PyPI 项目数据爬取和分析报告

王阳 2000010767

1 选题创意介绍

PyPI（Python Package Index）作为 Python 社区的核心资源库，包含了 50 多万个项目。然而，PyPI 的官网仅提供了简单的项目检索功能，并没有提供对其中数据进行精细的分析。本项目通过对 PyPI 中项目的编程语言、主题和维护者进行系统化的爬取和分析，可以很好地帮助我们了解开源社区的动态。具体而言，本项目有以下意义：

- **技术趋势分析：**通过分析各个项目所使用的编程语言，我们可以洞察当前哪些编程语言在 Python 生态系统中最受欢迎，从而对编程语言的技术趋势有所把握。
- **主题研究：**通过对不同主题下编程语言使用情况的分析，可以帮助我们理解在特定领域（如数据库、网络开发、操作系统等），哪些编程语言和技术框架是主流。
- **维护者网络：**分析项目的维护者信息，可以揭示出在开源社区中那些活跃的维护者以及他们之间的合作关系。这不仅有助于了解对开源社区贡献最大的个人或团队，还可以为新项目的协作提供网络基础。

2 代码使用说明

2.1 数据获取

fetch_pypi_list.py 文件用于获取全部 PyPI 项目的列表，并生成 pypi_projects.json 文件。

生成 pypi_projects.json 文件后，fetch_project_list.py 文件用于获取每个项目的详细信息，包括维护者列表，使用语言列表，以及主题列表，并存储于 pypi_project_info.json 中。（代码默认使用 32 线程抓取，耗时约 2 天）

2.2 数据分析、处理、保存

load_to_sqlite.py 文件用于将 pypi_project_info.json 中的数据导入 pypi_projects.db 数据库中。

language_counts_by_topic.py 文件用于获得各个主题下各语言的使用情况，并将结果保存在 language_counts_by_topic.json 文件中。

maintainers_cooperation.py 文件用于获得维护项目数量前 50 的维护者的具体信息已经他们之间的合作关系，并将结果保存在 maintainers_cooperation.json 文件中。

2.3 数据展示

create_bar_chart.py 文件用于生成各个主题下编程语言使用情况直方图。

create_word_cloud.py 文件用于生成维护者维护项目数量词云。

create_graph.py 文件用于生成维护者合作关系图。

3 效果展示

所有的图表都可以从该链接进入并查看

3.1 编程语言使用情况直方图

我们统计了不同主题下各个编程语言的使用情况，并将其做成了直方图。（由于 PyPI 中所有的项目都提供 Python 接口，因此 Python 语言不做统计）。此外，通过点击直方图上方的按钮，可以选择不同主题的数据查看。直方图横轴表示编程语言，纵轴表示该使用该编程语言的项目数量。我们的直方图很好地反映了各个主题下主流的编程语言。比如，对于网络页面开发，JavaScript 是主导语言，对于有关数据库的项目，SQL 是使用最多的语言。

