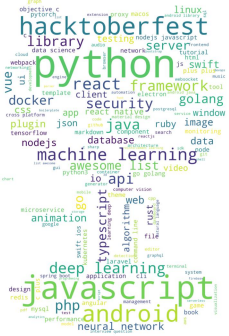
智能与计算学部
COLLEGE OF INTELLIGENCE AND COMPUTING

天津大学
Tianjin University

No One Knows GitHub Better than Data



汇报人：杨晨

小组成员：杨晨，刘冠中，胡天雨，旦增晋美

2021年01月08日

提　纲

1 课题背景及意义

GitHub拥有超过一千万个git仓库，是目前世界上最主流的开源平台之一。

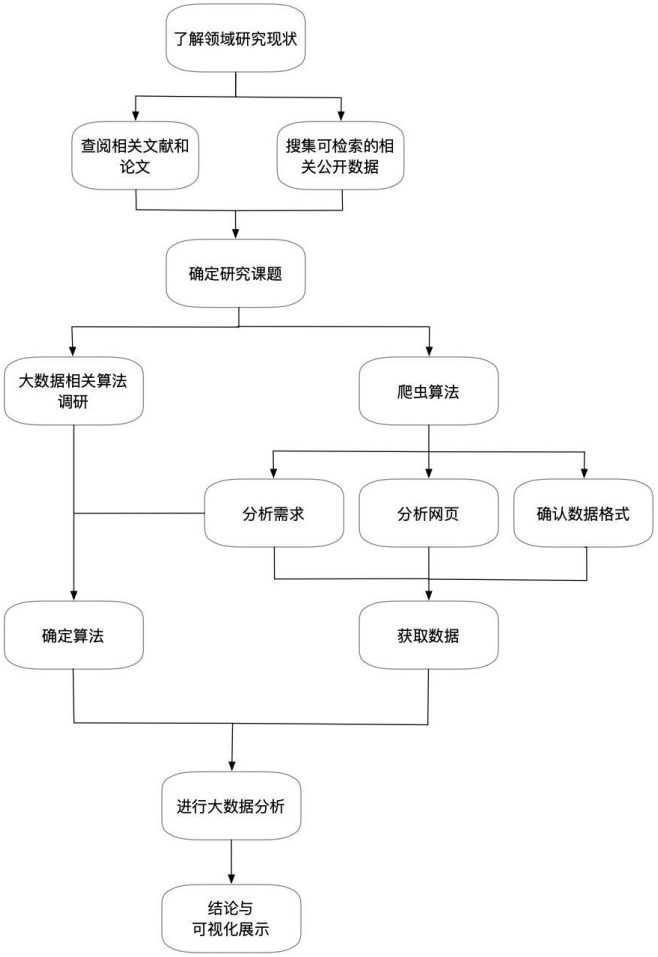
通过对GitHub大数据的分析可以让我们从各个方面对如今开源社区的发展现状，以及对一段时间以来开源社区的发展趋势有更加深刻的了解。

通过对GitHub成千上万Contributor数据的分析也可以让我们看到和我们从事相同工作的一群人的特征和群体画像，更好的了解自己所在的环境，拥有更多的角度来审视自己。

对开源社区发展的分析和技术趋势与技术间联系的分析，也对我们的学习有着一定的参考价值。

提　纲

2 技术路线


了解领域研究现状
查阅相关文献和论文
搜集可检索的相关公开数据
确定研究课题
大数据相关算法调研
爬虫算法
分析需求
分析网页
确认数据格式
确定算法
获取数据
进行大数据分析
结论与可视化展示

提　纲

3 数据采集

ossu / computer-science

Code | Issues 8 | Pull requests 3 | Actions | Projects | Wiki | Security | Insights

master | 2 branches | 5 tags | Go to file | Add file | Code

waciumawanjohi Remove prereq not mentioned by course creators 645917a yesterday 861 commits

master

Commits on Jan 6, 2021

Remove prereq not mentioned by course creators — waciumawanjohi committed yesterday — Verified — 645917a

Commits on Jan 3, 2021

Move CS50 to Courses/Extras — waciumawanjohi committed 5 days ago — d4c50dc

Commits on Dec 30, 2020

Remove direct link to issues — waciumawanjohi committed 9 days ago — Verified — 1d5b319

Commits on Dec 21, 2020

Sharpen FAQ answer language — waciumawanjohi committed 18 days ago — Verified — e94c974

Add section to FAQ about alts — Alaharon123 authored and waciumawanjohi committed 18 days ago — 33a03e7

add two new books on systems — Uniminin authored and waciumawanjohi committed 18 days ago — 0b2f86e

About

🎓 Path to a free self-taught education in Computer Science!

