



评估开源数据集的长期使用模式：一种引用网络视角

参赛选手：张雨昂 张雨欣 彭佳恒
指导教师：王伟

CONTENTS

目 录

1 / 介绍

2 / 方法

3 / 实验部分

4 / 开源协作

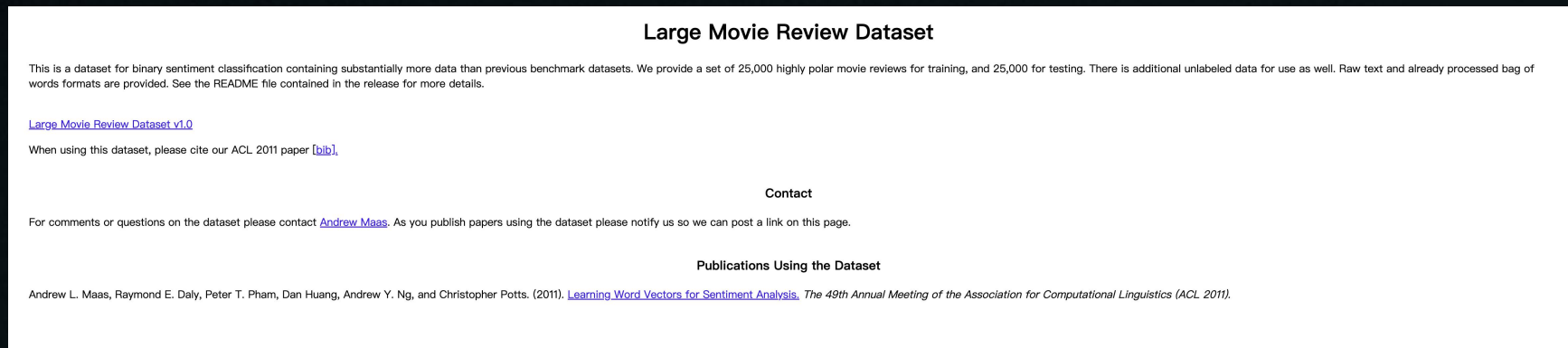


介绍

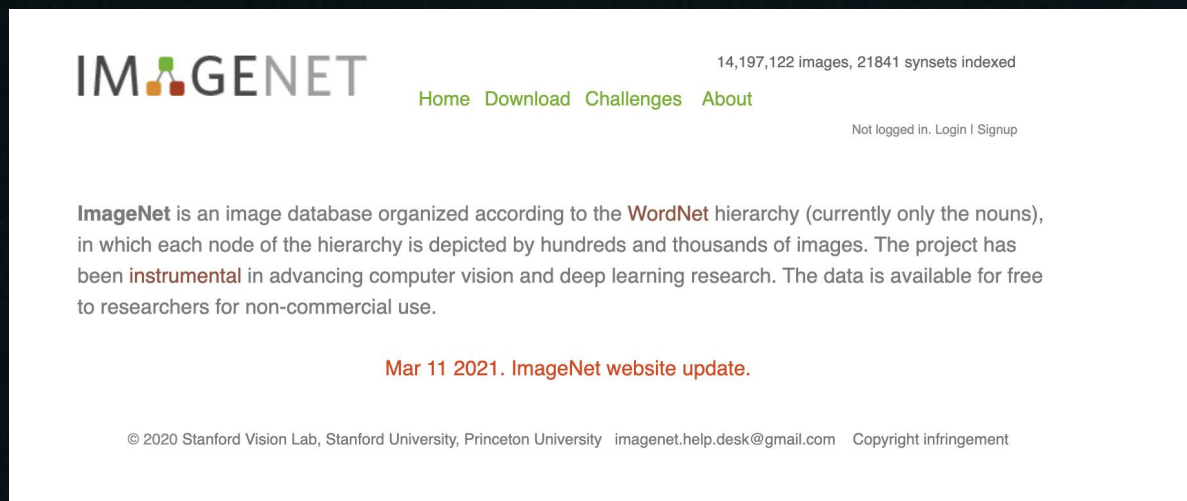


- 数据集的评估是评价学任务的基础和根本。
- 评估数据集的使用模式对选择合适的数据集以及背后的科学研究任务的展开具有重要影响。
- 研究人员在选择合适的数据集时经常遇到重大挑战，因为他们不了解这些数据集的使用方式。
- 许多著名的开源数据集已经很成熟，其官网和对应的Github仓库多年未更新，但仍被大量研究人员广泛使用。

