

Github仓库流行度

分析报告



2016-6-10

南京大学软件学院 崔浩

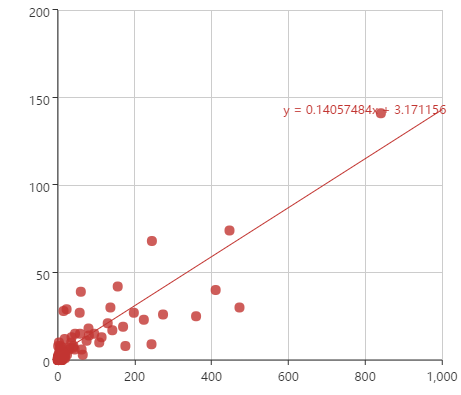
Copyright © 2016.CodeFairy

引言

通过Github提供的API，我们通过随机抽样获取了Github的98341个公共仓库，以及于此相关的330226位用户，1522728条贡献记录，将他们保存在云端服务器数据库中。我们依次为基础分析项目语言、项目领域和项目贡献者对项目star数量的影响，以及影响的显著性差异，并用逻辑分类算法给出了该项目流行概率的估计。

1. 流行度

在仓库可获取的数据中，我们认为仓库的star数量和fork数量最能代表仓库受欢迎的程度，我们通过统计发现，star和fork之间存在一定的关联，为了统计方便，我们便直接选取star的数量作为仓库流行度的衡量标准。



除此之外，我们定义star数量位于前5%的为流行仓库，这样的比例对应到star数量上约为大于等于50（star数量大于等于50的仓库共有5115个，占总数的5.20%）。

根据我们所能获得的数据，我们在这次实验中考虑以下因素对项目的影响：

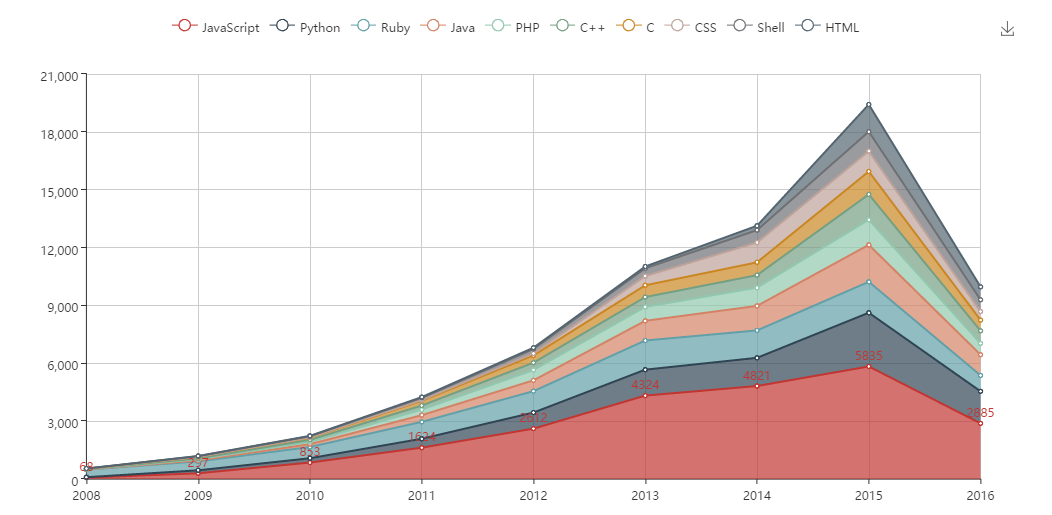
1. 项目语言：项目语言是指Github接口提供的项目的主要语言，在本次实验中没有考虑项目的其他语言分布。
2. 项目领域：根据我们能获取到的数据，我们对项目描述做了词频分析，并选取了约12个出现频率最多的领域，以此考察不同的领域对项目流行度的影响。
3. 项目贡献者：指的是项目的contributor，对项目贡献者的考察，是统计贡献者的follower数量，我们不妨推测用户的follower数量越多，这名用户则越优秀。