computer-science awesome-list courses curriculum

Septieme7

Unfollow ···

9 followers · 11 following · ☆ 27

天津大学/Tianjin University
Tianjin, China
septieme7.github.io

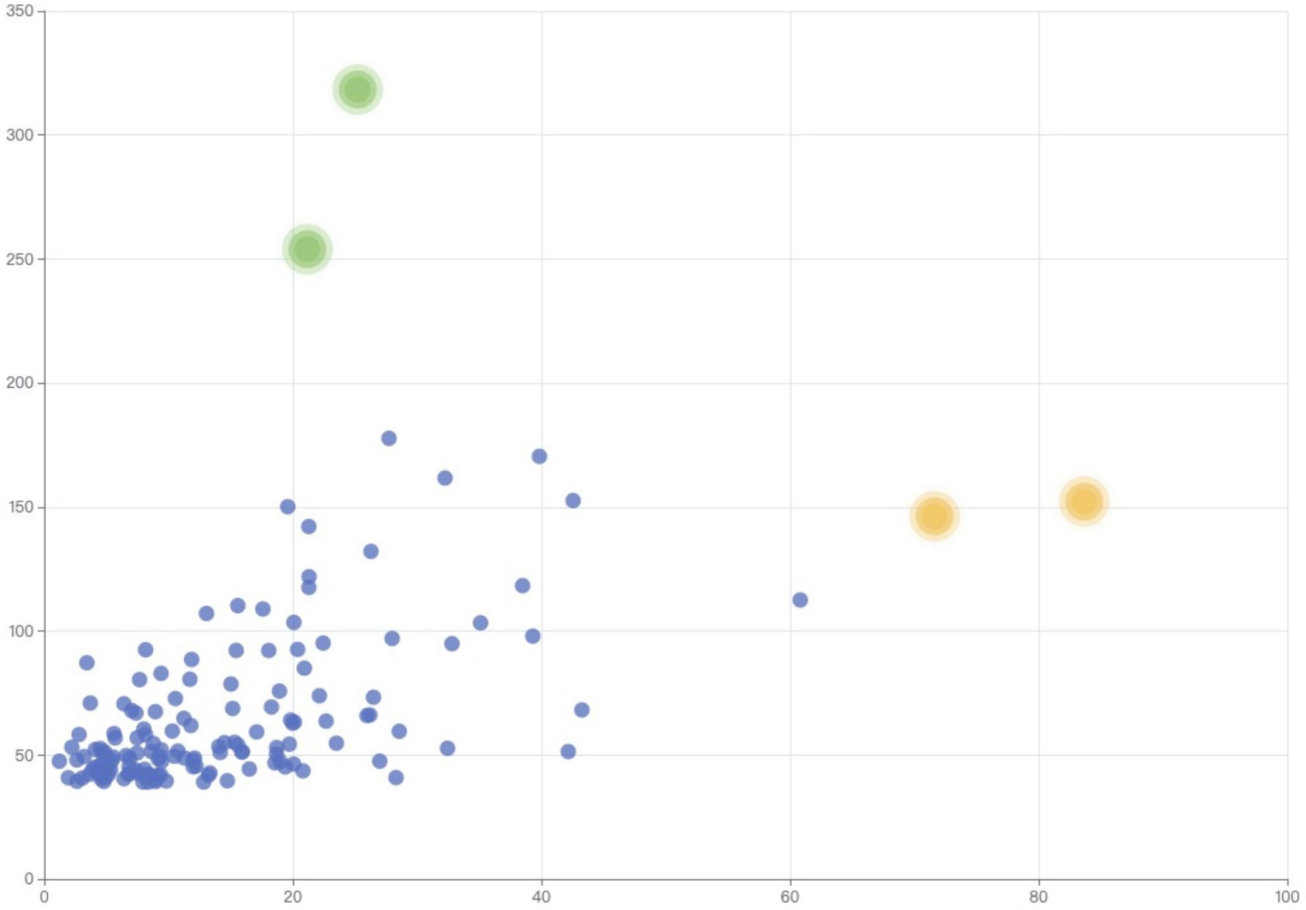
3 数据采集


```
{
    "author_repo": "EbookFoundation/free-programming-books",
    "desc": "\ud83d\udcda Freely available programming books",
    "labels": [
        "education",
        "list",
        "books",
        "resource",
        "hacktoberfest"
    ],
    "stars": 170445,
    "lang": null,
    "lic": "Other",
    "url": "https://github.com/EbookFoundation/free-programming-books",
    "fork": 39842,
    "watch": 9199,
    "issues": 35,
    "pull_req": 8,
    "ctbts": 1581,
    "commits": 6074,
    "date": "2013-10-11T06:50:37Z",
    "size": 10011
},
```