01 / 挑战 两大著名的数据集的官网截图



IMDB



ImageNet

这是两大著名的数据集的官网截图示例，官网基本上只提供了下载数据集的链接地址等少数信息。从官网上我们很难分析出其近期的被使用情况，以及长期的使用使用模式



01 / 挑战 两大著名的数据集的Github仓库

IMDB

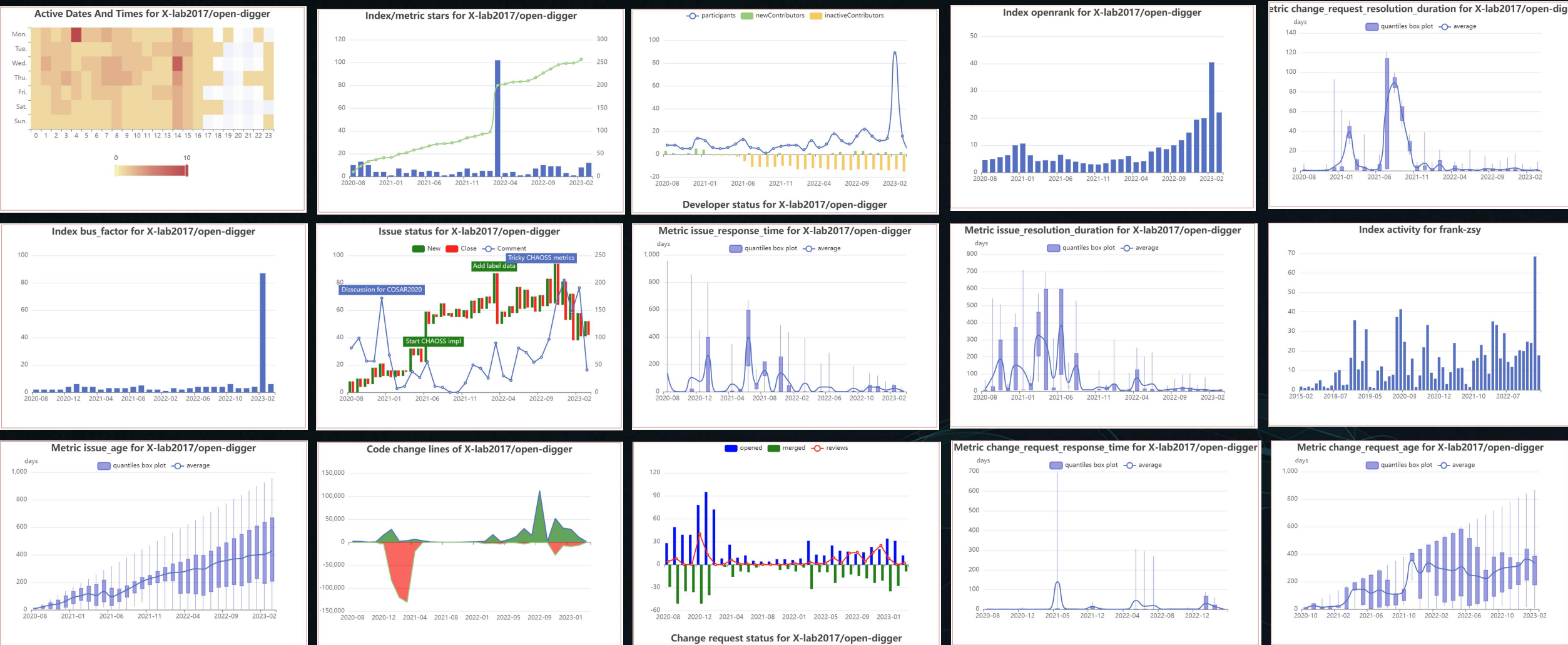
configs	add 3dcnn serarch (#8)	last year
docs	* update to v2 version	last year
modelscope	* update to v2 version	last year
requirements	* update to v2 version	last year
tinynas	Update super_res_k1kxk1_mutator.py	last year
tools	* update to v2 version	last year
.gitignore	* update to v2 version	last year
LICENSE	* update to v2 version	last year
MANIFEST.in	* update to v2 version	last year
NOTICE	* update to v2 version	last year
README.md	Update README.md	last year
get_started.md	* update to v2 version	last year
installation.md	* update to v2 version	last year
setup.cfg	* update to v2 version	last year
setup.py	* update to v2 version	last year

assets	Updated attention visualization photo.	5 years ago
models	Renamed tokenize() function to tokenize_and_encode().	5 years ago
utils	Changed to (conventional) method for acquiring [PAD] to...	4 years ago
LICENSE.md	Initial commit.	5 years ago
README.md	Updated account link.	5 years ago
baseline_main.py	Removed unnecessary line, and seperated loader and dat...	5 years ago
download_imdb_dataset.sh	Added shebang & directory creating lines.	5 years ago
main.py	Removed unnecessary line, and seperated loader and dat...	5 years ago
visualize_attention.py	Added more comments & fixed typos.	5 years ago

ImageNet

这是其对应的Github仓库截图示例，可以看出，这些数据集的相关文件已经很长时间没有更新了，其仓库背后的Github行为数据也寥寥无几

01 / 常见的Github洞察指标举例



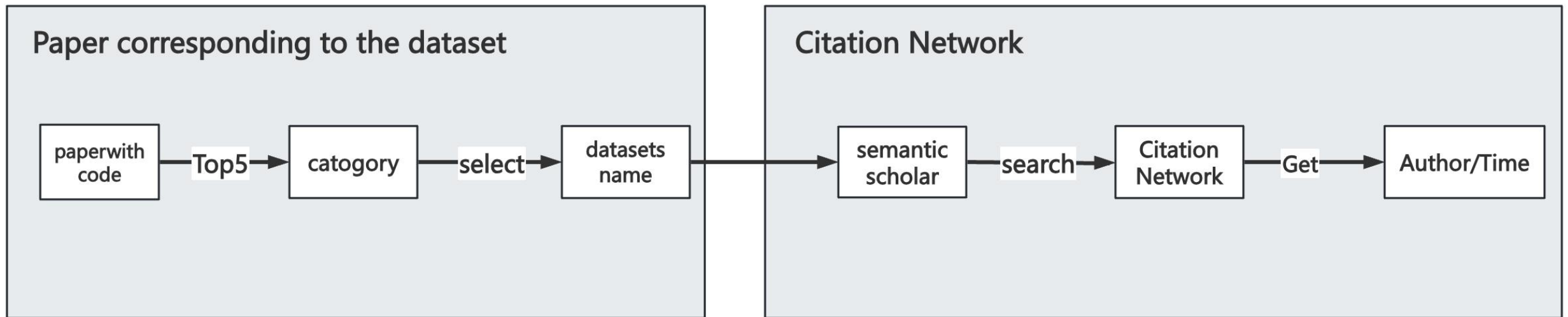
这是常见的Github日志数据洞察指标：如活跃度、Issue数、star数、OpenRank值等等，背后都需要Github行为数据支撑