下面我们对这三个因素进行一些简单、直观的统计。

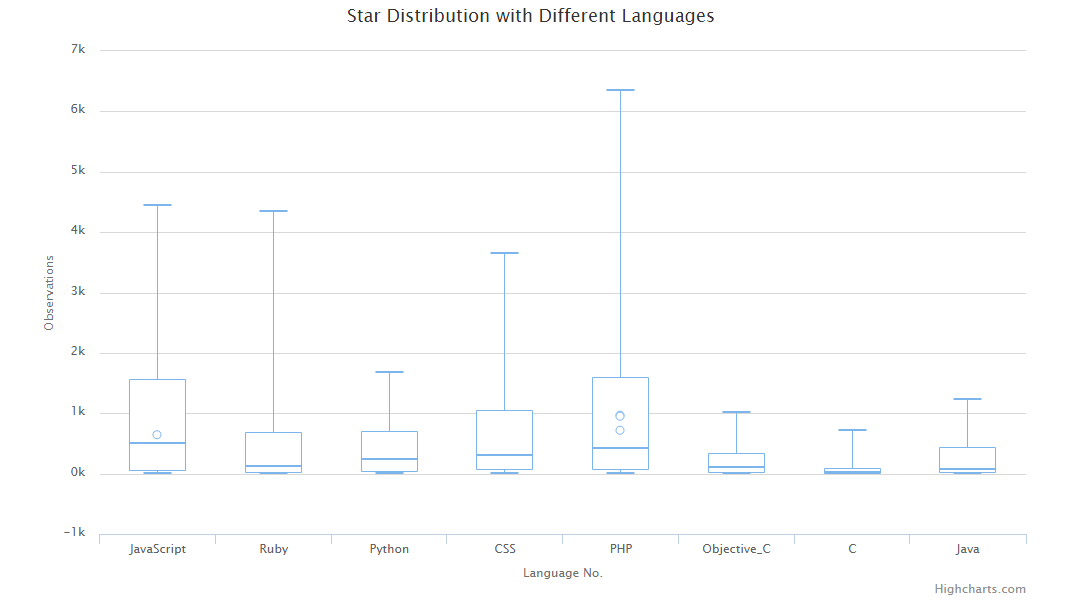
1. 项目语言

在考察流行仓库的语言之前，我们先统计了Github仓库语言的发展情况：我们将仓库按创建时间分类，并统计了每一年不同语言仓库的创建数量。

从下图中可以看到，Github仓库创建数量逐渐增多（2016年截止至5月），JavaScript和HTML随时间所占比例越来越大，而Github传统的Ruby语言所占比例却在不断下滑，这说明了Github开始告别传统的Ruby社区走向多元化，而WEB技术的流行使得Github上的WEB项目逐渐增多。

