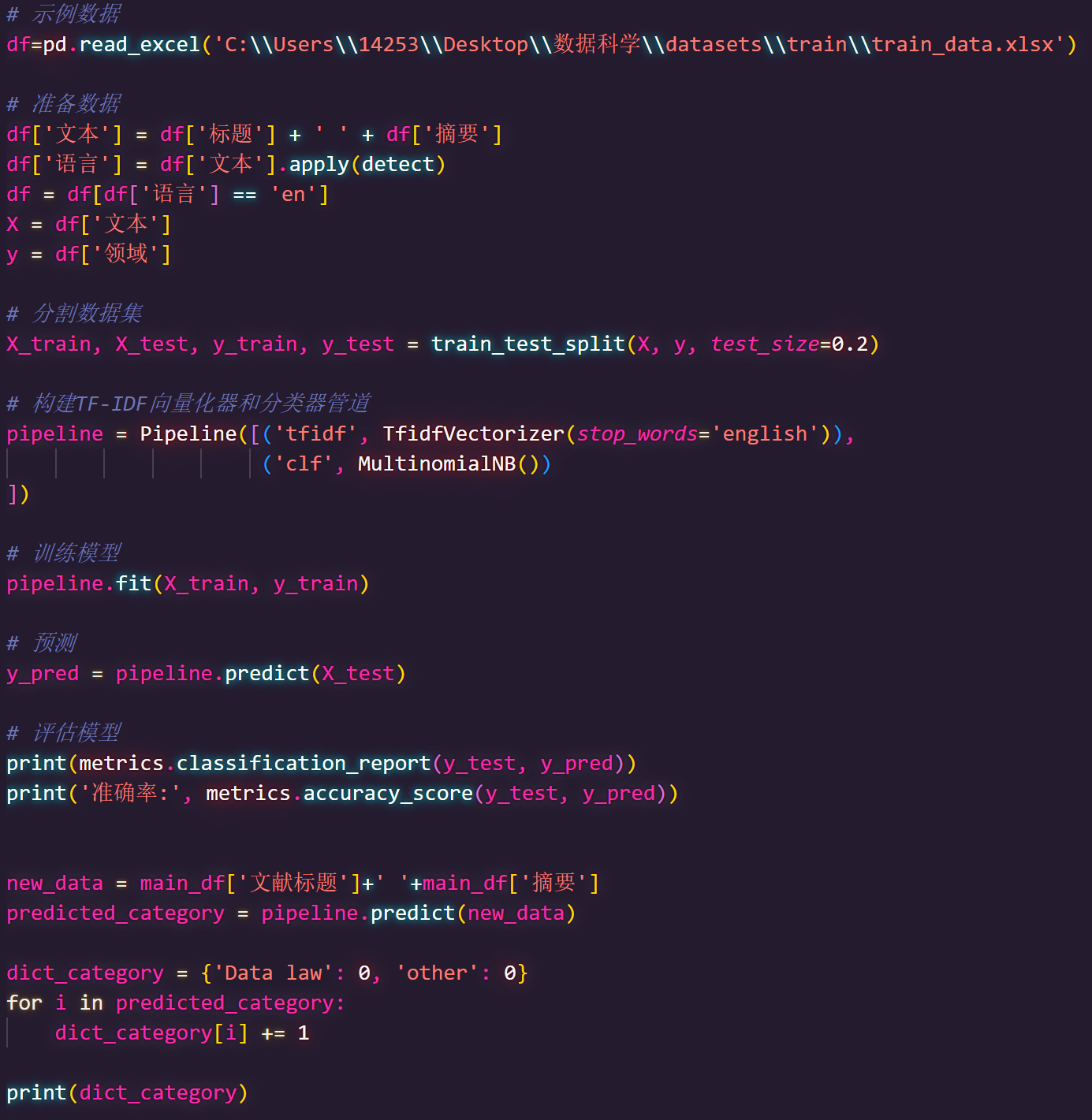
可以看到，p值远小于0.05，法学界数据科学应用的发文量与经济发展水平呈强相关关系。至此，这一主题的数据分析任务基本得到完成。

**（三）数据科学在法学领域的应用形式和细分领域分析**

要进行细分领域的分析，就必须引入自然语言处理进行文本分类。由于缺少适合的数据集作为训练集，笔者选择了50篇具有代表性的文本，在阅读后将他们按照细分领域进行了分类，由于数据集较小，因此此处只进行一个细分领域，即数据安全保护立法——这也是所有细分领域中最常见的一个应用方向——的分类，即分出属于数据安全保护立法领域的和不属于这一领域的两类。

导入表二和用于训练的数据集，对表二进行清洗，除去为翻译成英语的文献。对于可用英语的文献，取出其英语摘要和标题进行连接，并利用TF-IDF算法将文本向量化提取特征。由于训练数据集较小，分类算法选择上可以选择朴素贝叶斯分类。利用sklearn包的对象创建多项式贝叶斯分类器，并进行训练，并利用训练好的模型对表二中的数据进行分类，输出结果和具体代码如下图。