- 许多著名的开源数据集已经完善且多年未更新，但仍被大量研究人员广泛使用。
- 其对应的官方网站上也没有提供被使用情况的相应信息，同时对应的GitHub仓库中也没有足够的行为活动数据以支撑我们使用传统的开源指标（活跃度、Issue数、star数、OpenRank值等等）来评估长期使用模式。
- 但是我们发现，一般开源数据集都有一篇对应的学术论文（由数据集作者撰写）发表，可以将数据集与其对应的学术论文的引用网络建立连接，来分析此数据集的近期使用情况以及长期使用模式。



方法

02 / 方法框架



左侧：可以通过Paperwithcode网站获取这些开源数据集的模态/种类，每个开源数据集对应的Github仓库，数据集的名字

右侧：可以通过Semantic Ccholar网站，获取到这些开源数据集对应的学术论文，同时通过搜索其论文的引用网络，获取其被引用信息，来分析其长期使用模式

02 / 挑选具有代表性的数据集

category	small-dataset	medium dataset	large dataset
Images	CityFlow	Food-101	Fashion-MNIST
Texts	FinQA	CommonsenseQA	GLUE
Videos	CRVD	OTB	UCF101
Audio	XD-Violence	Common Voice	Librispeech
Medical	VerSe	ChestX-ray14	MIMIC-III

- 我们获取了paperwithcode网站上Top5模态类型的数据集：分别是Image、Texts、Videos、Audio和Medical这五个模态类别
- 从这五种类型中，每一种类别分别挑选具有代表性的小、中、大规模的数据集
- 小规模数据集被定义其对应的学术论文被引用少于500次的数据集，中等规模的数据集是指引用次数在500到5000次之间的数据集，大规模数据集是指引用次数超过5000次的数据集。

