计算机科学与技术学院（人工智能学院）本科生导师制创新项目总结报告

2023年4月6日

**1研究目的与意义**

港口是重要的交通基础设施，港口货物吞吐量与国民经济发展密切相关，港口发展受国家政策影响较大，对全世界港口信息做数据统计和可视化展示可以更好地了解全球港口的运营情况、货物流通情况以及各个港口之间的相互联系，为政府部门、航运公司、经济学家和政策制定者提供重要的参考依据，帮助他们更好地制定政策和经营策略[1]。

因此，本项目决定利用Python爬虫爬取港口信息网站数据。Python爬虫的应用不仅可以提高数据获取的效率和质量，还可以利用python内含的各种库对数据进行清洗、统计、可视化等操作，增强数据分析的能力和水平，实现数据的自动化处理和分析。数据可视化的方法也可使我们更直观地看到数据中的趋势和关系，进而发现数据中隐藏的规律和价值[2-3]。

**2 研究内容与总体方案**

**2.1研究内容**

**(1)港口信息爬取**

选择世界港口查询网站为数据来源，爬取的具体信息包括港口名称、锚地、航线、所属地区、国家、港口简介等。进而通过数据去重、数据格式统一等步骤对爬取的港口网站进行数据清洗，将处理后的数据存入到MySQL数据库中，为后续开展数据分析和可视化提供基础。

**(2)世界港口数据分析和可视化展示**

对获取后的世界港口数据采用统计分析方法，从地区、航线、国家、类型等多角度进行数据统计以及可视化展示。同时使用直观地图标注出各个港口所在位置，以便于分析世界各大港口的地理上的聚集特征。

**2.2总体方案**

**2.2.1爬取网页数据**

采用Requests库作为爬虫工具，爬取http://gangkou.00cha.net上的港口相关数据。在进行爬虫之前，需要阅读Requests库的官方文档，了解其基本用法和功能。然后，需要访问目标网站，分析其网页源代码和请求类型，确定如何从URL中传递参数和从HTML中提取数据。接着，需要在PyCharm中编写爬虫代码，使用Requests库和其他辅助库（如BeautifulSoup或者lxml）来发送HTTP请求，解析响应内容，并提取出港口信息。最后，需要将爬取到的港口信息保存到文件或者数据库中，以便后续的分析和处理。

**2.2.2解析并提取数据**

使用BeautifulSoup库作为HTML解析工具，从爬取到的网页源代码中提取港口基本信息。创建BeautifulSoup对象，并定位和提取元素来获取元素的文本和属性。然后，需要使用BeautifulSoup库解析爬取到的网页源代码，并根据网页的HTML结构，提取港口的名称，所属国家，地区，航线和锚地信息。最后，需要将提取到的港口信息存储到列表中。

**2.2.3清洗并存储数据**

使用MySQL作为数据库管理系统，将清洗后的港口基本信息存入数据库中。在存入数据库之前，需要对提取到的港口信息进行数据清洗，剔除无效和重复数据，保证数据的质量和完整性。然后，需要使用MySQL创建数据库，并根据提取到的数据定义合适的关系模式，如港口表，国家表等。接着，需要使用pymysql库连接MySQL数据库，并将清洗后的数据存入相应的表中，以便后续的查询、分析和可视化展示。

**2.2.4统计分析并可视化展示**

使用pandas库和pycharts库对港口数据进行统计分析和可视化展示。首先，需要使用pandas库从MySQL数据库中提取数据，并进行统计分析。然后，需要按照港口位置、航线、国家和类型等维度进行统计分析。接着，需要使用pycharts库将统计结果制作成方便数据直观展示的柱状图和方便地理位置直观展示的地图。最后，需要实现本地服务器上的网页可视化展示，以及简单的后端查询功能，以便可以通过网页查看港口数据的分析结果和详情。

**3 研究方法**

**3.1 python网络爬虫研究技术**

该项目使用了python网络爬虫技术爬取世界港口网站信息。网络爬虫是一种自动获取网页内容的程序，可以从互联网上抓取目标数据。网络爬虫技术的优点是可以快速、高效、灵活地获取大量的数据，而且可以根据不同的需求和场景定制不同的爬虫规则和逻辑。python是一种适合编写网络爬虫的脚本语言，具有简洁、易学、跨平台、丰富库等特点[4]。python的特点是可以让编程者用更少的代码实现更多的功能，而且可以方便地调用各种第三方库来增强爬虫的功能和性能。本研究利用python网络爬虫技术，从世界港口网上爬取了各个港口的名称、所属国家、航线、类型、锚地等信息，然后对这些数据进行了统计分析和可视化展示。