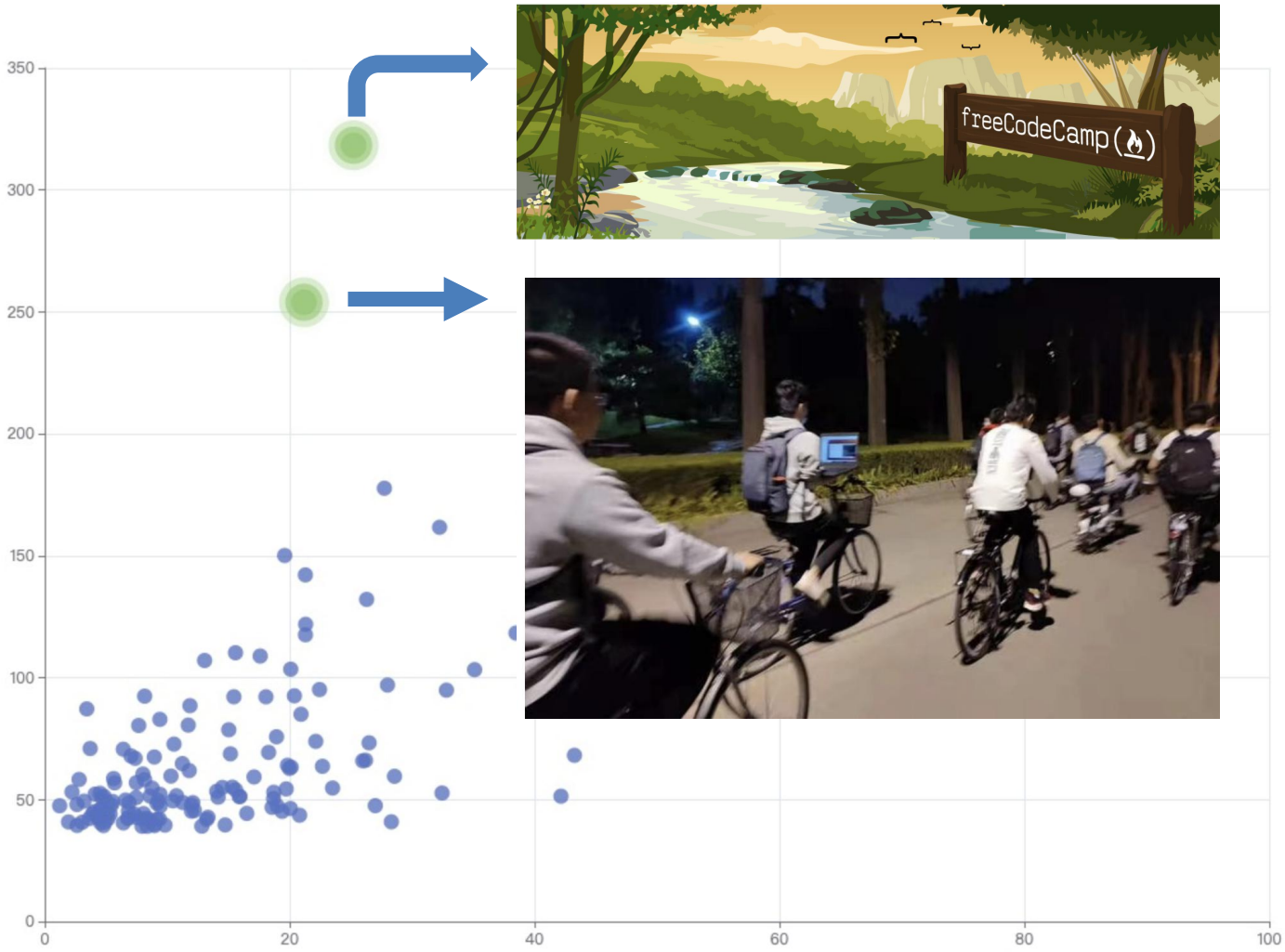
提　纲

4 数据可视化分析 - User Behavior

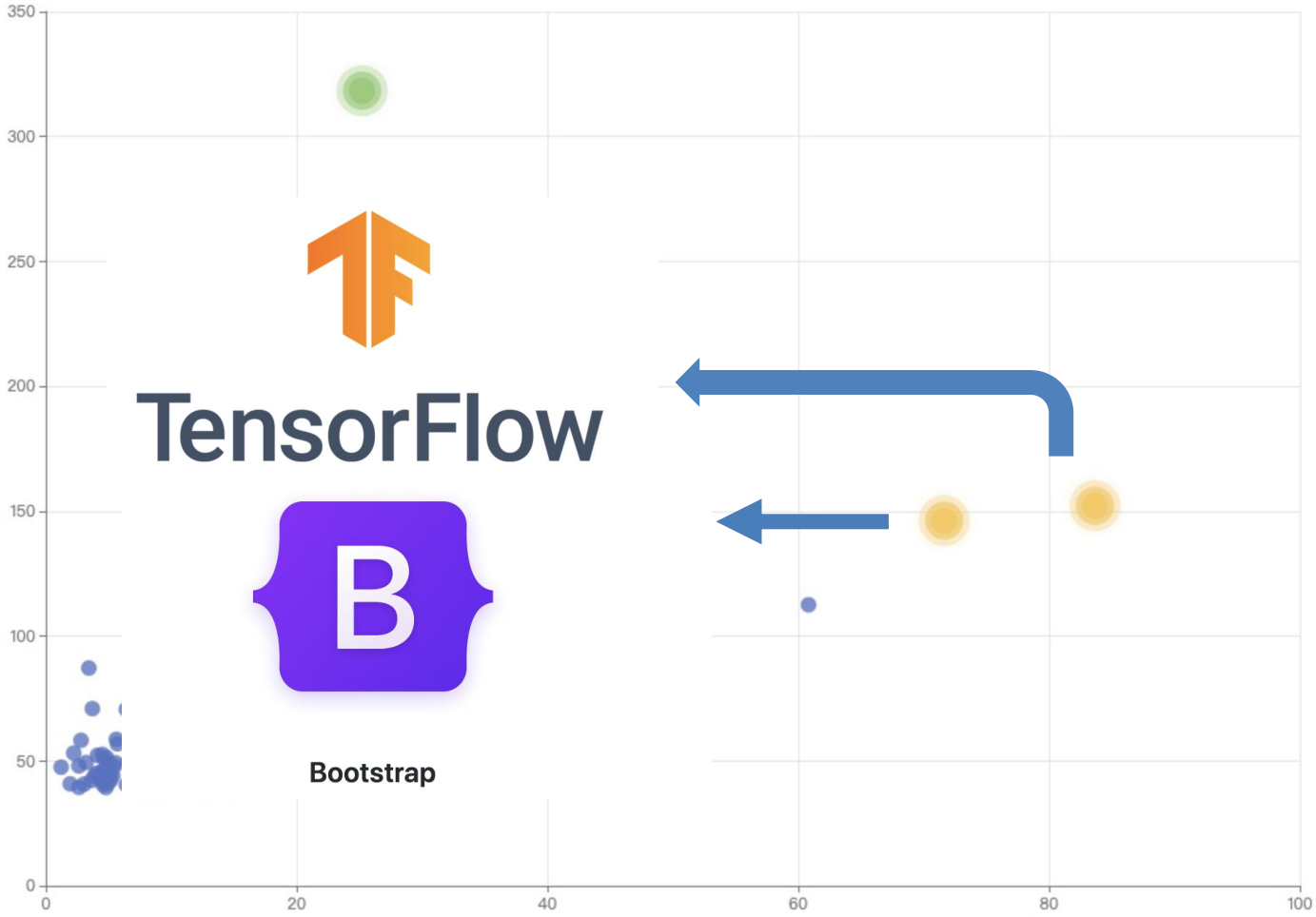
Stars and Forks



4 数据可视化分析 - User Behavior