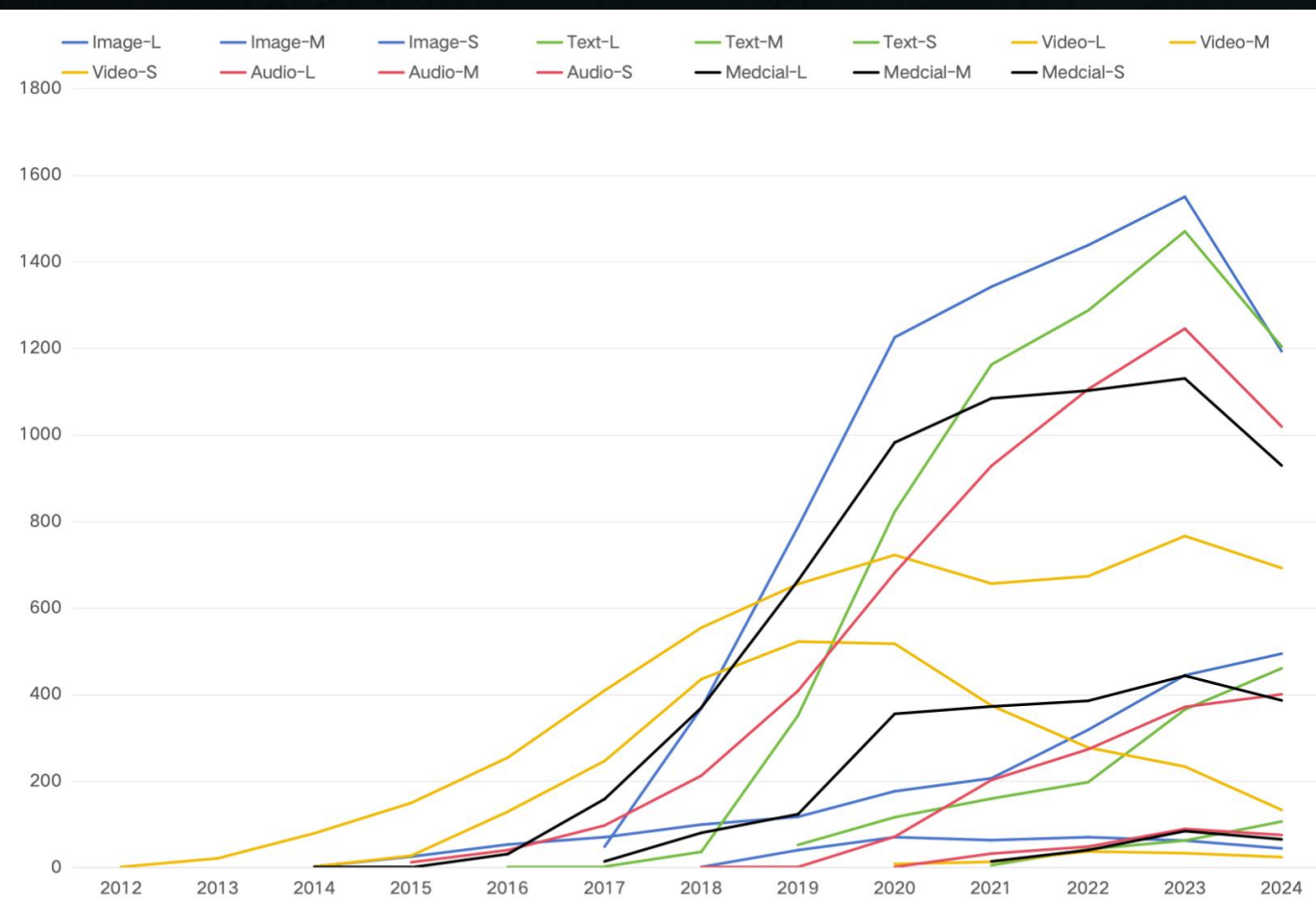


实验部分

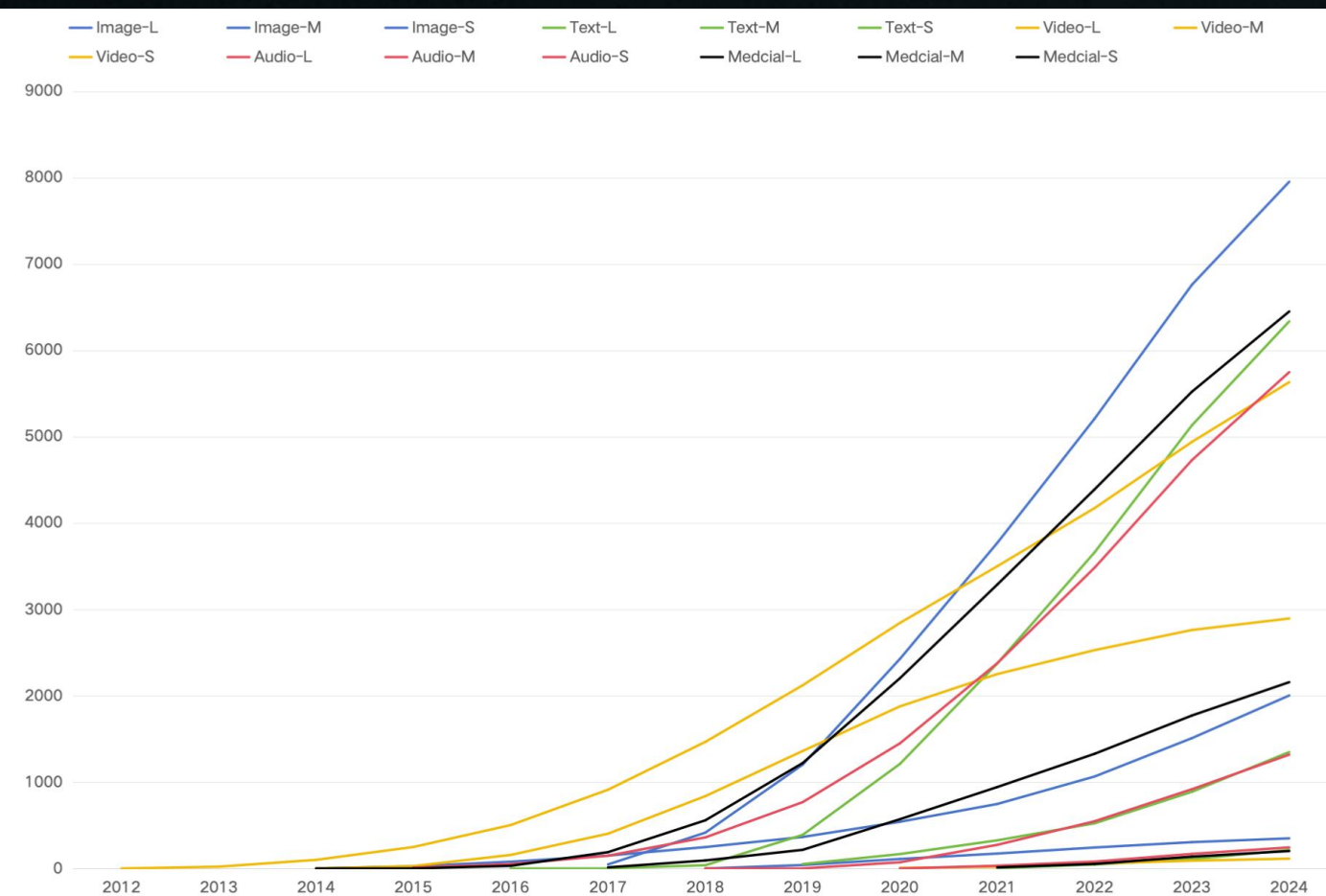


category	small-dataset	medium dataset	large dataset
Images	7949	2003	350
Texts	6334	1349	213
Videos	5629	2898	115
Audio	5752	1319	245
Medical	6449	2157	203

