Características de Repositórios Populares: Relatório Final

Laboratório de Experimentação de Software
2021/1
PUC Minas

Guilherme Diniz
Arthur Branco

Introdução

Neste laboratório, foi proposto o estudo de certas características dos sistemas open source mais populares da plataforma GitHub. A partir da análise de seus 1000 repositórios públicos com mais estrelas, serão respondidas às perguntas a seguir:

RQ 01: Sistemas populares são antigos?
RQ 02: Sistemas populares recebem muita contribuição externa?
RQ 03: Sistemas populares lançam releases com frequência?
RQ 04: Sistemas populares são atualizados com frequência?
RQ 05: Sistemas populares são escritos nas linguagens mais populares?
RQ 06: Sistemas populares possuem um alto percentual de issues fechadas?

A partir dessas indagações, foi criado uma série de hipóteses utilizando conhecimentos de desenvolvedores de sistemas para que fosse possível comparar com os resultados adquiridos posteriormente. Essas hipóteses foram:

RQ 1: Considerando que leva certo tempo para que a comunidade ligada a software possa conhecer e reconhecer certo repositório, calcula-se que cada um deles deve ter pelo menos 3 anos e a média de idade deva ser em torno de 5 anos.

RQ 2: Acredita-se que grande parte das novas funcionalidades de cada sistema (e de suas melhorias) vêm de contribuições externas. É difícil fazer previsões quantitativas nesse caso, mas infere-se que, pelo menos 100 Pull Requests aceitos, todo repositório apresenta.

RQ 3: Acredita-se que nem todos os repositórios mais populares lançam releases com frequência pois não é todo projeto que necessita de versionamentos constantes. Dito isso, a média geral deve ser bem alta. Novamente, é muito difícil fazer previsões quantitativas.

RQ 4: Acredita-se que que nem todos os repositórios que necessitam de atualizações constantes. Dito isso, infere-se que a maioria deve receber atualizações diárias ou, no máximo, semanais.

RQ 5: Com grande convicção acredita-se existir uma correlação entre as linguagens mais populares e os repositórios mais populares. Apesar de, possivelmente, existirem repositórios famosos de linguagens já entrando em desuso, calcula-se que a maior parte deve ser escrita em Javascript, Java e Python.

RQ 6: Por existir um enorme número de issues totais nesses repositórios, é extremamente difícil manter um alto percentual de issues fechadas. Mais especificamente, acredita-se que mais da metade dos issues continuam abertos.

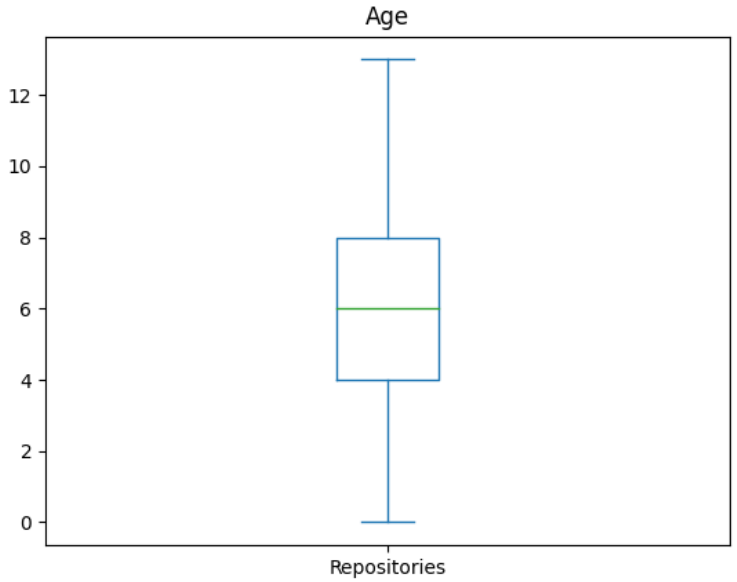
Metodologia

Utilizando a API GraphQL pública do GitHub, foi criado um script em Python onde foi realizada a coleta dos dados através do endpoint disponibilizado. Calculou-se então, a partir dos dados obtidos, a idade dos repositórios utilizando a própria lógica da linguagem e finalmente analisados os dados, foram gerados gráficos relevantes às perguntas realizadas pela pesquisa, esta etapa foi realizada utilizando as bibliotecas pandas e matplotlib.