由于笔者知识水平和时间有限，在模型选择、参数调整上都可能出现问题，另外训练集也较小，分类的可靠性并不强。但是总体也能看出，数据科学在法学领域的一你敢用基本停留在数据安全保护层次，和笔者预期的数据库、大数据辅助决策等存在一些偏差。

这种偏差实际上也反映了法学界在数据科学应用上的一些不足，并没有踏出传统的领域，只是将数据作为一种研究对象而并并没有认识到数据科学是辅助法学研究的重要工具和可作为应用的手段。未来可以研究的方向还有法律文本分类，立法需求分析，法律数据库，法律人工智能，大数据驱动决策等等。

至此，最后一个主题的任务也基本完成。

# 4.调研工作的心得体会

在这次的调研工作中，我获得了许多宝贵的经验和启示。从调研设计、数据收集到数据分析，每一个环节都不像看上去一样简单，有许多需要实际操作过后才会发现的问题和挑战。

首先，调研设计是整个调研工作的基础和指南。这一过程让我深刻体会到，科学合理的调研设计不仅能够提高调研效率，还能确保调研结果的准确性和可靠性。而且调研设计一旦确定了，后面再想改动就非常困难，所以这一环节必须要对细节的执行考虑的非常细致，同时又要有宏观的目标设计。在设计的过程中，既不能对调研对象有过度的要求和揣测，防止主观偏见的带入，又要对调研对象有一定的宏观把握和猜测，以明确方向，这要求分析者具有相当程度和知识面和数据科学素养。

在数据收集阶段，我主要采用了文献检索方法。通过在各大数据库中检索相关文献，我了解到当前数据科学在法学领域的研究热点和应用现状。这一过程中，我发现数据收集的难点在于如何筛选和提取有用信息，在大量的数据中检索出自己想要的数据耗时耗力，不同的检索方式和筛选手段会得到完全不同的效果。在面对海量的数据的时候，这个阶段重要的是思维的转变，也就是要有大数据时代的思维，不能希望自己能把所有数据看一遍，否则只会越来越无力而是要了解文献检索的底层逻辑，到底是关键词检索，还是文本词汇检索还是主题检索，每一种检索方式都有自己独特的效果。

数据分析和洞见则是调研工作的核心环节。在这个过程中，我运用了多种数据分析工具和方法，对收集到的数据进行了系统分析。在分析过程中，我深刻体会到数据的多样性和复杂性。这个阶段最重要的是对于不同类型的数据需要采用不同的分析方法和可视化手段，只有灵活运用各种工具，才能全面揭示数据背后的信息。除此以外，对数据的清洗和预处理也是非常重要的，这也是大数据思维的重要体现。最后，基于数据分析的结果得出结论的过程也要非常审慎，要注意每一步推导之间的逻辑关系，不能轻易的下结论，例如在第一个主题任务——根据时间分析数据科学在法学领域应用的趋势——中，不能只看数据科学相关的发文量，而要看它的占比有多少。对于数据可视化方式的选择也尤为关键，对于一个确定的分析目标要有确定的图表来最完美的反映分析者想要表达的意思，可视化要能解答读者的困惑。

最后，我也体会到了数据分析这一个过程在宏观上设计的重要性，也就是将数据故事化的能力。相互不能连接的分析是没有任何意义的，对于数据分析的宏观设计的精髓就是要形成一个讲述，形成一个故事。在这次调研过程中，我着重考虑了这一个因素，例如在第一部分的任务中，加入了对未来的预测，在第二部分的任务中，尝试分析是什么因素导致了数据科学在法学领域应用的发文量在国家之间会有区别，同时，第三部分原本是没有的，但是仅有前两个部分显然过于浅显，难以满足读者对于深入理解数据的要求，因此要再设计一个专门的主题来看数据科学的应用领域。在数据分析形成讲述的过程中，我认为最重要的是要预测到读者的疑问，看到一组数据后，读者会有什么思考？读者会有什么疑问？如何引导读者去思考？这些问题的解答就是将数据分析形成故事的重要因素。但是在这个领域，我也是初出茅庐，理解难免还过于浅显，实际的效果也不是那么完美，未来我也会在这个领域多多探索，这也是提升数据科学的素养的重要途径。

作为总结，通过这次调研工作，我不仅掌握了大量关于数据科学在法学领域应用的知识，还提高了自身的调研能力和科学素养。特别是对数据分析工具和方法的运用，使我在面对复杂问题时更加自信。我也深刻体会到了数据分析工作的不易，面对大量数据的无力感，面对分析结果不如人意时的挫败感都是非常宝贵的财富。再解决困难，转变思维的过程中，我也体会到了学习技术的成就感。此外，通过调研工作，我学会了如何在实际工作中灵活运用理论知识，解决实际问题。这些经验和收获将对我今后的学习和研究产生深远的影响。

注：本项目所有数据分析源代码、原数据、可视化文件和分析报告均已在Github上传，链接：https://github.com/YichenShen0103/Datascience\_survey