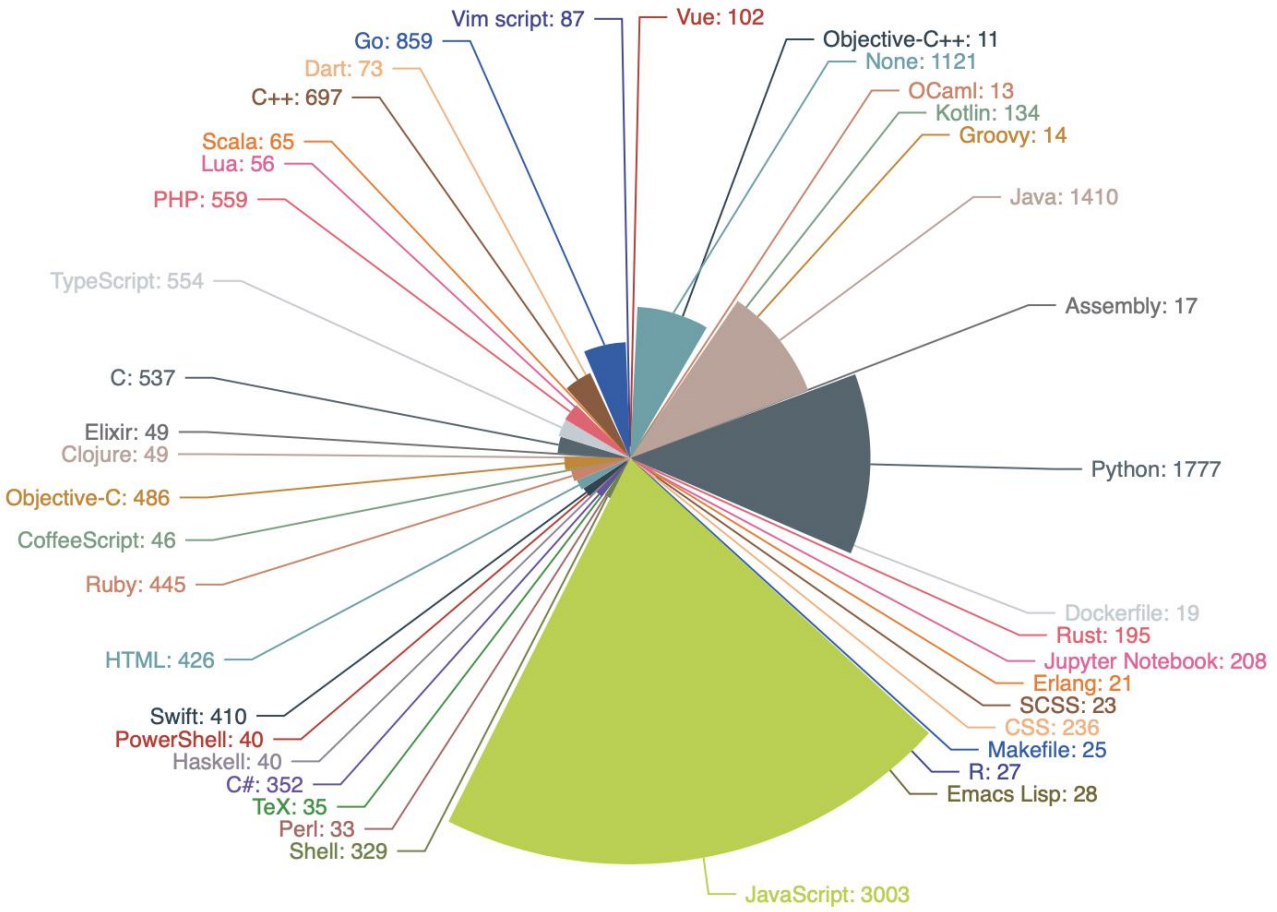


4 数据可视化分析 - User Behavior



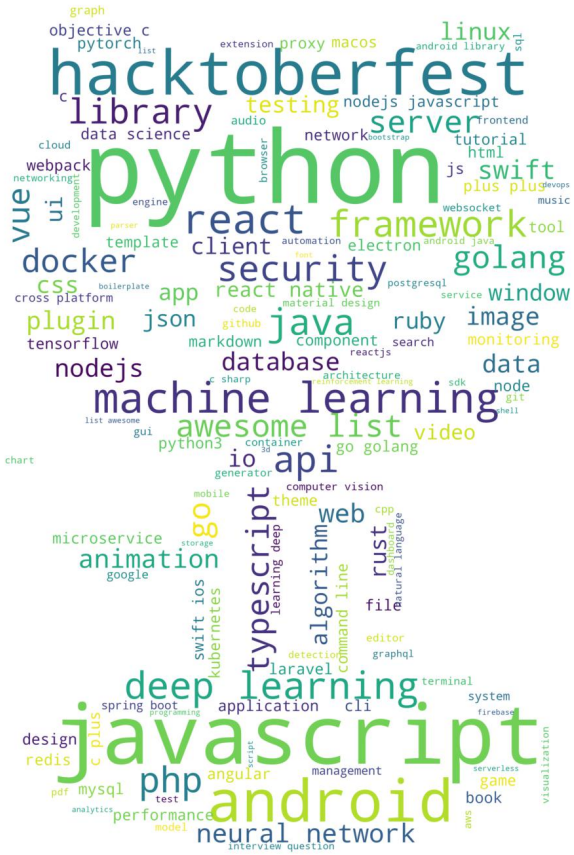
4 数据可视化分析 - Summary

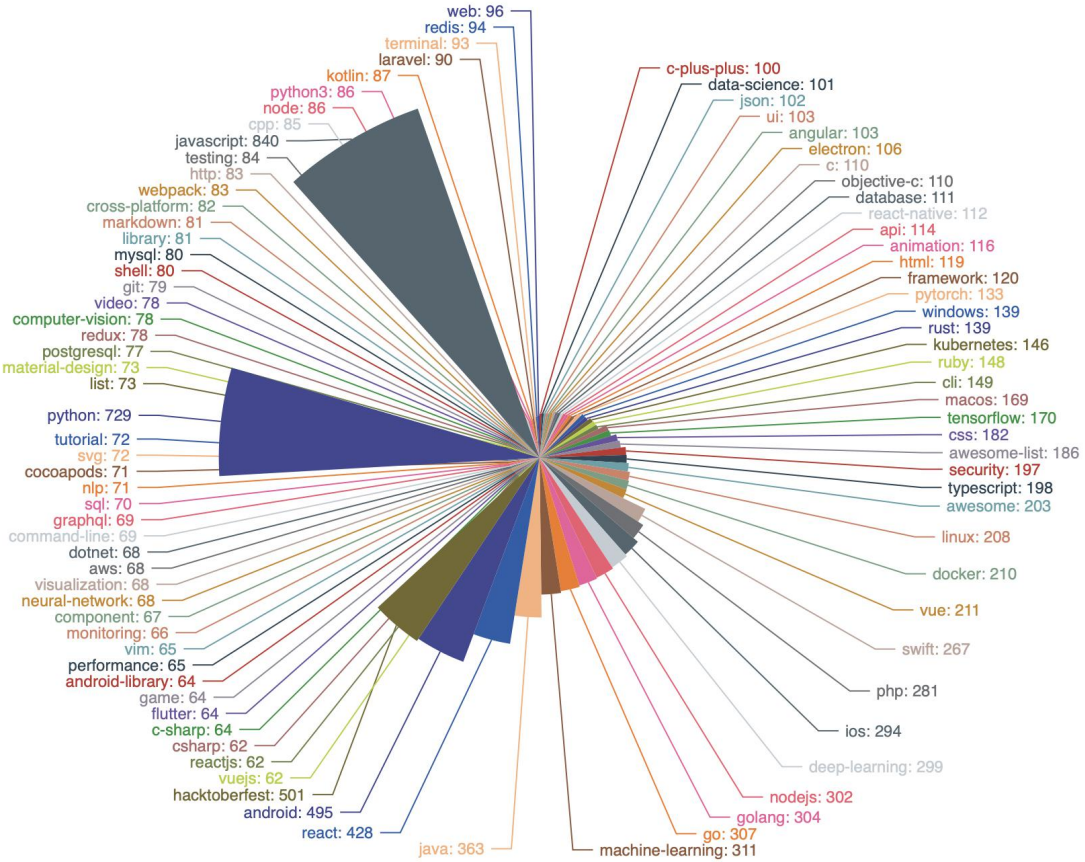
Language Proportion



