

数据科学导论第三周作业

Part 1 爬虫

本项目计划采用Midscene的开源框架来完成整个爬虫过程，为了控制本实验报告的长度，我将把开源框架的配置+编写办法+维护办法，沉淀到另外的文件中，下面的部分仅为简单的概括

Midscene是什么？

以Playwright为执行引擎，以Midscene作为连接AI与浏览器的桥梁，通过AI模型理解自然语言指令，最终实现高度智能化的UI自动化测试

Midscene的优势是什么？

智能+可视化

其融合了AI的推理能力+Playwright的模拟操作，为用户提供多个ai函数，支持其高效的可视化的完成整个爬虫过程

Midscene该如何用？

在整个实践的过程中，Midscene的爬虫过程可以分为以下几步：

- 1.先梳理好整个模拟操作的流程
- 2.将操作流程转换为代码（在Midscene中是以.ts的格式来写）

3.多次微调，修改.ts代码，使得执行的操作符合预期（Midscene有回放的功能，可以辅助我们去进行调试，也就是我们可以及时的知道是哪一步代码，出现了问题）

Midscene源代码在哪里？（官方）

[midscene/README.zh.md at main · web-infra-dev/midscene](https://github.com/web-infra-dev/midscene)

使用指南在哪里？（官方）

[API 参考 - \(AI UI Automation, AI Testing, Computer Use, Browser Use, Android Use\)](#)