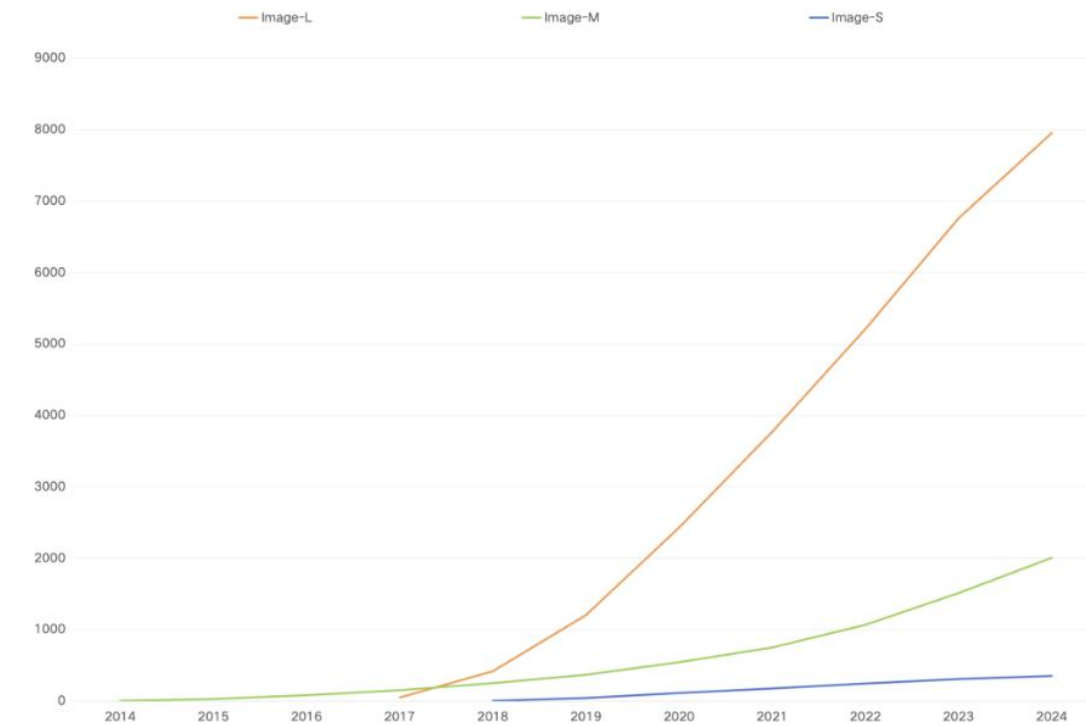
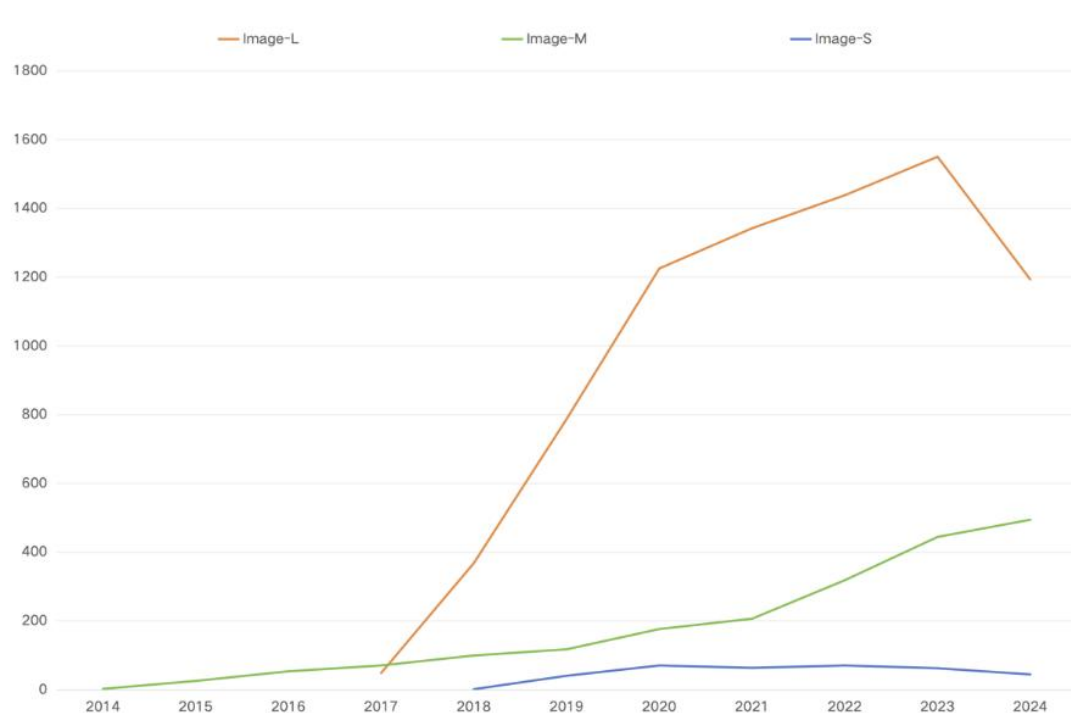
- 除了使用OpenDigger的开源行为日志数据外
- 我们还通过上述方法补充了这些引用数据信息
- 一共补充的数据量为：45965条
- 所有的引用信息均以开源至参赛仓库中：<https://github.com/TIAI-ThisisAI/OpenRank2024>



- 从累计引用量来看，Image-L也是最突出的一类。
- 自2017年以来，其总被引量呈指数级增长，到2024年远远超过其他类别，这表明大规模图像数据集在研究领域占据中心和主导地位。
- 同样，Text-L和Medical-L的累积被引量也快速上升，尤其是Text-L，其增长轨迹自2020年以来几乎与Image-L平行，表明大规模文本数据集与大规模图像数据集的差距正在逐渐缩小。



- 大规模数据集，如Image-L、Text-L和Medical-L，显示出显著的引文增长，其中Image-L在2022年达到峰值，但保持了较高的引文数量。
- 尽管近年来有所放缓，但在NLP和医疗人工智能的推动下，大规模图像和文本数据集仍在增长。
- 与图像和文本数据集相比，Audio-L和Video-L数据集表现出更慢、更稳定的增长。
- 中等规模的数据集，如Image-M和Text-M，显示引文以较慢的速度逐渐增加。
- 小规模数据集（如音频和视频）最初出现增长，但在2020年后停滞或下降，反映出人们的兴趣降低。