4 数据可视化分析 - Summary

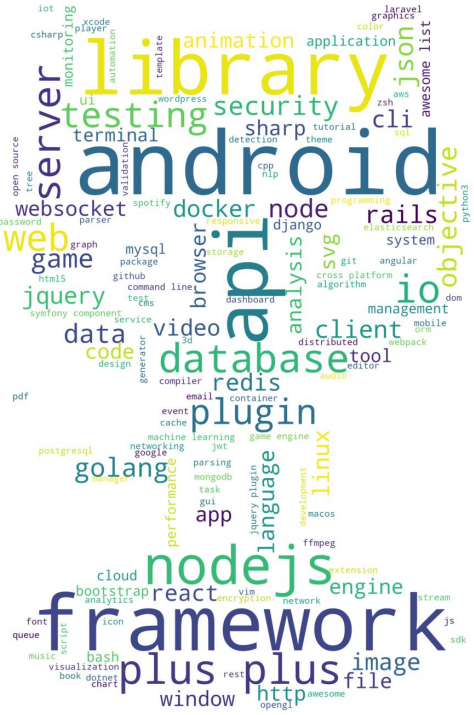
Labels Proportion



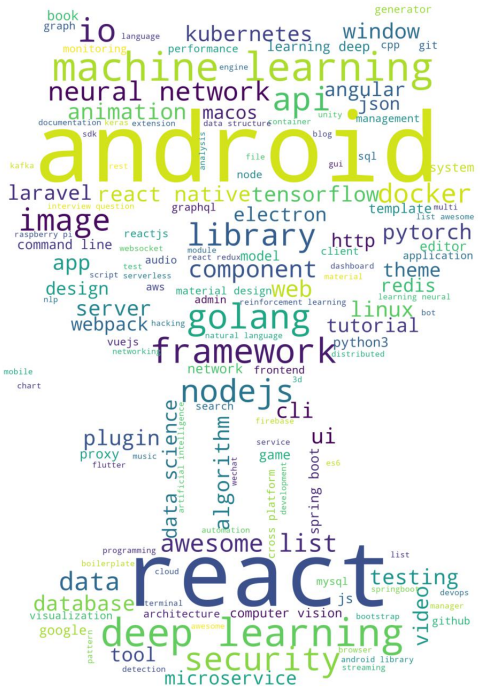


4 数据可视化分析 - Trends

What’s the differences ?