1.1 模拟操作流程

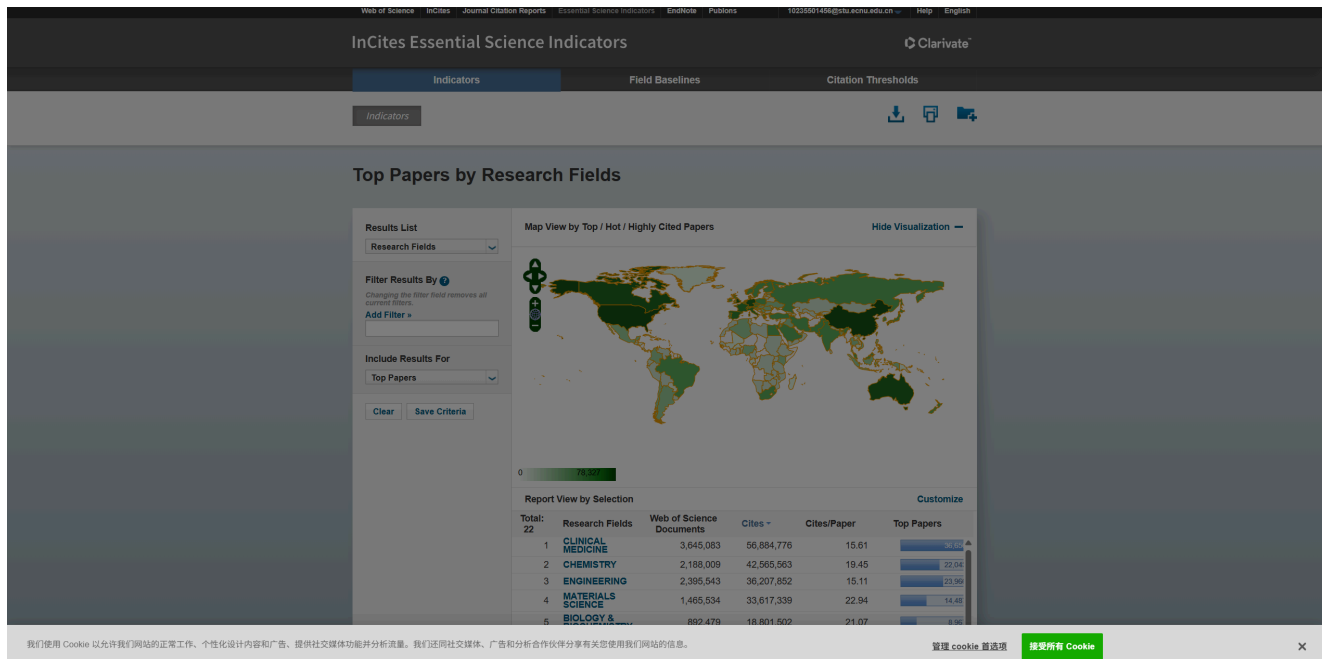
核心目的是为了前往[新建标签页](#)能够自动的获取ESI学科排名的相关数据

1.登录界面

Essential Science Indicators™

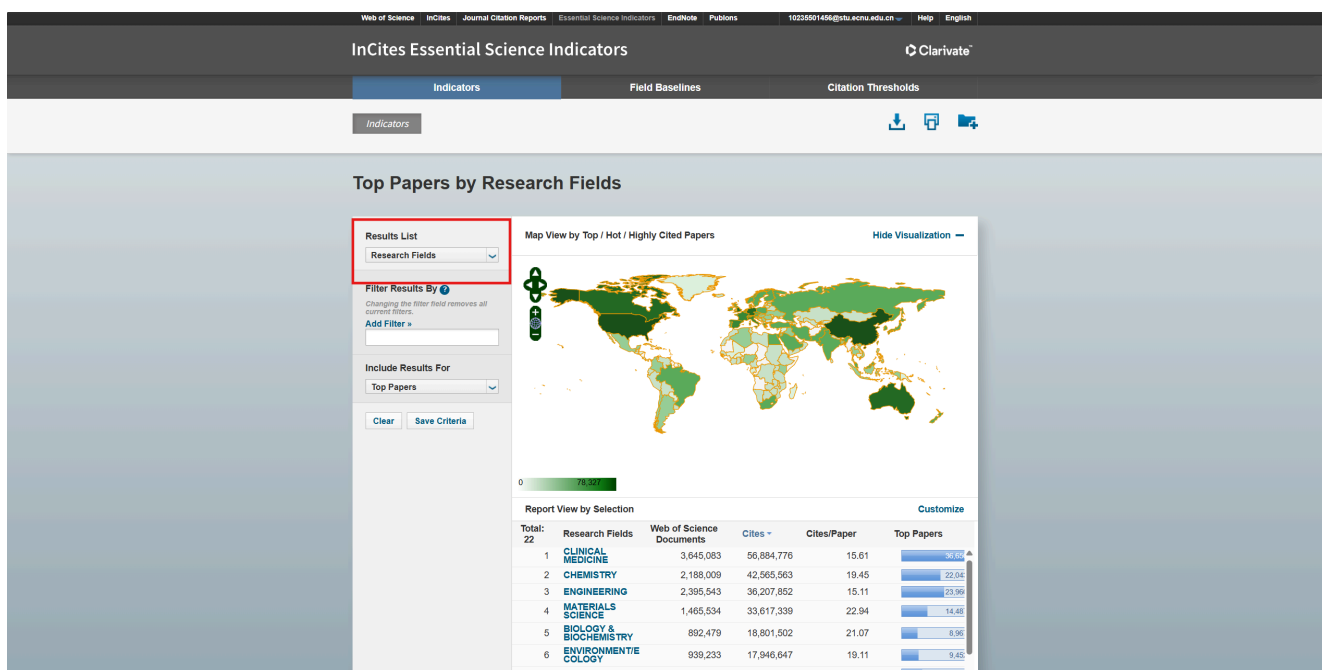
2.进入主页

需要先点击下方的“Accept All Cookies”



3.切换Result List

进入主页后，可以看到，它默认状态下，是按照 Research Fields来划分的，需要对当前的结果进行一个切换



切换成Institutions从而可以按照科研机构来筛选



4.按照学科门类，依次查询并下载

先点击Add Filter,后点击 Research Fields，从而支持按学科门类来筛选



可以看到Research Fields里包含如下所示的学科门类

Top Papers by Institutions

Results List

Institutions

Filter Results

Changing the filters changes the current filters.

Add Filter »

Include Results

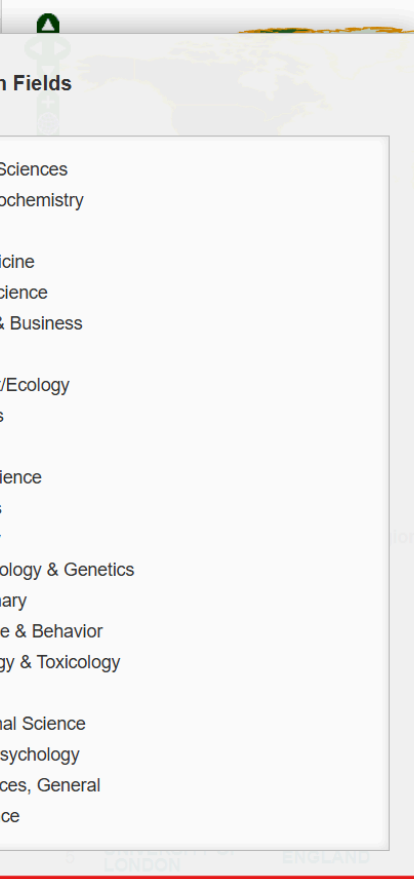
Top Papers

Clear

Save

Map View by Top / Hot / Highly Cited Papers

Hide Visualization —



Back

Search Fields

- + Agricultural Sciences
- + Biology & Biochemistry
- + Chemistry
- + Clinical Medicine
- + Computer Science
- + Economics & Business
- + Engineering
- + Environment/Ecology
- + Geosciences
- + Immunology
- + Materials Science
- + Mathematics
- + Microbiology
- + Molecular Biology & Genetics
- + Multidisciplinary
- + Neuroscience & Behavior
- + Pharmacology & Toxicology
- + Physics
- + Plant & Animal Science
- + Psychiatry/Psychology
- + Social Sciences, General
- + Space Science

Customize

	Web of Science Documents	Cites	Cites/Paper
1	637,297	15,566,361	24.43
2	456,676	14,593,904	31.96
3	275,741	10,789,920	39.13
4	402,207	8,832,204	21.96
5	273,539	8,260,184	30.20
6	188,552	7,339,267	38.92
7	212,893	6,483,838	30.46

下一步，我们应当是点击一个学科，选中，点击下载

InCites Essential Science Indicators Clarivate™

Indicators Field Baselines Citation Thresholds

Indicators

Top Papers by Institutions

Results List
Institutions

Map View by Top / Hot / Highly Cited Papers

Hide Visualization —

Filter Results
Changing the filters changes the current filters.