**3.2 可视化地图**

该项目根据港口信息的经纬度制作可视化地图。地图是一种常用的数据可视化形式，可以直观地表达出数据之间的空间关系和分布特征。根据港口信息的经纬度制作地图的方法是利用echart库，将经纬度数据转换为地理坐标，然后在地图上绘制出各个港口的位置和其他相关信息[5-6]。这种方法可以方便地对港口数据进行分析和展示，从而得到有关世界港口地点地分布的有价值的知识。

**3.3 MySQL数据库**

该项目了使用MySQL作为统计数据的数据库的研究方法。MySQL是一种开源的关系型数据库管理系统，可以存储和管理大量的结构化数据，支持多种操作系统和编程语言，具有高性能、高可靠性、高扩展性等特点。使用MySQL作为统计数据的数据库的方法是利用pymsql库，将爬取到的港口信息导入到MySQL数据库中，然后利用SQL语句对数据进行查询、分析、处理和导出[7-8]。这种方法可以方便地对港口数据进行统计和管理。

**3.4 pandas统计分析**

该研究使用pandas对数据进行统计分析。pandas是一种基于Python的数据分析工具，可以提供高效的数据结构和操作方法，支持多种数据源的读取和处理，具有强大的数据清洗、分组、聚合、合并等功能。使用pandas对数据进行统计分析，将爬取到的港口信息导入到pandas中，然后利用pandas提供的函数对数据进行描述性统计、相关性分析、异常值检测等，最后利用pyecharts提供的接口对数据进行柱状图的绘制和展示[9-10]。这种方法可以方便地对港口数据进行探索性分析和可视化呈现，从而更直观具体的了解世界港口相关信息。