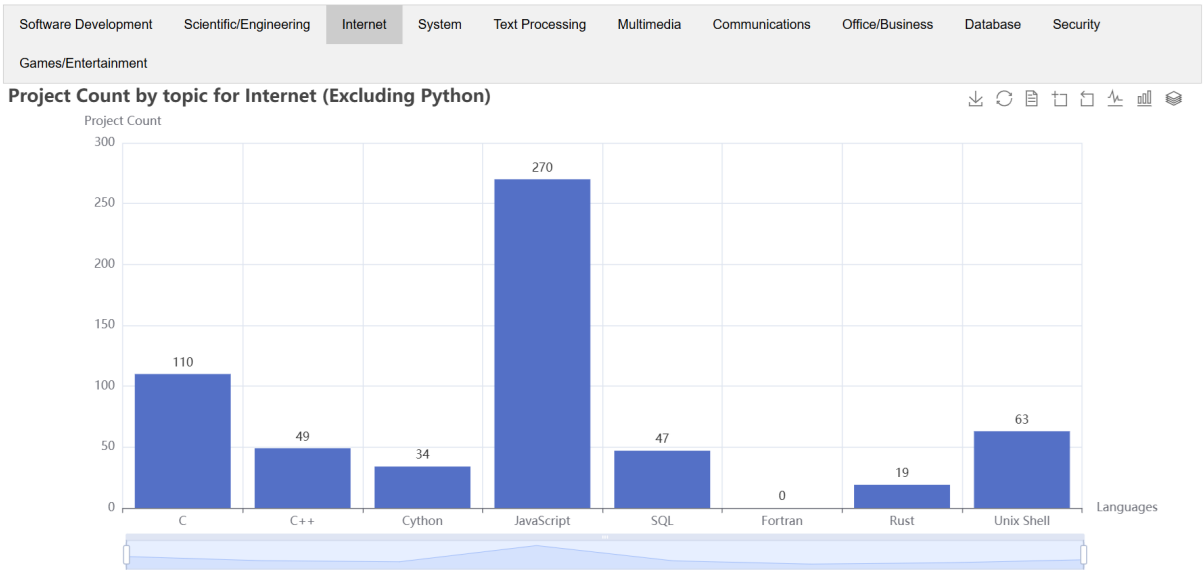

图 1: 网络主题下所使用编程语言情况

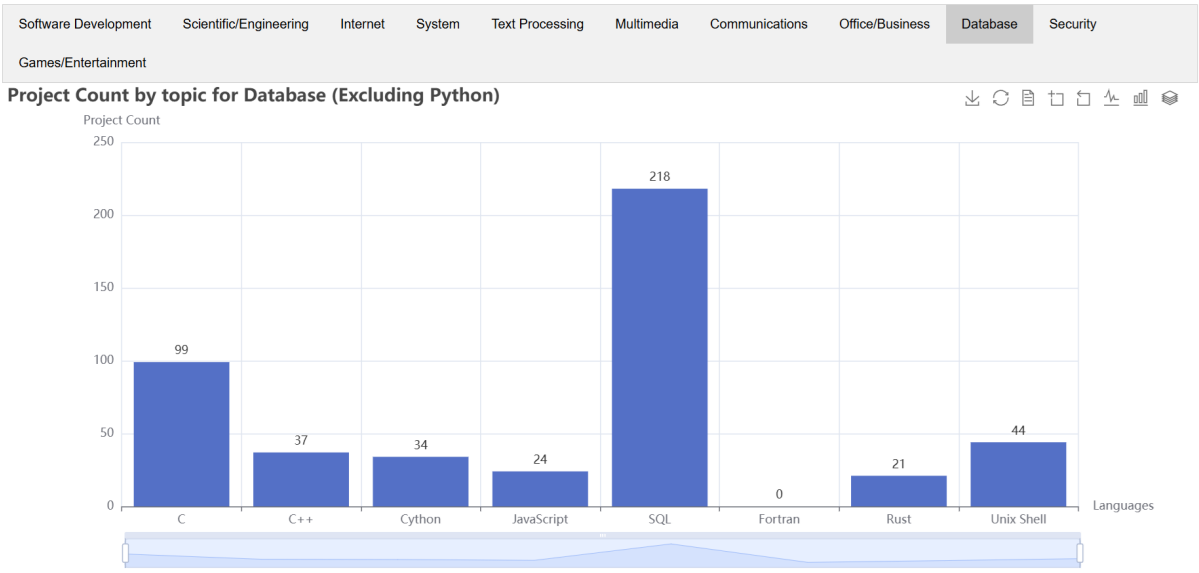

图 2: 数据库主题下所使用编程语言情况

3.2 维护者词云

我们选取了维护项目数量前 50 的维护者，根据其维护项目数量的多少制作了词云。在该词云中，可以很直观地看到 OCA 组织维护了最多项目，很多公司如 Google, Microsoft, Nvidia 等也维护了很多 PyPI 的开源项目。当然，该词云只是通过维护项目数量这一简单标准制作，并不能精确地反映用户的贡献。