Add Filter »

× Agriculture

Include Results
Top Papers

Clear

Search Fields

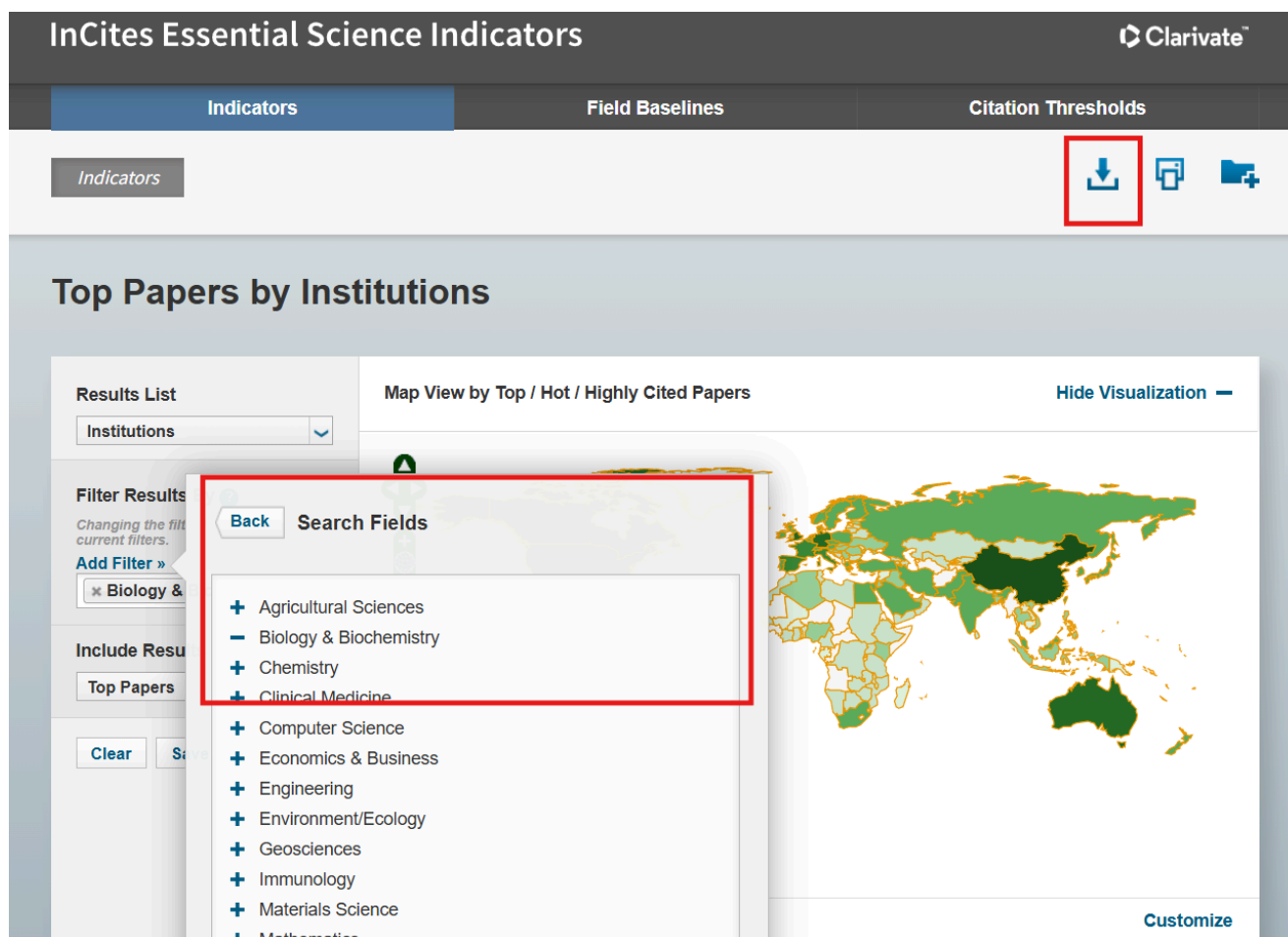
- Agricultural Sciences
- + Biology & Biochemistry
- + Chemistry
- + Clinical Medicine
- + Computer Science
- + Economics & Business
- + Engineering
- + Environment/Ecology
- + Geosciences
- + Immunology
- + Materials Science
- + Mathematics
- + Microbiology
- + Molecular Biology & Genetics
- + Multidisciplinary
- + Neuroscience & Behavior

Web of Science Documents Cites Cites/Paper

15,661 332,254 21.22

Customize

对于要下载下一个学科，需要先点击“-”删除当前的选中，然后点击“+”选中新的，之后再下载



接下来只要不断重复以上步骤，我们就可以完成整个爬虫过程

1.2 基于Midscene将操作过程转化为.ts文件

1.2.1 用到的几个ai函数

首先，我先用简短的篇幅，概括一下，在本实验中用到的ai函数

- **ai() 或 aiAction():**（适用于比较复杂，别的操作还不支持的情况）
 - 核心：用自然语言描述多步UI操作，AI自动执行整个流程。

- 实际场景：如"点击登录按钮，然后输入用户名"，适合复杂任务链。
- 简写用法：await agent.ai('描述操作步骤'); // 快速执行多步。
- **aiTap():**
 - 核心：点击指定元素。
 - 实际场景：触发按钮或链接，如登录/提交。
 - 简写用法：await agent.aiTap('登录按钮'); // 如'登录按钮'。
- **aiInput():**
 - 核心：在元素中输入文本。
 - 实际场景：填写表单或搜索框，常需后跟Enter键触发。
 - 简写用法：await aiInput('10235501456@stu.ecnu.edu.cn', 'Email address input');（第一个参数是实际要输入的地方；第二个参数是定位，也就是输入在哪里）
- **aiWaitFor():**
 - 核心：等待指定元素出现或变化。
 - 实际场景：由于前端的前面涉及到加载的过程，需要一段时间，为了避免在执行操作的