**4 研究结果**

**4.1世界港口信息数据库**

利用python爬虫成功爬取数据，清理完毕后将总共3290条数据存入MySQL数据库中，爬取的数据分两张表进行存取。

第一张表（图1）：

基本信息包括：港口代码、所在城市、港口中文名、港口英文名、所在国家、所属地区、航线中文、航线英文、锚地以及港口类型：

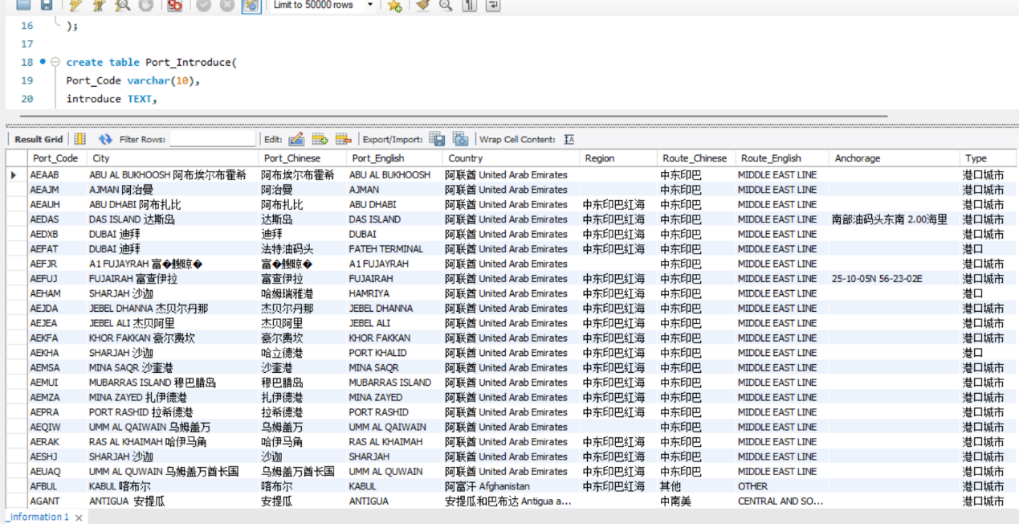
****

图1港口基本信息表

第二张表（图2）：

基本信息包括：港口代码和该港口的相关介绍

****

图2港口介绍信息表

**4.2数据可视化分析**

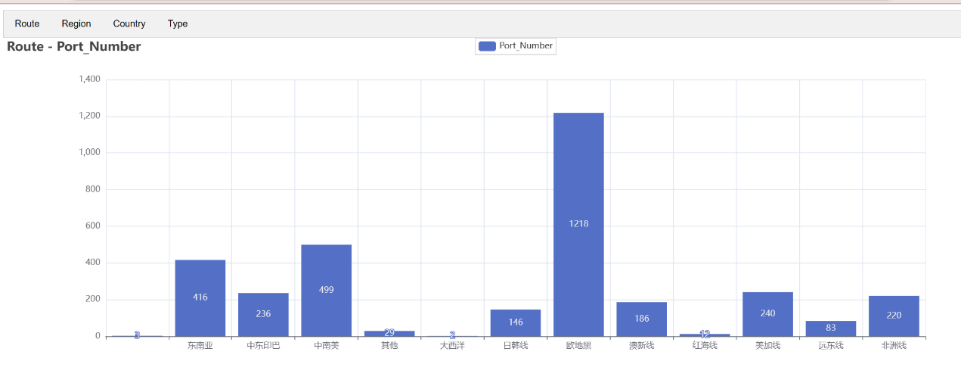
****

图3按航线分组柱状图

如图3，在爬取的3290个港口中，它们所在航线最多的有1218条在欧地黑航线，499条在中南美线，416在东南亚线，最少的航线是大西洋线。这些数据反映了以下几个方面的信息：

欧地黑航线是连接欧洲、地中海和黑海的重要航线，覆盖了世界上经济发达和发展中的国家和地区，具有较高的贸易量和货运需求。这些港口具有较强的经济实力和贸易活力，也反映了欧洲和地中海地区在世界港口体系中的重要地位。

中南美线是连接中国、东南亚和中南美洲的重要航线，主要承担了中国和中南美洲之间的贸易往来。在中南美航线的这些港口与中国有着密切的贸易联系，也反映了中国在世界贸易中的重要角色和影响力。

东南亚线是连接东亚、东南亚和澳大利亚的重要航线，主要承担了东亚和东南亚之间以及东南亚和澳大利亚之间的贸易往来。这些港口与东亚和澳大利亚有着密切的贸易联系，澳大利亚作为澳洲唯一的国家，与其他国家之间的交流贸易高度依赖航线，因此会有更多港口处于该航线。

大西洋航线是连接北美、欧洲、非洲和南美的重要航线，覆盖了世界上不同经济发展水平和地理位置的国家和地区。但大西洋航线的港口却不多，一方面可能由于爬取数据源的原因，但更多的是因为：大西洋航线作为时加上最长的海运航线之一，风险较高，竞争较激烈。大西洋沿岸国家和地区的经济发展水平和贸易需求不均衡。发达国家可能更倾向于使用空运或者陆运来进行贸易往来。大西洋沿岸国家和地区的港口建设和管理水平不高。一些港口可能由于地理位置不佳、规模过小、设施陈旧、效率低下等原因而无法吸引更多的船只停靠

