# **全国高校名单网络爬虫需求文档**

1. 数据采集的意义
2. 教育研究与决策支持：为政府部门、研究机构分析高等教育资源的地理分布均衡性、院校结构（公办/民办、本科/专科）、发展历程提供数据基石，支持宏观教育政策的制定与评估。
3. 公众信息服务与透明度：为学生、家长提供一份权威、全面的高校参考名单，消除信息不对称，辅助其进行升学和择校决策。
4. 商业市场分析：为教育科技公司、招聘平台、出版机构等提供目标市场分析的数据依据，例如定位特定区域或类型的院校。
5. 动态监控与预警：定期运行爬虫，可及时发现并跟踪高校的新设、合并、更名、撤销等动态变化，为相关方提供第一时间的信息更新。
6. 数据来源
7. 主要来源
8. ****中华人民共和国教育部 政府门户网站****

* ****URL：****<http://www.moe.gov.cn/jyb_xxgk/s5743/s5744/>
* ****说明****：教育部会以“教育部公报”的形式，不定期发布《全国高等学校名单》。该页面汇集了历年来的名单公告，是**最权威的数据源头**。爬虫将从此处获取最新的官方名单PDF或HTML页面。

1. ****教育部直属事业机构 - 教育管理信息中心****

* ****URL****：<http://www.jsj.edu.cn/>
* ****说明****：该网站公布了中国政府承认的各类院校名单，特别是包含“中外合作办学”机构与项目的信息，可作为重要补充。

1. 补充来源：

****阳光高考信息平台（教育部高校招生指定平台）****

* ****URL：****<https://gaokao.chsi.com.cn/>
* ****说明：****这是教育部下属的面向学生和家长的官方信息平台。其中的“院校库”数据全面、更新及时，且具有良好的结构化特征，非常适合作为爬取详细信息的来源，用于补全和校验从官方公报中提取的数据。

1. 采集数据项

|  |  |  |  |
| --- | --- | --- | --- |
|  | **数据项** | **说明** | **示例** |
| **基础标识信息** | school\_id | 学校唯一标识码（教育部编码） | 4111010001 |
| school\_name | 学校官方全称 | 北京大学 |
| school\_english\_name | 官方英文名称 | Peking University |
| **地理信息** | location\_province | 所在省份 | 北京市 |
| location\_city | 所在城市 | 北京市 |
| location\_address | 详细地址 | 北京市海淀区颐和园路5号 |
| longitude | 经度（通过地理编码API获取） | 116.310316 |
| latitude | 纬度（通过地理编码API获取） | 39.998743 |
| **属性信息** | administration | 主管部门 | 教育部 |
| level | 办学层次 | 本科 |
| type | 办学性质 | 公办 |
| establish\_year | 成立年份 | 1898 |
| **联系信息** | website\_url | 学校官方网站 | http://www.pku.edu.cn |
| **标签与分类** | tags | 特色标签 | ["双一流","985工程","211工程"] |
| category | 院校类别 | 综合类 |
| **元数据** | data\_source | 数据来源URL | http://www.moe.gov.cn/... |
| update\_time | 数据采集/更新时间 | 2024-05-27 10:00:00 |

1. 采集规模
2. 初始规模：预计采集约2800-3200所院校（涵盖普通高等学校和成人高校）。
3. 更新频率：

* 全量更新：每6个月执行一次，与教育部发布新名单的周期同步。
* 增量监控：每月对主要来源进行一次扫描，检查是否有新的公告发布，实现动态监测。

1. 数据量估算：初始数据表大小约在1MB以内，但随着采集历史版本数据，数据量会逐步增长。
2. 输出格式
3. 主输出-结构化文件：

* universities.csv：包含上表所有数据项的CSV文件，使用UTF-8编码。
* universities.json：包含相同数据的JSON文件，结构为 {"data":[{...},{...}]}，便于Web应用调用。

1. 持久化存储：导入至MySQL或PostgreSQL数据库中的china\_universities表。​
2. 日志与元数据输出：每次爬取生成一个日志文件（.log），记录爬取状态、遇到的错误（如无效链接）和数据校验结果。
3. 数据预处理
4. 去重与合并：基于 school\_id 和 school\_name，对从多个来源爬取的数据进行去重。如果同一学校信息有多个来源，以教育部来源为准。
5. 缺失值处理：

* 对于非关键字段（如英文名、成立年份）的缺失，标记为 NULL。
* 对于关键字段（如学校名称、ID）的缺失，该条记录将被标记为“待核查”，并记录日志。

1. 格式标准化：

* 地址：使用正则表达式将地址统一为“省+市+区+详细地址”的格式。
* 年份：统一为YYYY格式。
* 标签：将“985、211、双一流”等标签从文本描述中提取出来，转化为标准化的数组或布尔值字段。

1. 数据校验与增强：

* 连接有效性验证：使用HTTP HEAD请求抽查或验证学校官网链接的有效性。
* 地理编码：调用百度地图或高德地图的Geocoding API，将标准化后的地址转换为经纬度坐标，用于地图可视化。

1. 数据一致性检查：对比教育部名单和阳光高考的名单，对于名称不一致或一方缺失的记录，进行人工复核标记。
2. 数据分析且结果可视化

|  |  |  |
| --- | --- | --- |
| **分析主题** | **分析内容** | **可视化方案** |
| 空间分布分析 | 分析全国高校的密度分布，识别教育资源的集中与匮乏地区。 | ****全国热力图****或****点密度图****，以经纬度为基础，直观展示高校在全国的分布情况。 |
| 省份竞争力排名 | 统计各省份的高校总数、本科院校数、"双一流"高校数。 | ****多系列柱状图****，横向比较各省份的高教实力。 |
| 院校结构分析 | 分析全国公办与民办、本科与专科的比例。 | ****饼图****或****环形图****，清晰展示各类别院校的构成比例。 |
| 历史变迁分析 | 统计不同年代成立的高校数量，反映我国高等教育的发展历程。 | ****折线图****或****面积图****，展示高校数量随时间的变化趋势。 |
| 顶尖院校分布 | 专门分析“双一流”高校在各省的分布情况。 | ****中国地图chloropleth图****，用颜色深浅表示各省拥有的“双一流”高校数量。 |
| 院校类型分析 | 对高校的“综合类”、“理工类”、“师范类”等类别进行词频分析。 | ****词云图****，突出显示我国高校的主要类型。 |