过程中，由于模块还没加载出来，导致发生错误的情况，可以采用这个函数

- 简写用法：await agent.aiWaitFor('页面已经加载完成'); // 如'加载完成提示'。

这些函数组合使用时，先用aiWaitFor等待，再aiInput输入，aiTap点击，最后aiAction处理复杂流。添加重试逻辑可提升稳定性。

1.2.2 将先去确定好的流程转化为.ts文件（核心）

1.定义好要使用的ai函数并导航到目标网站
用page.goto的方式

```
import { test } from './fixture';

test('download esi data', async ({ page, ai, aiTap, aiInput, aiWaitFor, aiAction }) => {
  // 步骤 1: 导航到目标网站
  await page.goto('https://esi.clarivate.com');
  await aiWaitFor('界面加载成功', { timeoutMs: 100000 });
  await aiInput('10235501456@stu.ecnu.edu.cn', 'Email address input');
  await aiInput('Liuziyang225678!', 'Password input');
  await aiTap('Sign In button');
```

2.等待登录成功并加载主界面
用aiWaitFor

```
// 步骤 2: 智能等待登录成功并加载主界面
// 等待直到 "Top Papers by Institutions" 标题出现，表明已进入主界面
await aiWaitFor('界面中包含 "Top Papers by Research Fields"', { timeoutMs: 500000 });
```

3.将结果列表设置为“Institutions”
用aiTap

```
// 步骤 3: 确保结果列表设置为 "Institutions"  
// 这是一个保险步骤, 通常默认就是这个选项  
await aiTap('点击Result lists下方的"Research Fields"选择框');  
await aiTap('选择"Institutions"选项');
```

4. 获取每一个case

以其中一个case来举例

用aiTap、aiAction、aiWaitFor

```
await aiTap('the "Add Filter" link');  
await aiTap('the "Research Fields" link');  
await aiTap('Environment的"-"号');  
await aiAction('滚动到Geosciences的"+"号');  
await aiTap('Geosciences的"+"号');  
await aiAction('滚动到页面上方');  
await aiTap('Indicators那一行的第一个按钮下载数据');  
await aiTap('点击"CSV"');  
await aiWaitFor('界面中包含 "Top Papers by Institutions"', { timeoutMs: 100000 });
```

如上图所示,

1. 使用aiTap通过点击“-”来删除上一个选中的学科 (比如这里的“Environment”)
2. 通过aiAction来定位寻找下一个学科; 因为有的时候, 对应的学科并不在当前看到的界面上, 需要用滚轮往下滚, 来去找到它, 对于这种复杂的逻辑, 考虑用aiAction
3. 使用aiTap选中下一个选课
4. 通过aiTap选中下载的按钮
5. 通过aiWaitFor等待下载的过程