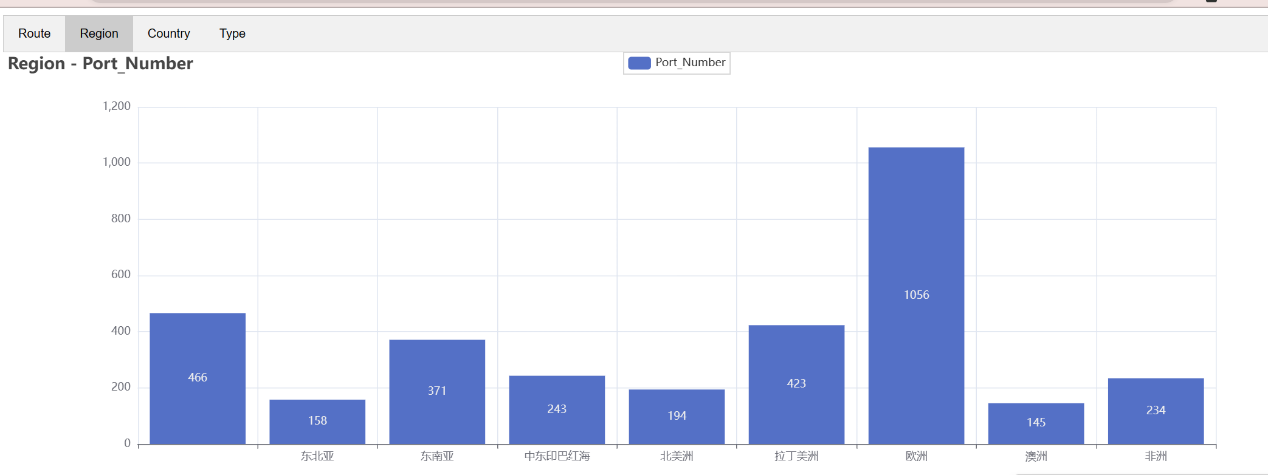
****

图4按地区分组柱状图

如图4，在爬取的数据中有1056个港口位于欧洲，欧洲的港口数量和分布离不开欧洲的地理环境和经济发展，大致由于下面几方面原因：

欧洲位于欧亚大陆的西部，北临北冰洋，西临大西洋，南临地中海，东与亚洲相连，是世界上海岸线最为曲折的大洲。欧洲的海岸线长达3.8万公里，占全球海岸线的20%以上。欧洲的海岸线上分布着许多深水良港，如鹿特丹港、汉堡港、安特卫普港、马赛港、伦敦港等，这些港口不仅为欧洲本身的贸易提供了便利，也为欧洲与其他大洲的贸易提供了重要的通道。

欧洲是世界上经济最发达、工业化程度最高、贸易量最大的大洲之一。欧洲拥有强大的工业基础和消费市场，是全球重要的商品生产和消费地区。欧洲的港口不仅承担着欧洲内部的货物运输，也承担着与亚洲、美洲、非洲等其他大洲的货物运输。欧洲的港口具有较高的吞吐量和效率，能够满足欧洲经济发展和贸易需求。

欧洲是世界上区域一体化程度最高的大洲之一。欧盟是由27个成员国组成的政治和经济联盟，是全球最大的单一市场和货币联盟。欧盟通过实施统一的法律、政策、标准和规则，促进了成员国之间的自由流动和合作。欧盟还通过建设跨国交通网络（如泛欧交通网络），加强了成员国之间以及与邻近国家和地区之间的互联互通。欧盟为欧洲港口提供了有利的政策环境和基础设施支持，增强了欧洲港口的竞争力和协调能力。

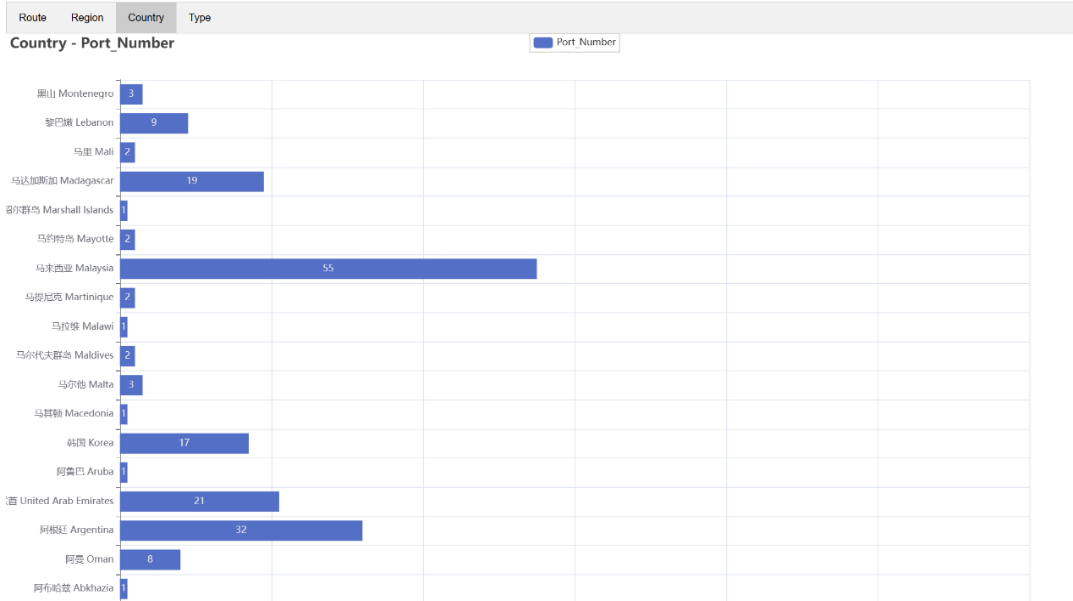
****

图5按国家分组柱状图

如图5，在按国家区分港口的柱状图中可以轻易看出，中国、美国、印度尼西亚和菲律宾等都有不少的港口数量，这些国家有这么多港口的原因可能有以下几方面：

地理位置和海岸线。这些国家都位于亚洲大陆的东部或南部，临近太平洋或印度洋，是世界上海洋贸易最活跃的地区之一。这些国家的海岸线都比较长，有些还有许多岛屿，为港口的建设和发展提供了自然条件。例如，美国的海岸线长达1.9万公里，中国的海岸线长达1.8万公里，日本的海岸线长达2.9万公里，印度尼西亚的海岸线长达5.4万公里，菲律宾的海岸线长达3.6万公里。