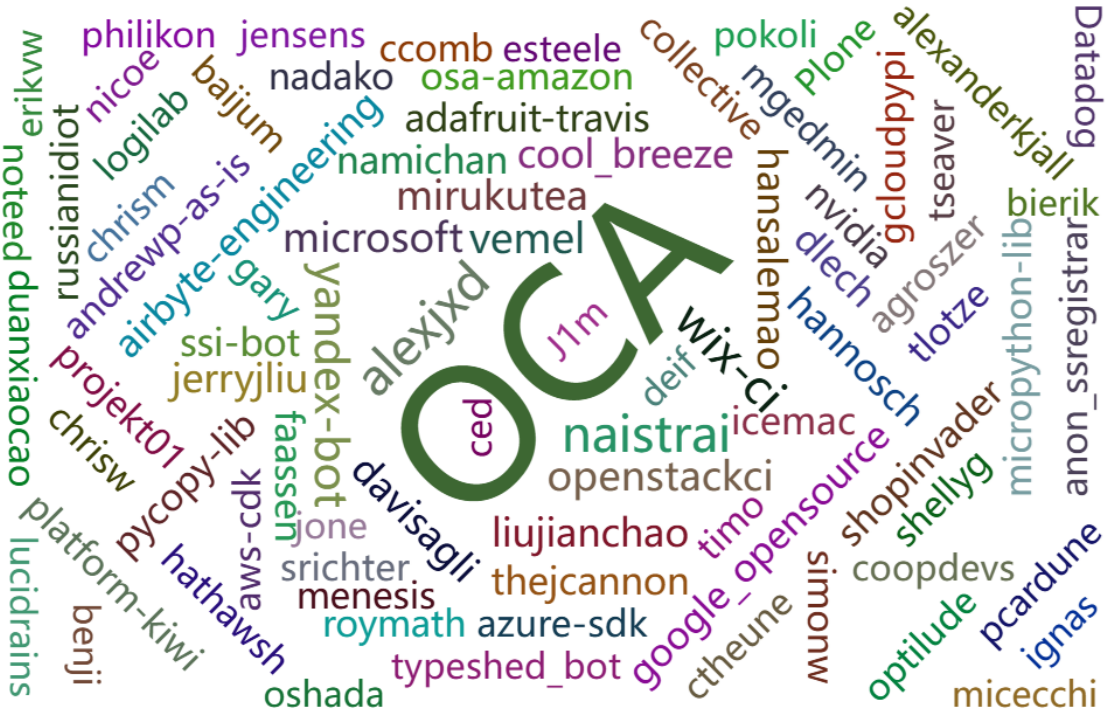

图 3: 数据库主题下所使用编程语言情况

3.3 维护者合作关系图

在维护项目数量前 50 的维护者中，我们根据他们的合作情况制作了关系图。具体而言，图的节点大小代表了该维护者所维护项目的数量，边的粗细代表了两个维护者共同维护项目数量的多少。此外，为了更好的可视化效果，我们去除了图中的孤立点，也就是没有和其他维护者合作的维护者。该图很好地展示了各维护者之间的合作关系。如图中间的团展示了众多个人维护者之间紧密地合作，图边缘的 google_opensource 和 gcloudpypi 展示了 Google 公司内部之间的合作。

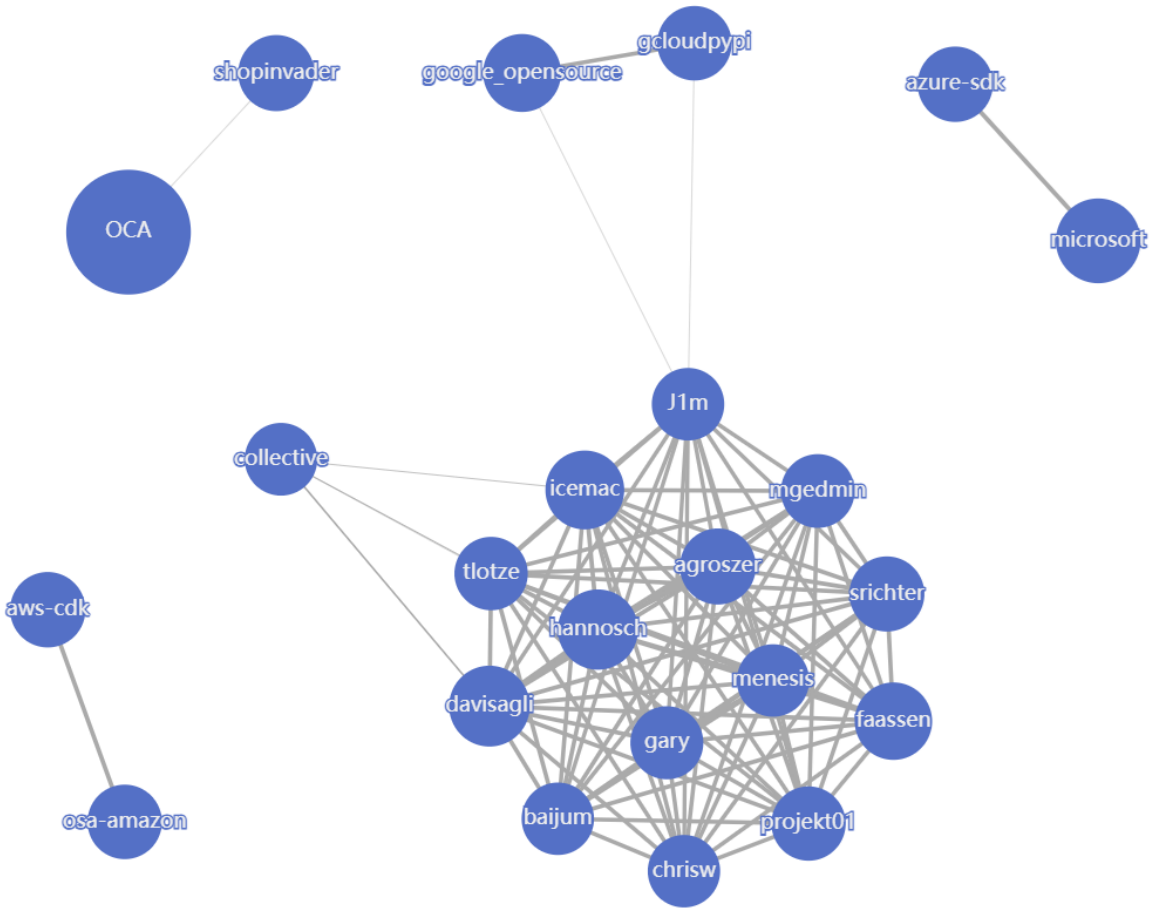

图 4: 数据库主题下所使用编程语言情况