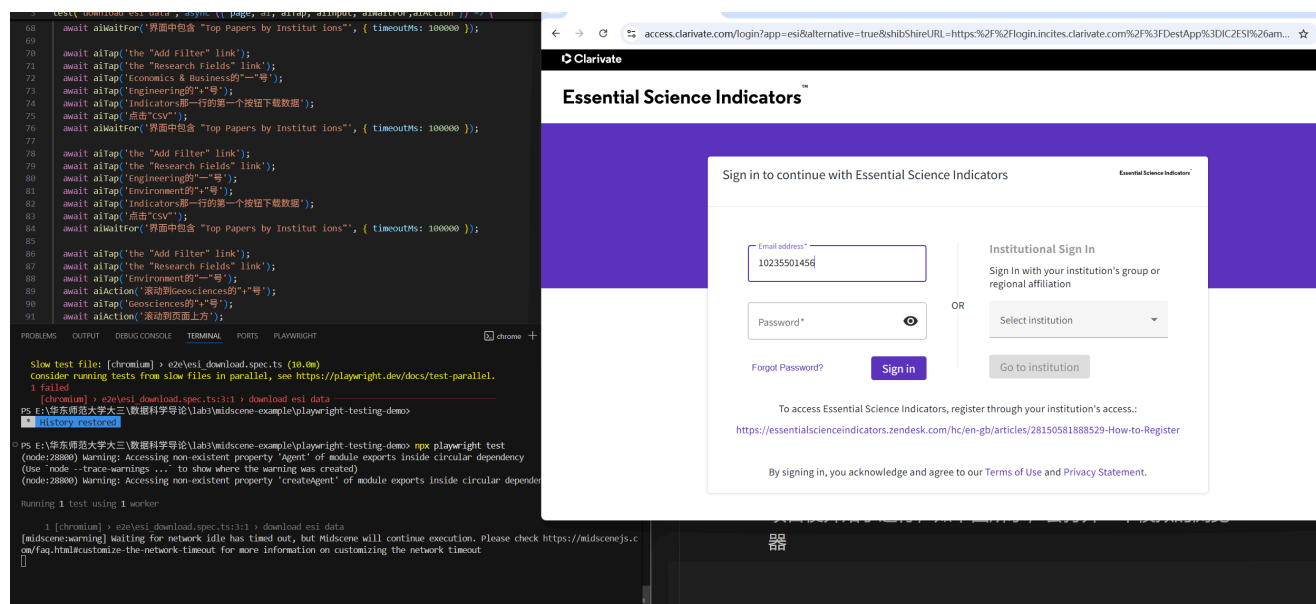
1.3 爬虫的运行和回放录制

Midscene的一个巨大优势就是可以实时看到，整个执行的过程，并且每一次执行，都会生成回放，告诉我们是执行到哪一步发生了错误，使得我们可以精准的定位问题并解决

只需要在命令行中输入

```
npx playwright test
```

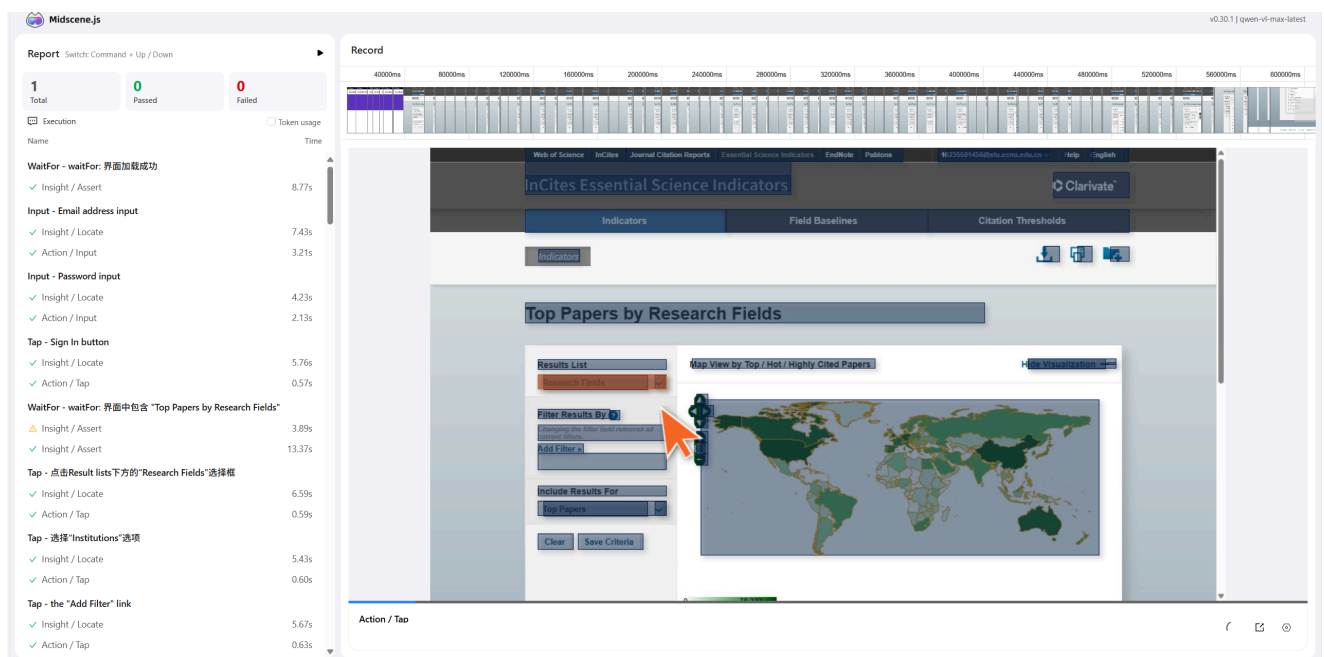
项目便开始了运行，如下图所示，会打开一个模拟的浏览器



它会自动的执行我们给他的每一步操作

在执行完成后，我们可以查看保存在本地的回放
路径是

midscene_run/report


















是一个动态的过程

具体请见文件夹中的完整回放，是html的格式

至此，我们已经完成了整个的爬虫的过程

这个是所得到的数据

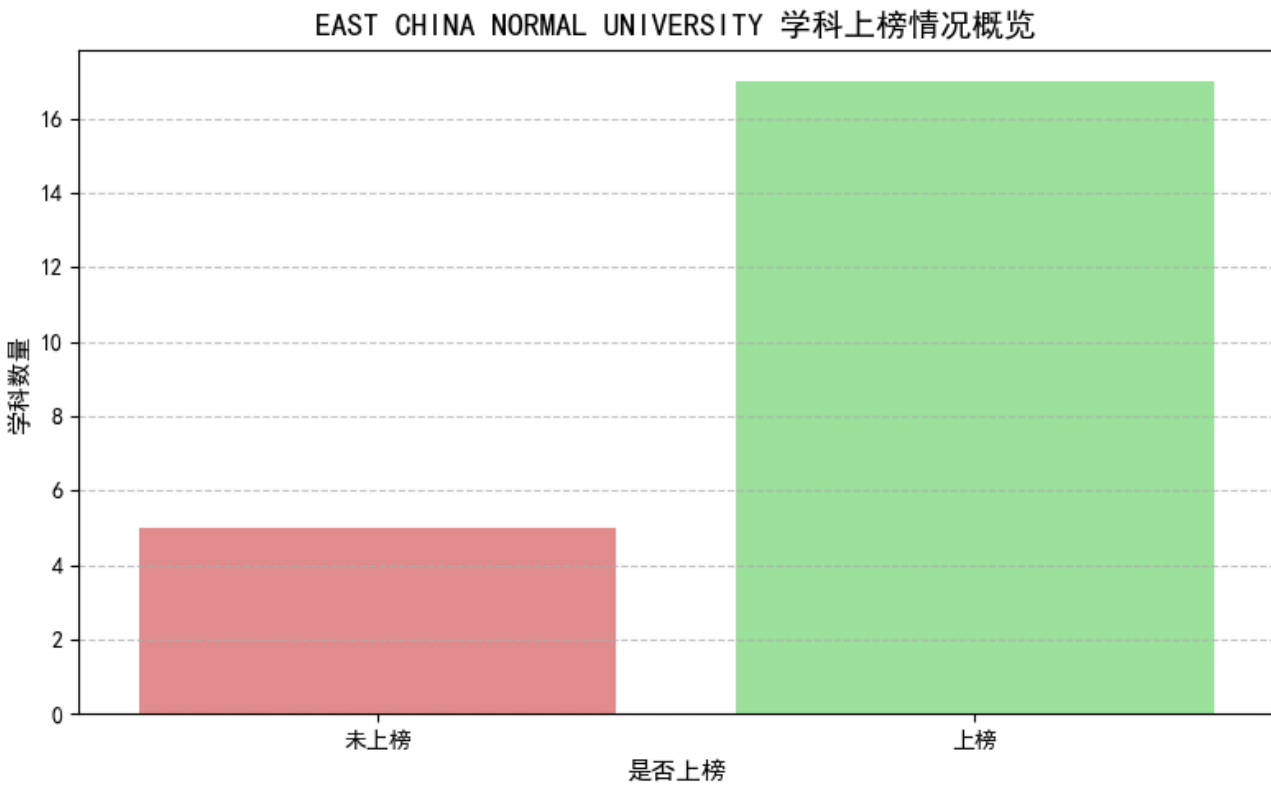
 IndicatorsExport (21).csv	2025-10-08 21:38	Microsoft Excel Co...	18 KB
 IndicatorsExport (20).csv	2025-10-08 21:38	Microsoft Excel Co...	1 KB
 IndicatorsExport (19).csv	2025-10-08 21:38	Microsoft Excel Co...	1 KB
 IndicatorsExport (18).csv	2025-10-08 21:38	Microsoft Excel Co...	149 KB
 IndicatorsExport (17).csv	2025-10-08 21:38	Microsoft Excel Co...	77 KB
 IndicatorsExport (16).csv	2025-10-08 21:37	Microsoft Excel Co...	102 KB
 IndicatorsExport (15).csv	2025-10-08 21:37	Microsoft Excel Co...	96 KB
 IndicatorsExport (14).csv	2025-10-08 21:37	Microsoft Excel Co...	16 KB
 IndicatorsExport (13).csv	2025-10-08 21:37	Microsoft Excel Co...	88 KB
 IndicatorsExport (12).csv	2025-10-08 21:37	Microsoft Excel Co...	60 KB