经济发展和贸易需求。这些国家都是世界上经济规模较大或增长较快的国家，有着强大的工业生产和消费市场，需要大量的进出口贸易来满足经济发展和人民生活的需要。这些国家的港口不仅承担着本国内部的货物运输，也承担着与其他国家和地区的货物运输。例如，2020年前11月，美国、中国、日本、印度尼西亚、菲律宾的货物进出口额分别为3.4万亿、4.1万亿、1.3万亿、2.7千亿、2.8千亿美元。

区域一体化和合作机制。这些国家都参与了一些区域一体化和合作机制，如亚太经合组织（APEC）、东盟（ASEAN）、东盟加三（ASEAN+3）、中日韩自贸区（CJK FTA）等，通过降低关税壁垒、简化贸易程序、建设跨境基础设施等措施，促进了区域内外的贸易往来和互联互通。这些机制为港口的发展提供了政策支持和市场机遇。例如，2020年前11月，美国、中国、日本、印度尼西亚、菲律宾与APEC成员国之间的货物进出口额分别占其总进出口额的70.6%、64.7%、74.5%、76.8%、83.9%。

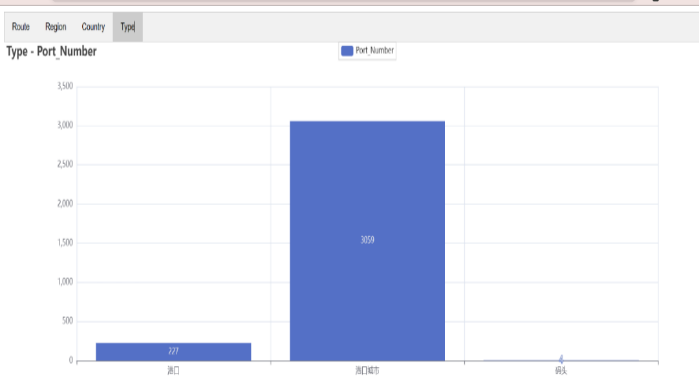
****

图6港口类型分组

如图6可见在本项目所爬取的数据中港口城市数量较多

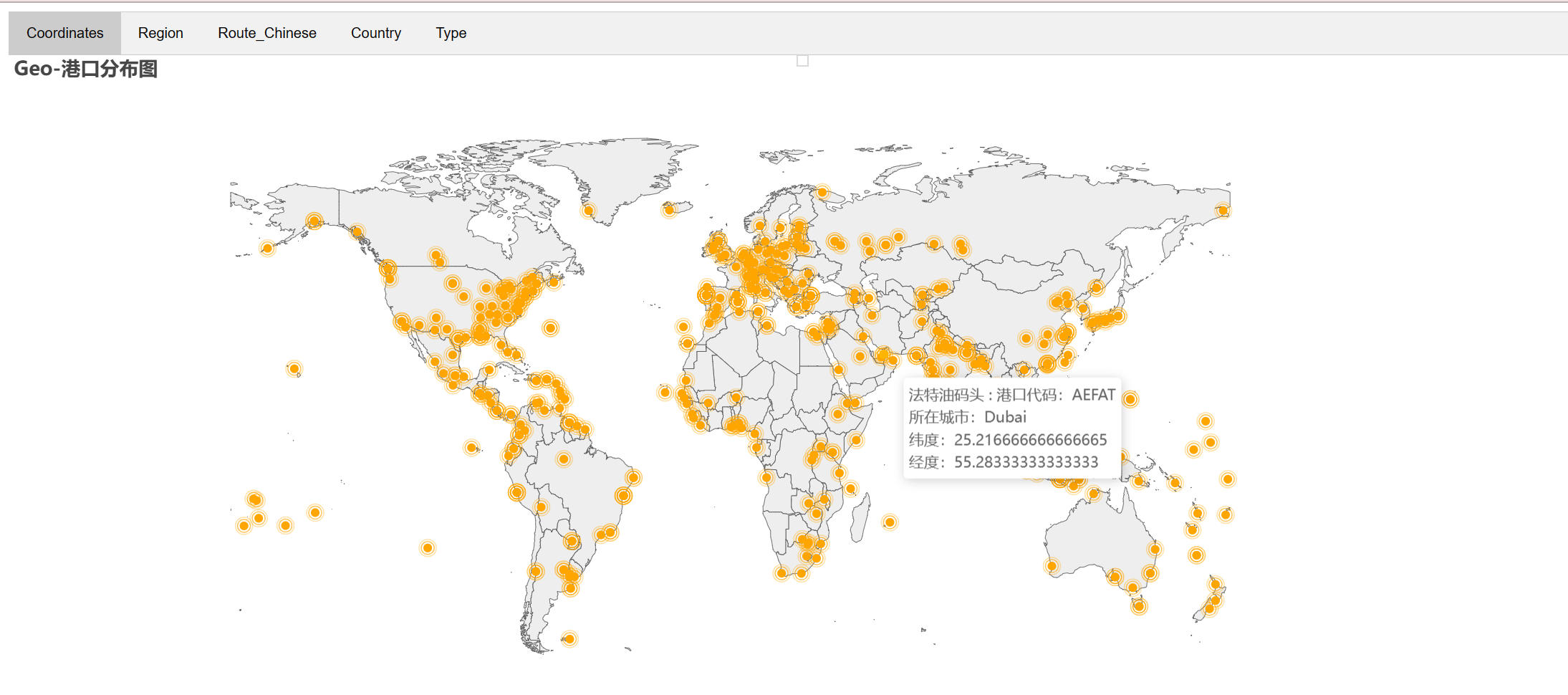
****

图7港口位置分布图

如图7，该可视化地图上标注了相关港口的信息，从地图上可以看出，港口的数量和分布与海岸线的长度和曲折程度有关，一般来说，海岸线越长越曲折的国家和地区，港口的数量越多，分布越密集。例如，美国、中国、日本、印度尼西亚、菲律宾等国家和地区都有很多港口，而非洲内陆国家、澳大利亚西部等地区则很少有港口。另外，港口的数量和分布也与经济发展水平和贸易需求有关，一般来说，经济发展水平越高，贸易需求越大的国家和地区，港口的数量越多，分布越广泛。例如，欧洲、北美、东亚等地区都有很多港口，而非洲、南亚、中东等地区则相对较少。



图8网页查询功能展示

如图8，这是网页上设置的一个简单的后端查询功能，你可以通过输入港口代码查询相关的港口基本信息，港口代码可以通过将鼠标悬浮到可视化地图上具体的点位中得到。

**参考文献**

[1] 千际投行.港口行业价值分析，看这篇就够了[EB/OL].(2021-01-13)[2023-04-06]. https://m.21jingji.com/article/20210113/herald/157609b3aaea6db62b9701fefc873468.html