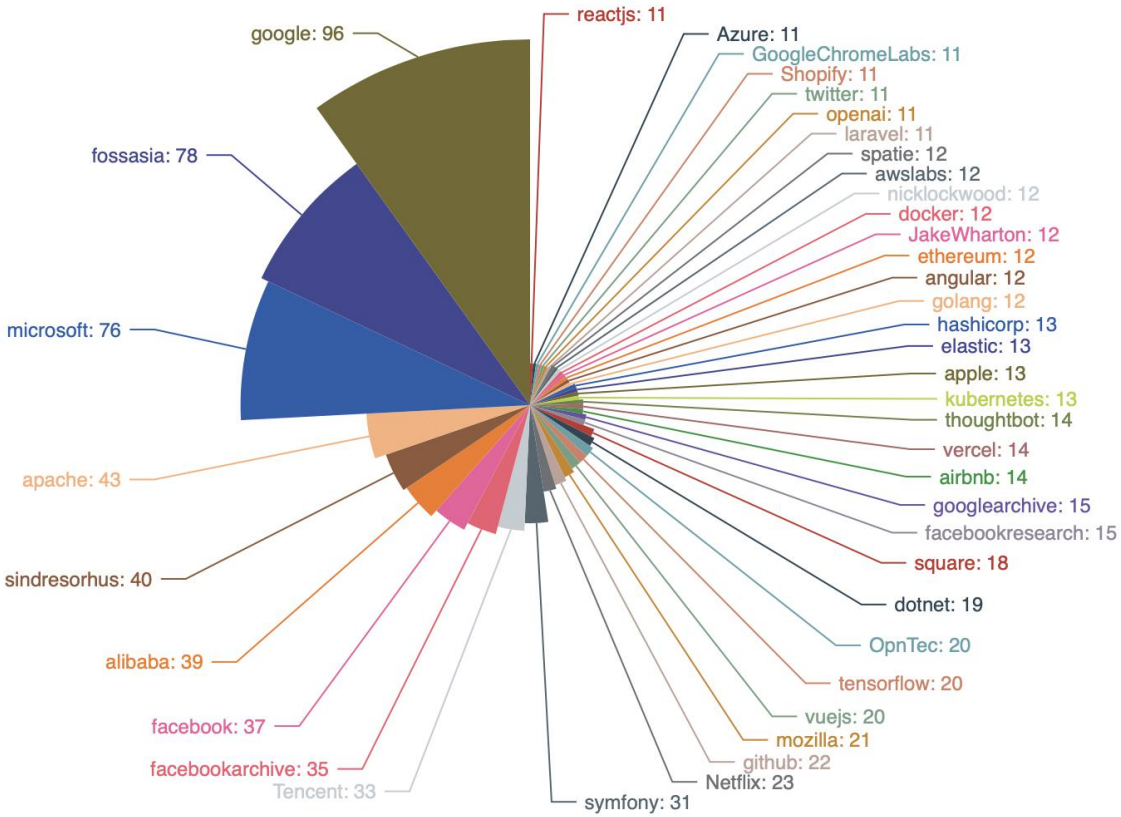
Before 2015

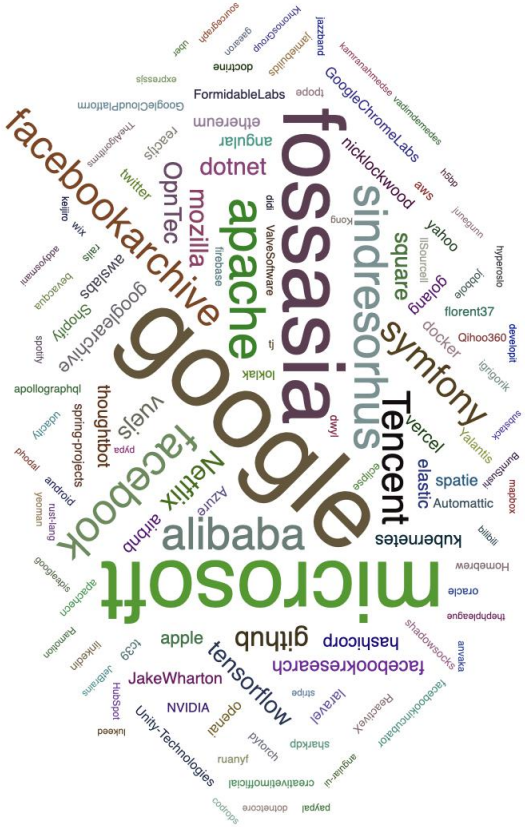


After 2015

4 数据可视化分析 - Trends

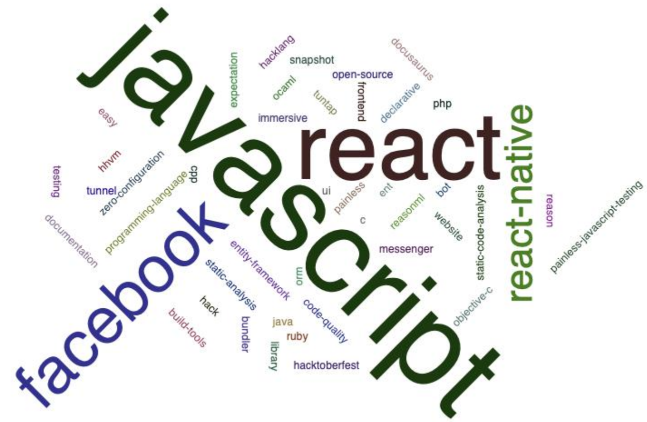
Repo amounts maintained by companies





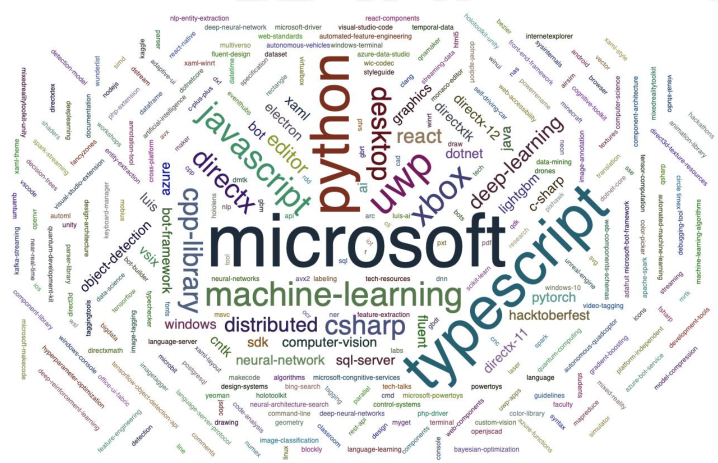
4 数据可视化分析 - Trends

Google



Facebook

Tencent

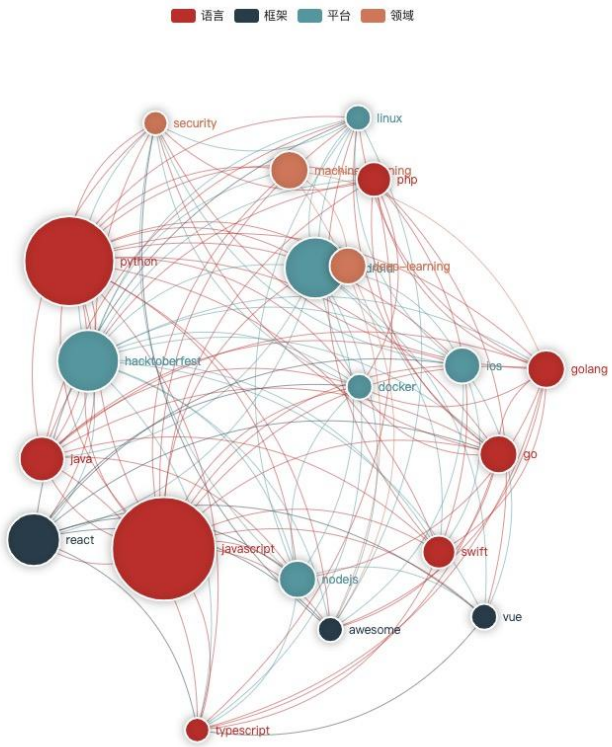


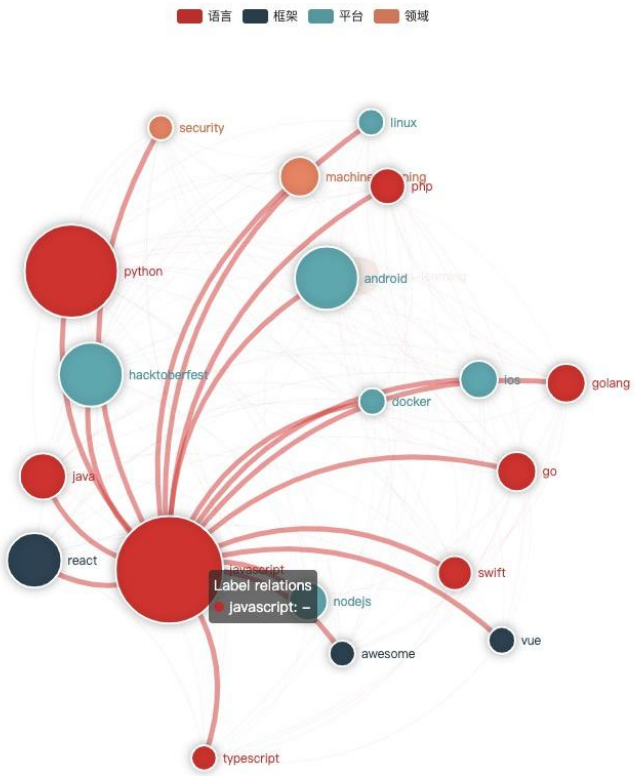
Microsoft

提　纲

5 大数据分析算法 – PageRank