 IndicatorsExport (11).csv	2025-10-08 21:37	Microsoft Excel Co...	29 KB
 IndicatorsExport (10).csv	2025-10-08 21:37	Microsoft Excel Co...	122 KB
 IndicatorsExport (9).csv	2025-10-08 21:36	Microsoft Excel Co...	87 KB
 IndicatorsExport (8).csv	2025-10-08 21:36	Microsoft Excel Co...	91 KB
 IndicatorsExport (7).csv	2025-10-08 21:36	Microsoft Excel Co...	1 KB

Part 2数据分析与处理

本部分将基于对22个已经下载好的学科门类数据进行深入分析，详细阐述华东师范大学在科研领域的整体表现和各学科的相对竞争力

2.1 学科上榜情况概览

在本次分析的22个学科门类中，华东师范大学共有**17个学科成功上榜**（即进入了相应学科的全球机构榜单），同时有**5个学科未上榜**。这表明华东师范大学在全球范围内拥有较广泛的学科覆盖和一定的科研影响力，尤其在多个学科领域展现出显著实力。



2.2 上榜学科详细数据分析

下表详细列出了华东师范大学在17个上榜学科中的各项关键指标数据：

Discipline	Rank	Total_Institutions_
Chemistry	90.0	2141
Mathematics	115.0	395
Environment/Ecology	130.0	2066
Materials Science	196.0	1580
Computer Science	207.0	863
Geosciences	275.0	1175
Social Sciences, General	314.0	2407
Engineering	317.0	2787
Plant & Animal Science	395.0	1950
Psychiatry/Psychology	467.0	1147
Physics	522.0	995
Biology & Biochemistry	721.0	1649
Agricultural Sciences	845.0	1381
Neuroscience & Behavior	853.0	1298
Molecular Biology & Genetics	867.0	1169