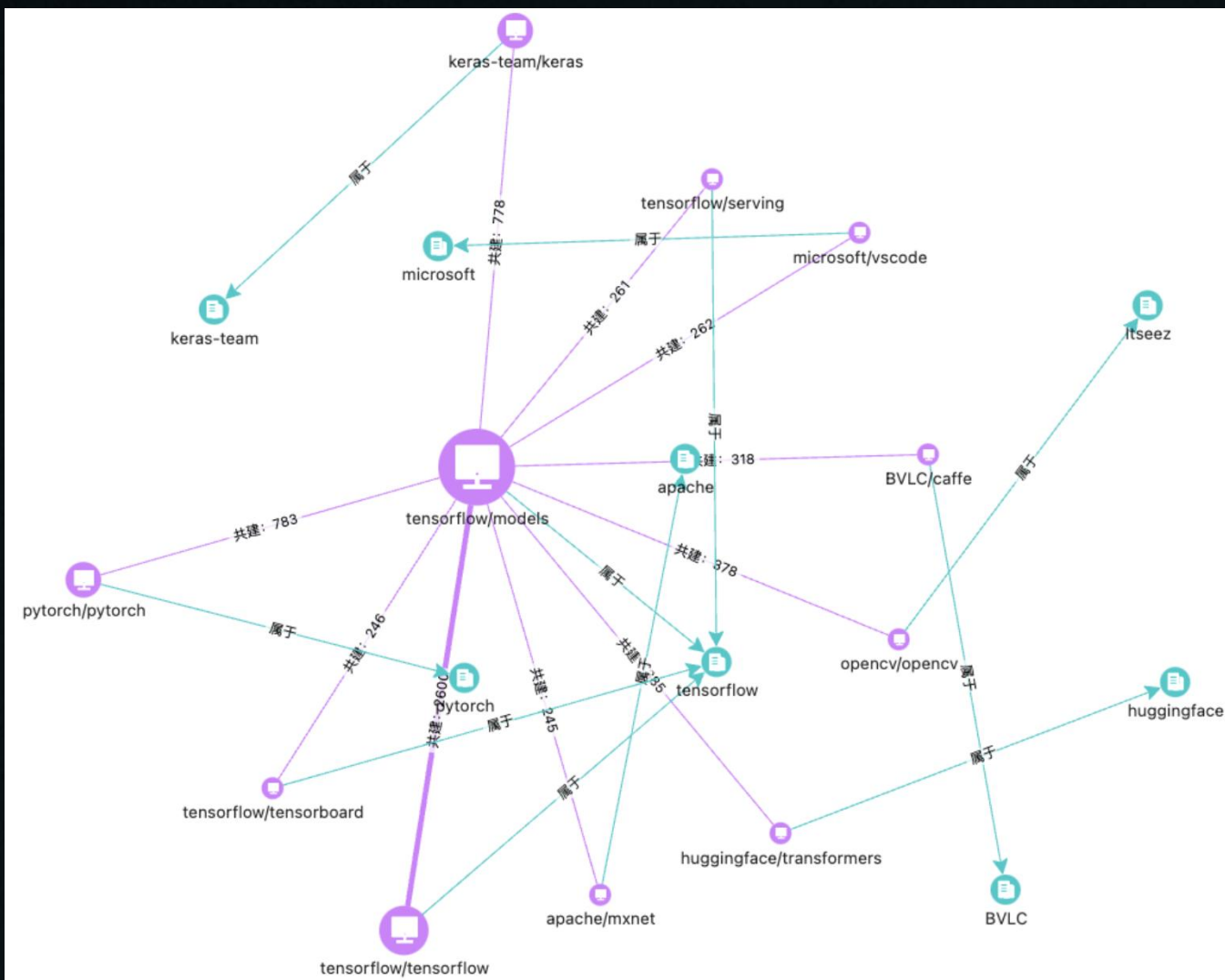


与中型（Image-M）和小型（Image-S）数据集相比，大型图像数据集（Image-L）的增长曲线明显更陡峭。Image-L 的累计引用量呈指数级增长，始于 2017 年左右，并在 2020 年后急剧加速。到 2024 年，Image-L 的总引用量超过 8,000，远远超过 Image-M 和 Image-S。快速增长凸显了大型图像数据集的持续流行和影响力，这得益于它们在图像识别、目标检测等深度学习任务中的关键作用。