Top 20 labels and their relationships in 20,000 repos

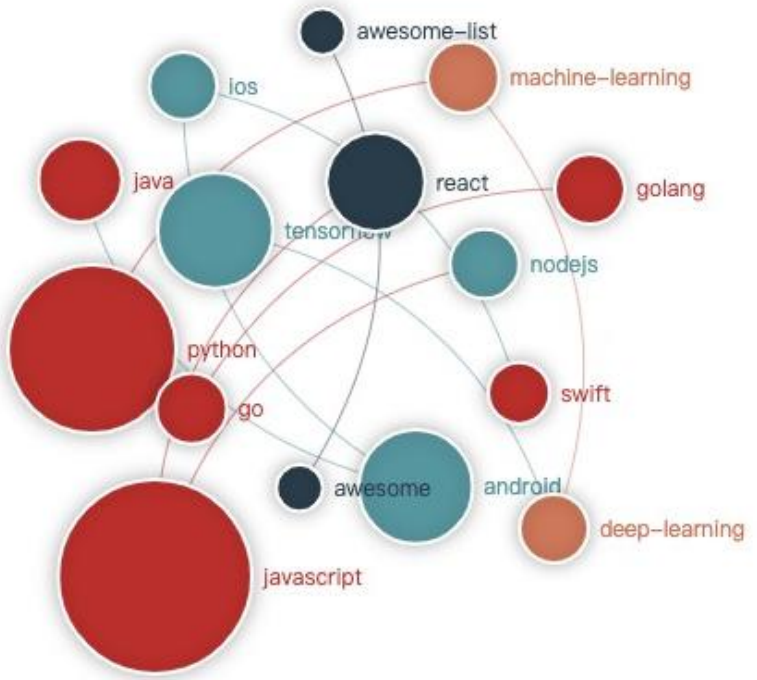




5 大数据分析算法 – Apriori

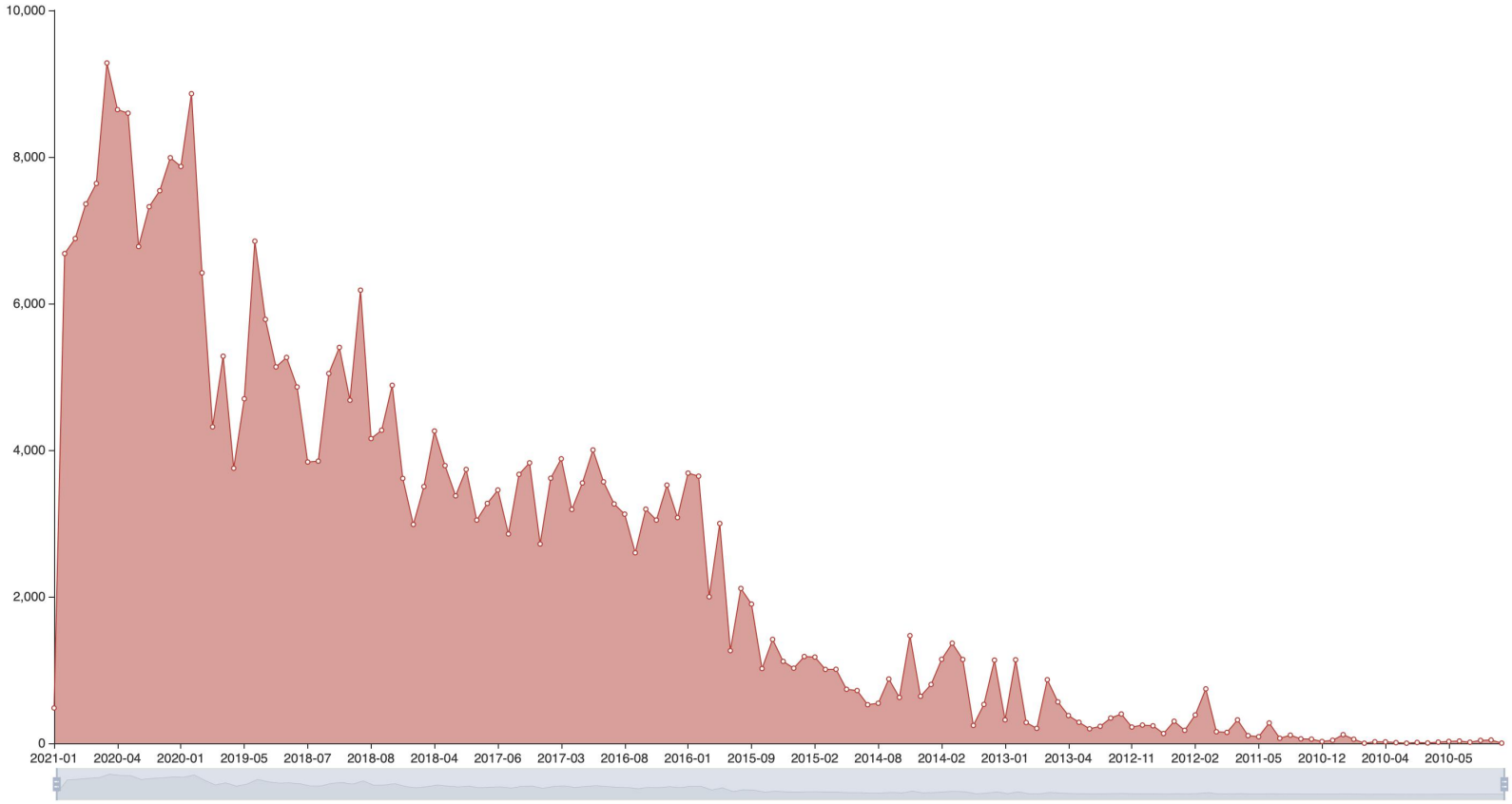
Top 10 association rules in 20,000 repos

[['android', 'ios']
['android', 'java']
['awesome-list', 'awesome']
['machine-learning', 'deep-learning']
['tensorflow', 'deep-learning']
['go', 'golang']
['swift', 'ios']
['javascript', 'nodejs']
['react', 'javascript']
['machine-learning', 'python']]