Discipline	Rank	Total_Institutions_
Pharmacology & Toxicology	1064.0	1389
Clinical Medicine	2852.0	6754

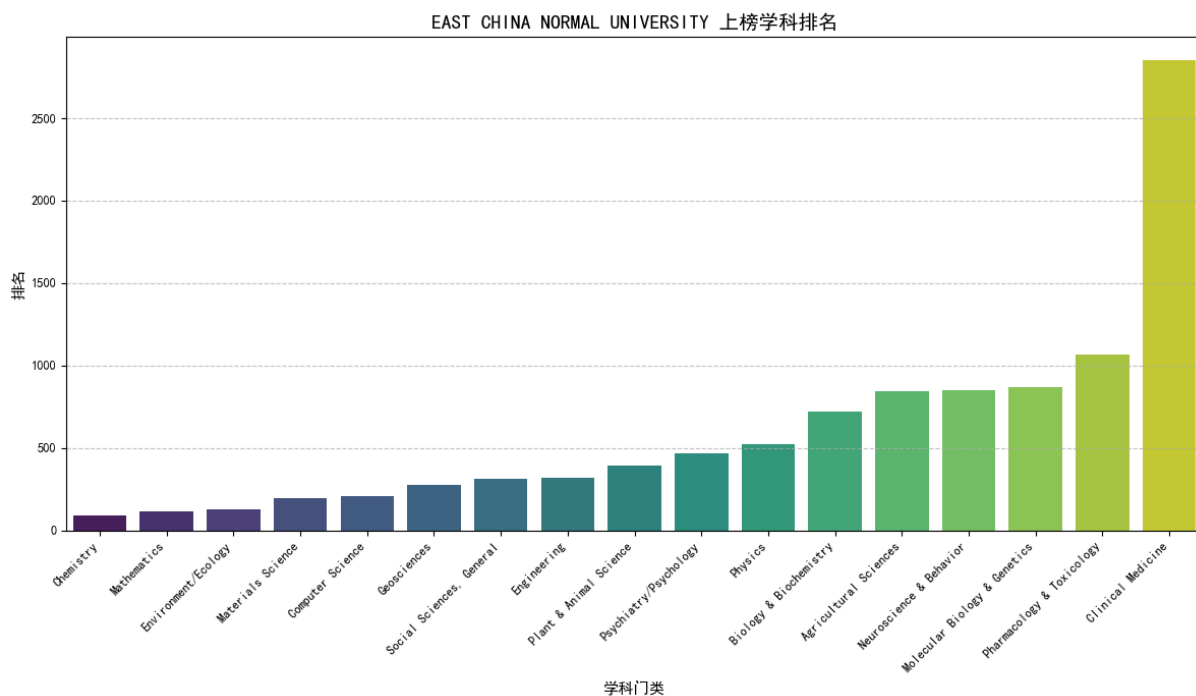
2.2.1 排名最优学科

化学 (Chemistry) 是华东师范大学排名最优的学科，位列全球第 **90** 位（在2141个机构中）。

- WOS文档数: 5420
- 总引用数: 164390
- 篇均引用: 30.33
- 高被引论文数: 157

这表明华东师范大学在化学领域的研究产出量、引用影响力及顶尖研究成果方面均表现卓越，是

学校的优势学科之一。

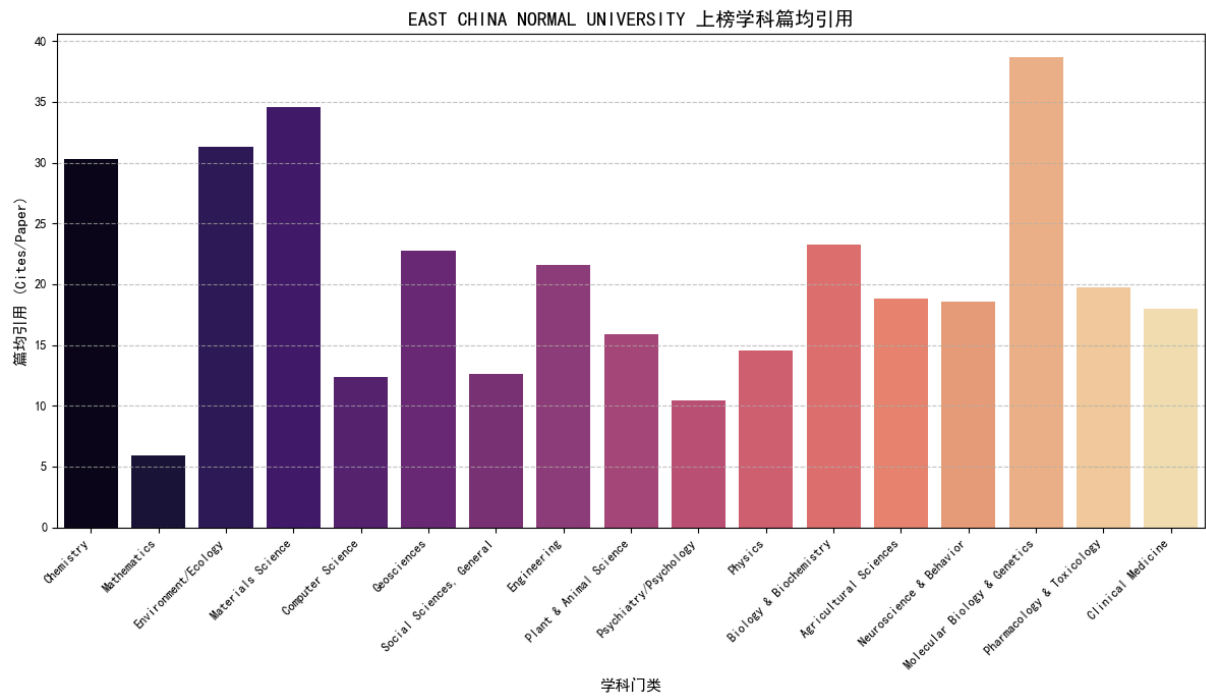


2.2.2 篇均引用最高学科

在科研成果的影响力方面，**分子生物学与遗传学 (Molecular Biology & Genetics)** 表现突出，篇均引用达到 **38.66**。尽管该学科的全球排名为867位，但其研究成果的质量和被引频率非常高，显示出强大的学术影响力。

- 排名: 867 (总计 1169 个机构)
- WOS文档数: 532
- 总引用数: 20568

- 高被引论文数: 6

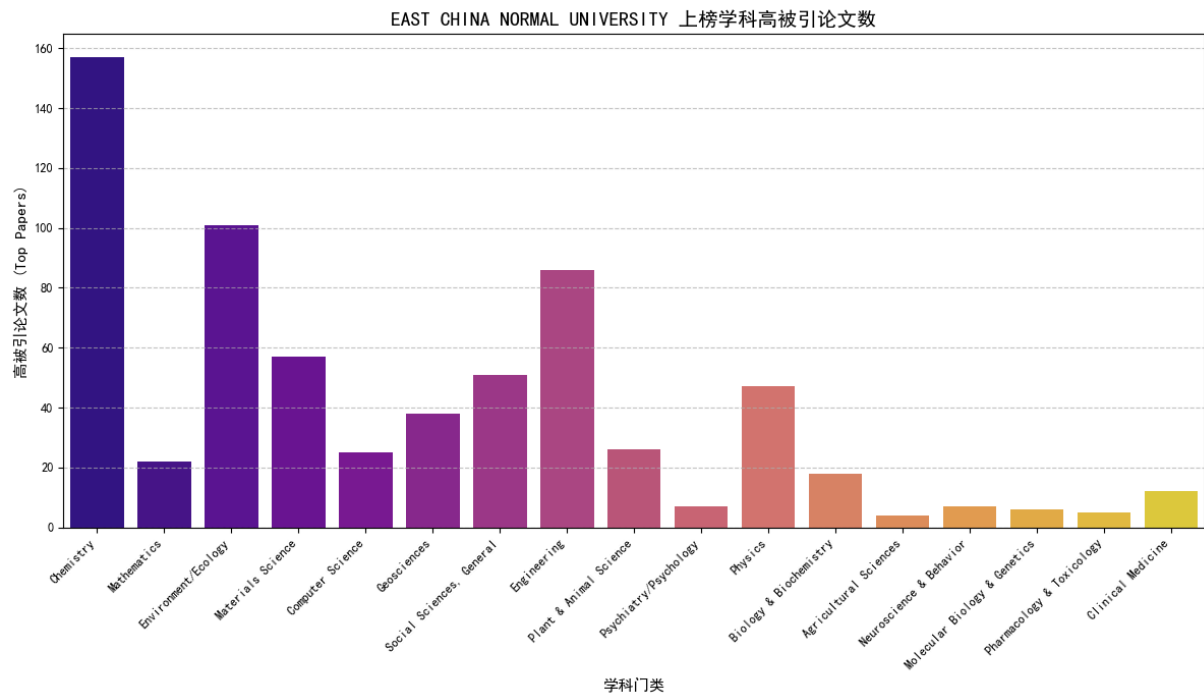


2.2.3 高被引论文数最多学科