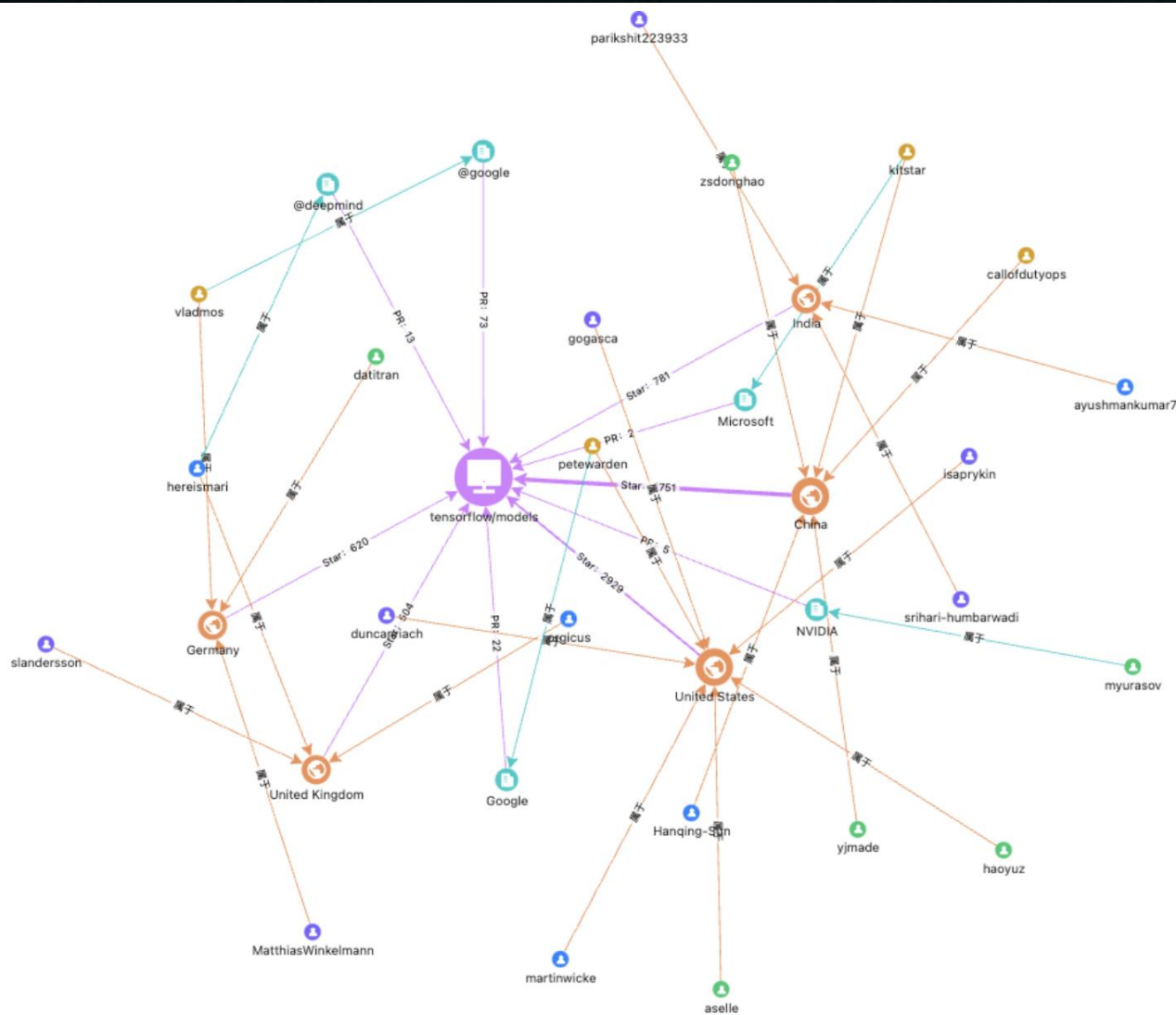


- 以image中的大型数据集对应的最高star数的Github仓库为例，分析了项目的贡献协作网络
- 可以看出此数据集吸引了大量的有影响力的贡献者进行贡献，包括MarkDaoust, nealwu, cshjtn 等
- 同时还吸引到了googlebot这种知名公司的机器人以及tensorflowbutler 等进行自动化协作，显示了项目中 CI/CD 流程和自动化协作的重要

03 / 实验结果 项目生态协作网络



- 以image中的大型数据集对应的最高star数的Github仓库为例，分析了该仓库的项目生态协作网络
- 可以看出此数据集吸引了大量的知名仓库之间的互相协作：如pytorch、微软的vscode、huggingface社区以及huggingface的transformers等等
- 同时也可以看出越被大量引用的知名数据集和仓库，他们能更吸引到更知名的仓库和开发者来产生协作关系，从而带来“富俱乐部”效应



- 以image中的大型数据集对应的最高star数的Github仓库为例，分析了该仓库的项目社区协作网络
- 可以看出此数据集吸引了大量知名的开发者和社区/公司，除了中国以外，还有德国、英国、美国、印度等大量的开发者和社区/公司与该仓库进行了协作
- 同时还吸引到了google、nvidia、microsoft这种知名公司产生协作关系