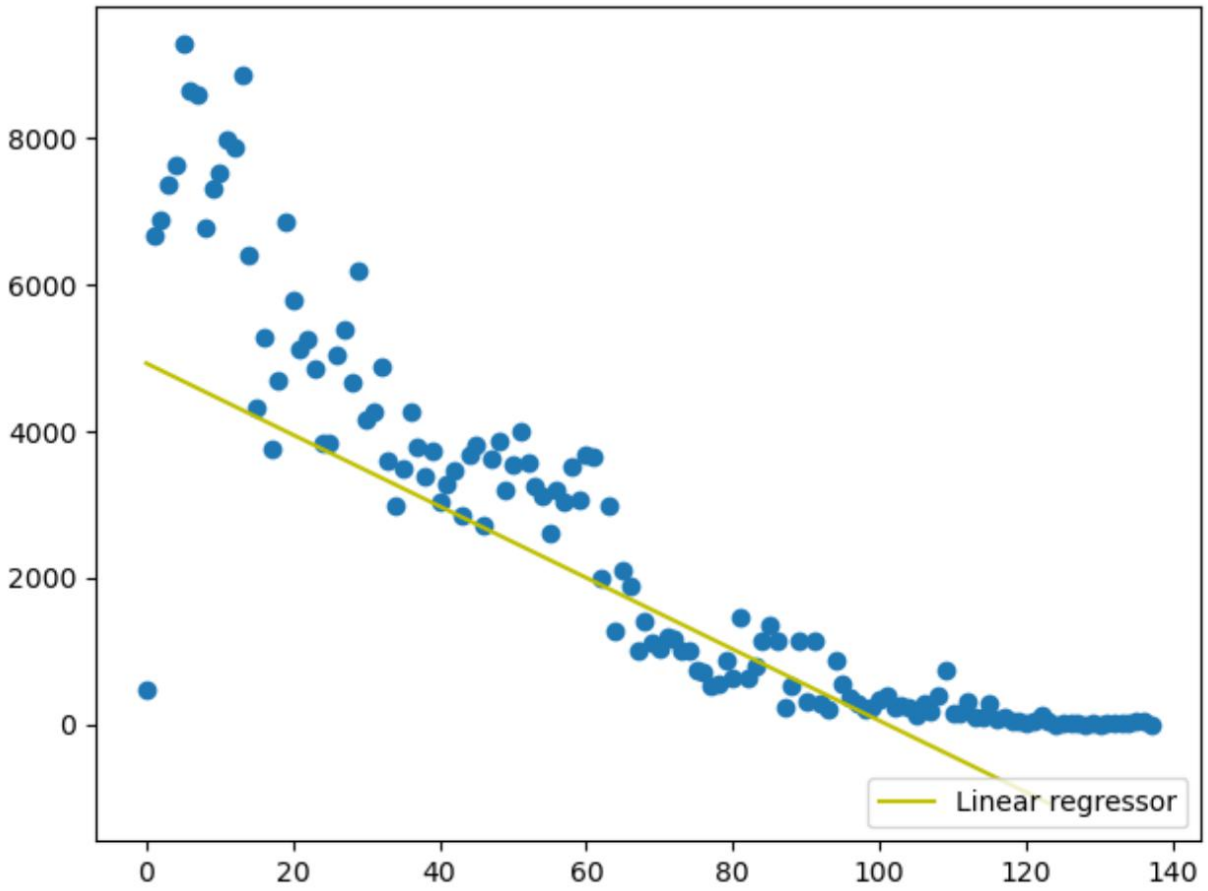
5 大数据分析算法 – Linear Regression

300,000 commit logs from 20 repositories



5 大数据分析算法 – Linear Regression

Linear regression results and the scatter plots



5 大数据分析算法 – Linear Regression

Linear regression for contributors ‘ region distribution

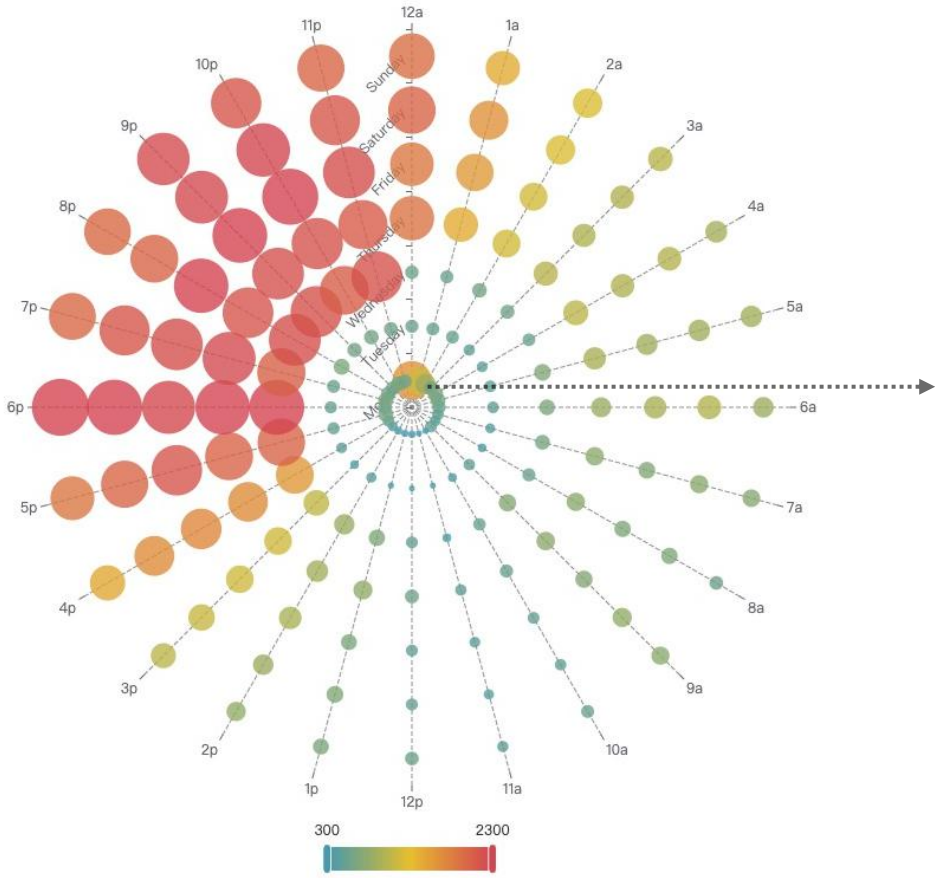
2015

2020

2030

5 大数据分析算法

Contributors' real working time



提　纲

6 结论

- Functional repositories tend to have more forks than stars
- You can learn in advance the community and technology stack that your desired company maintains
- Machine-learning， Deep-learning and Cross-platform is developing rapidly
- Some technologies are closely related and can be used for reference in learning
- Commit logs has an obvious upward trend, and the open source community is booming
- China and India are more and more involved in the open source community
- Most programmers are night owls （DDL： No.1 productivity）