[2] ChatGPT领航者. 如何理解数据可视化的意义?[EB/OL].(2023-01-26)[2023-04-06]. https://www.zhihu.com/question/580547256

[3] CnOpenData-ybl. 如何理解数据可视化的意义?[EB/OL].(2022-01-20)[2023-04-06]. https://zhuanlan.zhihu.com/p/460023344

[4] [ChenBinBini](https://blog.csdn.net/ChenBinBini). Python爬虫史上超详细讲解（零基础入门，老年人都看的懂）[EB/OL].(2020-11-17)[2023-04-06]. https://blog.csdn.net/ChenBinBini/article/details/109739116

[5] sunchengquan.python调用echart交互式可视化[EB/OL].(2018-03-04)[2023-04-06].https://blog.csdn.net/sunchengquan/article/details/79438730

[6] Summer是只白喵.超简单，用Python Pyecharts制作可视化大屏[EB/OL].(2021-12-16)[2023-04-06].https://zhuanlan.zhihu.com/p/445991406

[7] 民工哥. MYSQL的主要特点有哪些？[EB/OL].(2022-11-10)[2023-04-06]. https://www.zhihu.com/question/45638237

[8] Samven\_7.Python.用pymysql连接MySQL数据库[EB/OL].(2019-07-06)[2023-04-06].https://blog.csdn.net/qq\_42780289/article/details/94837708

[9] 爬遍天下无敌手.如何使用Pandas进行数据分析！最详细的数据分析教程！[EB/OL].(2020-09-26)[2023-04-06]. https://blog.csdn.net/weixin\_43881394/article/details/108811809

[10] 庄蔚. 【十分钟Python知识点】pandas最详细教程[EB/OL].(2020-09-16)[2023-04-06]. https://zhuanlan.zhihu.com/p/9988992