与排名表现一致，**化学 (Chemistry)** 也是华东师范大学高被引论文数量最多的学科，达到 **157** 篇。这进一步印证了化学学科在全球范围内的领先地位和其产出高质量研究的能力。

- 排名: 90 (总计 2141 个机构)
- WOS文档数: 5420
- 总引用数: 164390

- 篇均引用: 30.33



2.3 未上榜学科分析

在本次分析的22个学科门类中，华东师范大学有5个学科未上榜。这些学科包括：

Economics & Business (经济学与商学)

Immunology (免疫学)

Microbiology (微生物学)

Multidisciplinary (多学科)

Space Science (空间科学)

2.4 总结与建议

2.4.1 总结

华东师范大学在科研领域展现出显著的实力，尤其在化学、数学、环境/生态学 等多个传统优势学科中位居全球前列，科研产出质量和影响力均表现不俗。特别地，化学学科不仅排名靠前，在高被引论文数量上也居于首位；分子生物学与遗传学则以其高篇均引用展现出深远的研究影响力。然而，学校在经济学与商学、免疫学、微生物学、多学科和空间科学等五个学科仍有待加强。

Part 3 总结

在本次实验中，我充分学习和体会了使用Midscene框架来完成对页面的爬取；其结合了AI的推理能力+Playwright 的模拟浏览器操作，使得爬虫变得更加容易和简单，可谓是收获颇丰！！！！