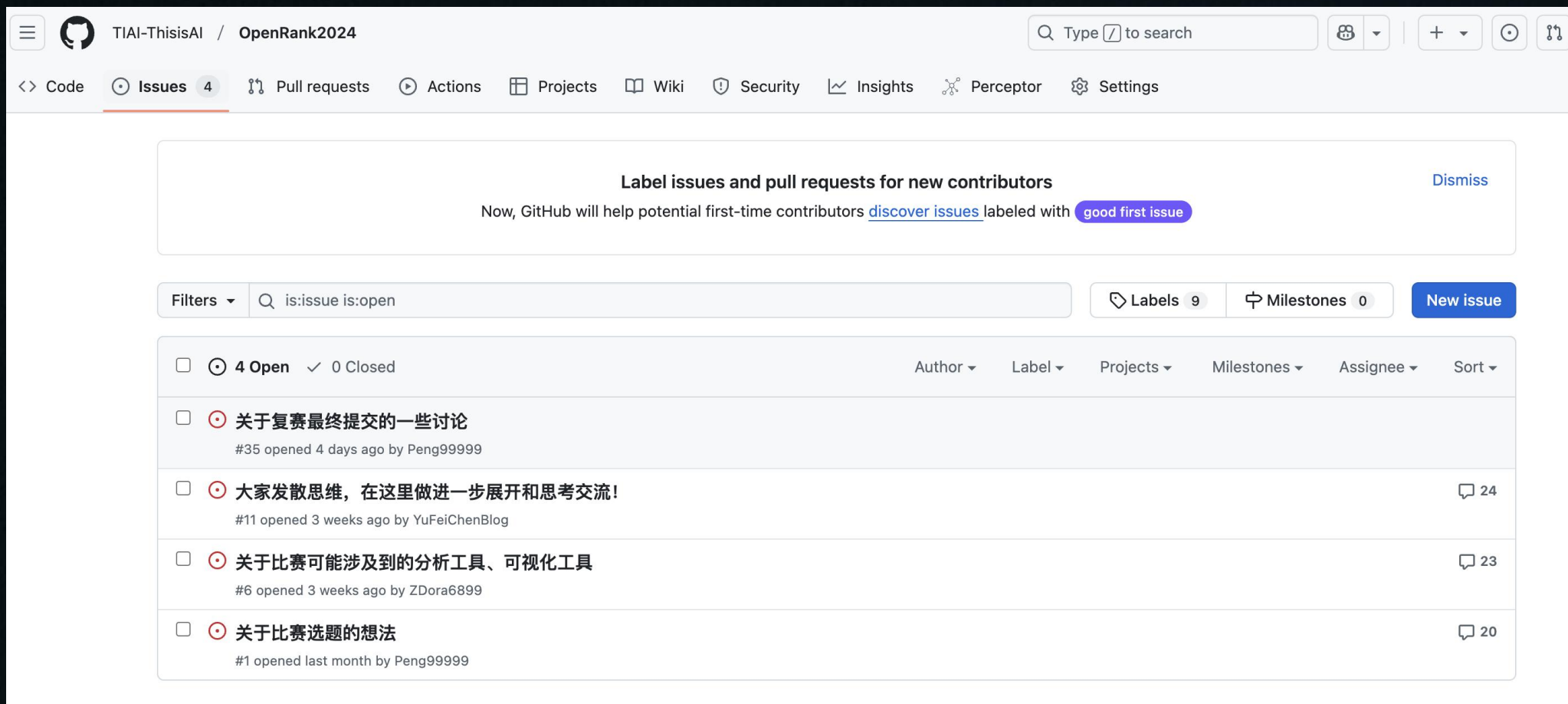


开源协作



04 / 开源协作 issue讨论

- 本团队三个成员自从参加比赛开始就一直采取开源协作的方式
- 从讨论选题开始就在github上以issue的形式进行讨论，共计4个issue，六十多条评论回复：



The screenshot shows the GitHub interface for the repository 'TIAI-ThisisAI / OpenRank2024'. The 'Issues' tab is selected, showing 4 open issues. A notification banner at the top encourages labeling issues for new contributors. The issues list includes titles, opening dates, authors, and comment counts.














Label issues and pull requests for new contributors
Now, GitHub will help potential first-time contributors [discover issues](#) labeled with **good first issue** [Dismiss](#)















Filters Labels 9 Milestones 0 [New issue](#)

<input type="checkbox"/>	4 Open ✓ 0 Closed	Author	Label	Projects	Milestones	Assignee	Sort
<input type="checkbox"/>	关于复赛最终提交的一些讨论 #35 opened 4 days ago by Peng99999						
<input type="checkbox"/>	大家发散思维，在这里做进一步展开和思考交流！ #11 opened 3 weeks ago by YuFeiChenBlog						24
<input type="checkbox"/>	关于比赛可能涉及到的分析工具、可视化工具 #6 opened 3 weeks ago by ZDora6899						23
<input type="checkbox"/>	关于比赛选题的想法 #1 opened last month by Peng99999						20

04 / 开源协作 PR贡献

- 本项目也是完全以开源协作的方式进行代码迭代，以共计31个PR的形式完成本项目的开发

<input type="checkbox"/>	 Cross-field citation #22 by ZDora6899 was merged 3 weeks ago
<input type="checkbox"/>	 Multi-dimensional evaluation indicators #21 by YuFeiChenBlog was merged 3 weeks ago
<input type="checkbox"/>	 Thoughts on developing a standardized evaluation system #20 by YuFeiChenBlog was merged 3 weeks ago
<input type="checkbox"/>	 Quantifying the activity and influence of datasets #19 by ZDora6899 was merged 3 weeks ago
<input type="checkbox"/>	 Add Get the citation part function #18 by Peng99999 was merged 3 weeks ago
<input type="checkbox"/>	 Specific meaning #17 by ZDora6899 was merged 3 weeks ago
<input type="checkbox"/>	 Application scenarios and practical significance #16 by YuFeiChenBlog was merged 3 weeks ago
<input type="checkbox"/>	 Citation Networks and Analysis #15 by ZDora6899 was merged 3 weeks ago
<input type="checkbox"/>	 A brief introduction to the methodology #14 by YuFeiChenBlog was merged 3 weeks ago
<input type="checkbox"/>	 Add the description part #13 by Peng99999 was merged 3 weeks ago
<input type="checkbox"/>	 Add the SearchPaper function #12 by Peng99999 was merged 3 weeks ago
<input type="checkbox"/>	 Update readme.md #10 by ZDora6899 was merged 3 weeks ago
<input type="checkbox"/>	 Update readme.md #9 by YuFeiChenBlog was merged 3 weeks ago

<input type="checkbox"/>	 0 Open  31 Closed
<input type="checkbox"/>	 Count the Total and each year citations #34 by Peng99999 was merged 4 days ago
<input type="checkbox"/>	 Add a new method of Get Citation #33 by Peng99999 was merged last week
<input type="checkbox"/>	 Construct test_data without domain labels #32 by ZDora6899 was merged last week
<input type="checkbox"/>	 Construct train_data with domain labels #31 by YuFeiChenBlog was merged last week
<input type="checkbox"/>	 Detailed Introduction #30 by ZDora6899 was merged last week
<input type="checkbox"/>	 Summary of the code #29 by YuFeiChenBlog was merged last week
<input type="checkbox"/>	 Cross-field citation analysis #28 by YuFeiChenBlog was merged last week
<input type="checkbox"/>	 Sampling inspection and manual verification #27 by ZDora6899 was merged last week
<input type="checkbox"/>	 Code example for trend analysis #26 by YuFeiChenBlog was merged 2 weeks ago
<input type="checkbox"/>	 Code example for automated classification #25 by ZDora6899 was merged 2 weeks ago
<input type="checkbox"/>	 Methods for calculating cross-field citations #24 by YuFeiChenBlog was merged 2 weeks ago
<input type="checkbox"/>	 Interpretation of cross-field references #23 by ZDora6899 was merged 2 weeks ago



THANK YOU