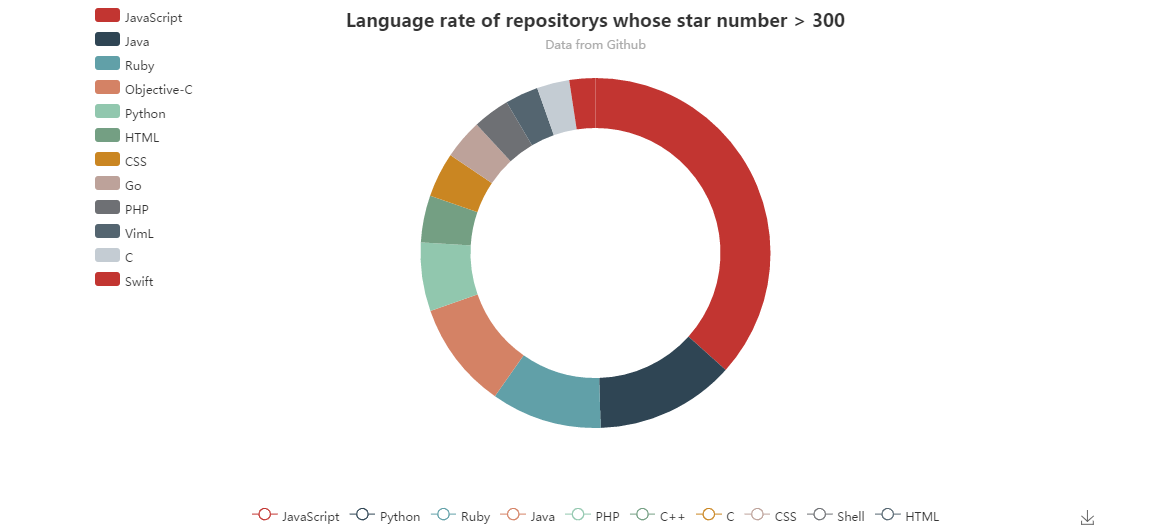


语言对流行度的预测究竟有什么样的影响呢？我们仅仅改变项目语言这一变量，进行了抽样，统计了不同语言的项目的star数情况。

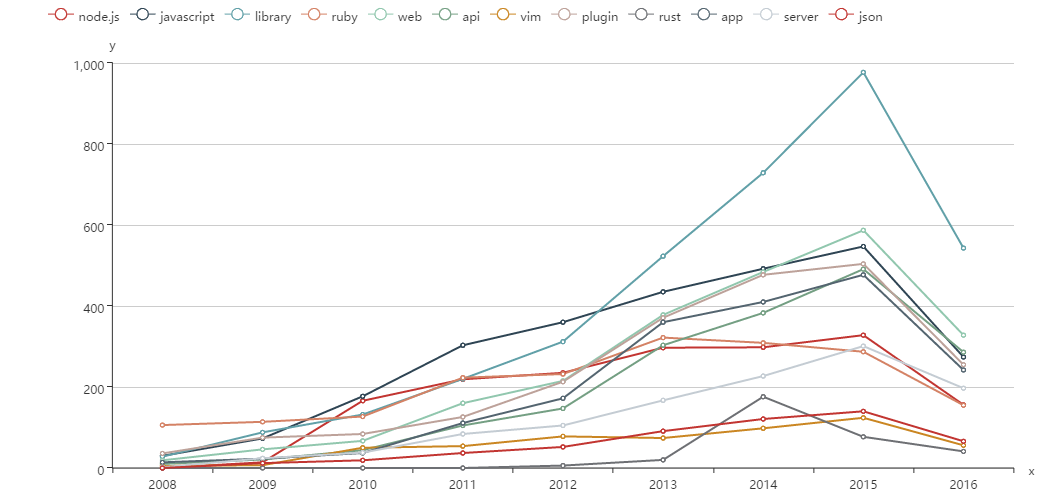


从结果来看，如果一个项目的语言是JavaScript或PHP等语言，我们更愿意相信他们比一个C语言的项目更容易受到关注（尽管每种语言都有不受关注的项目）。

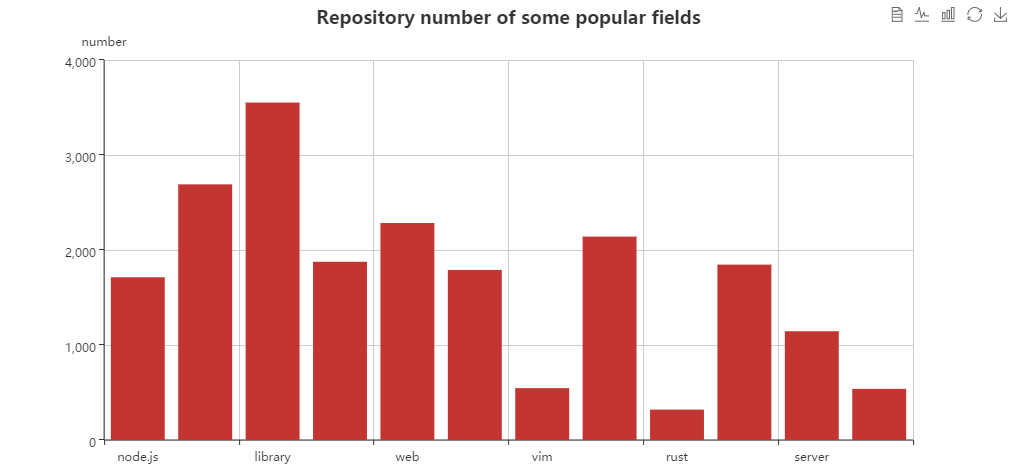
此外，我们统计了流行仓库中各个语言所占的比例：



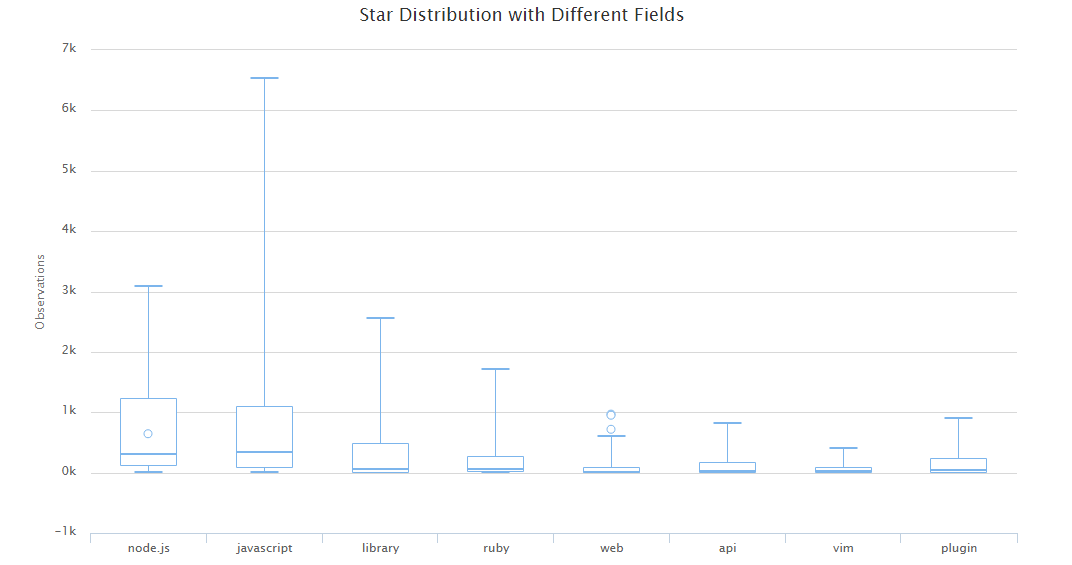
1. 项目领域



可以看到，标有web和library关键词的项目在2012年以后增长迅猛，标注node.js或json关键词的项目增长平稳，vim和json的增长趋势几乎完全一致，我们没有具体探究这样的原因，仅仅描述这样的现象。



在较流行的项目中，library，web，app，plugin等关键词频频出现。



然而，在针对不同领域的随机抽样中，node.js， JavaScript， library， plugin却表现地更加出色。

1. 项目贡献者

敏捷软件开发一直把优秀的团队成员放在更重要的位置，那么，项目贡献者的优秀程度对项目流行的影响究竟有多大呢？

我们将流行项目的贡献者关注人数的分布与普通Github项目贡献者关注人数分布做了对比。