Resultados

RQ 1:

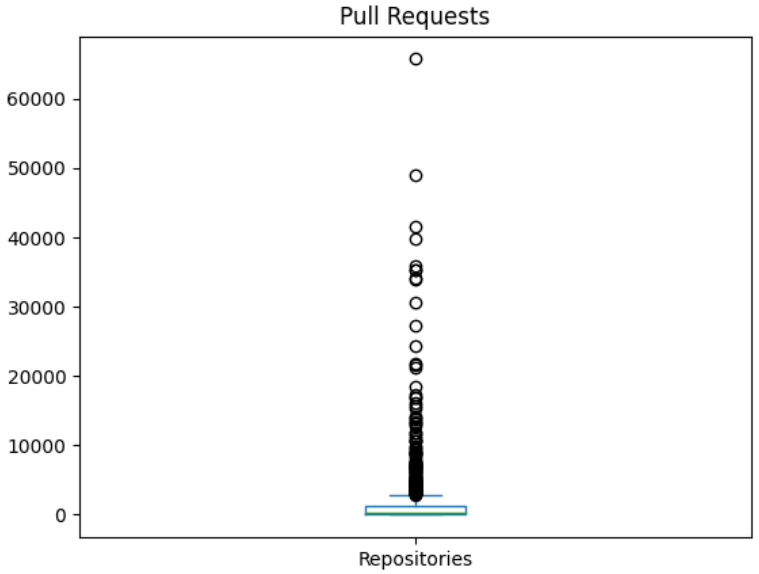
Média de idade: 5.965



Ao observar o gráfico gerado e a partir da média adquirida, os repositórios selecionados podem ser considerados antigos quando comparados com a velocidade da evolução da tecnologia hoje em dia. Dito isso, a média encontrada está levemente acima da previsão proposta de 5 anos. Foi encontrada uma média de aproximadamente 6.

RQ 2:

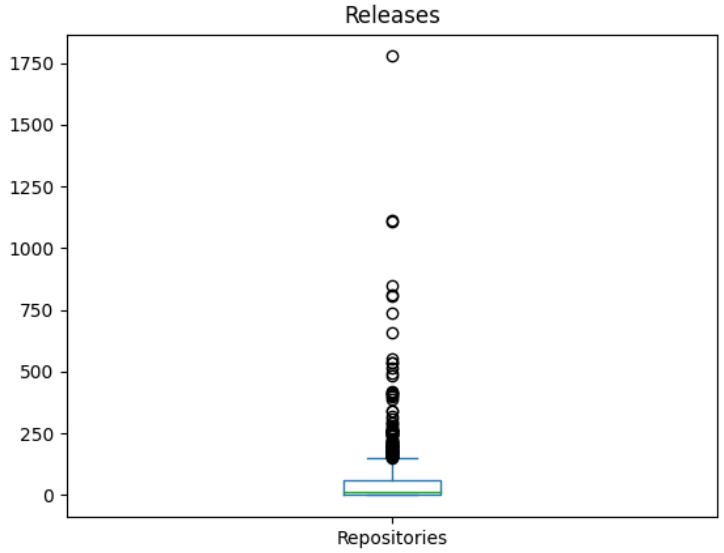
Média de pull requests aceitos: 1818.254



Com uma média próxima de 2000 pull requests aceitos, 20 vezes mais do que previsto, observamos que os repositórios selecionados recebem sim uma grande quantidade de contribuições externas. É interessante notar pela análise do gráfico como existem valores dispersos se comparados aos valores próximos da média. Isso mostra que existem grandes repositórios com um número extremamente alto de pull requests aceitos.

RQ 3:

Média de release: 53.7

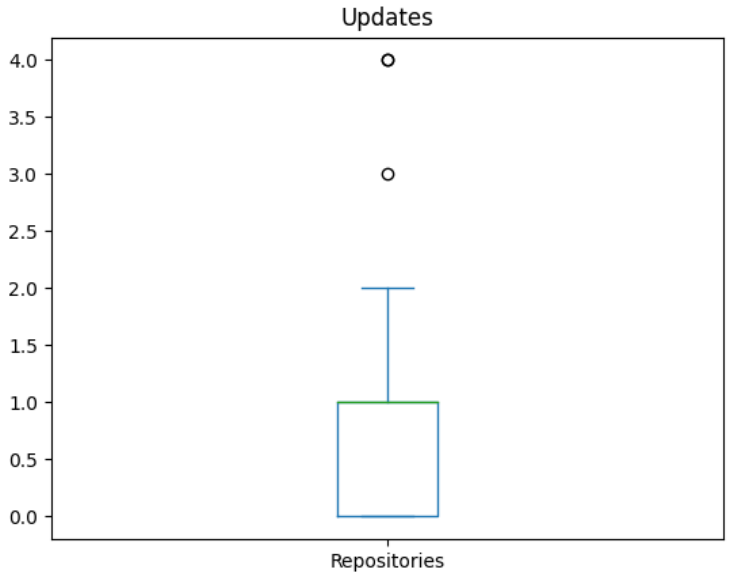


A partir da análise do gráfico e da média obtida, podemos observar que a previsão previamente realizada está de acordo com os dados obtidos, uma vez que 50 releases representa uma quantidade significativa de novas versões sobre os repositórios estudados. Vale ressaltar que nessa análise também foi possível observar a existência de uma grande

disparidade entre os menores e maiores valores encontrados, gerando este gráfico de grande amplitude de resultados.

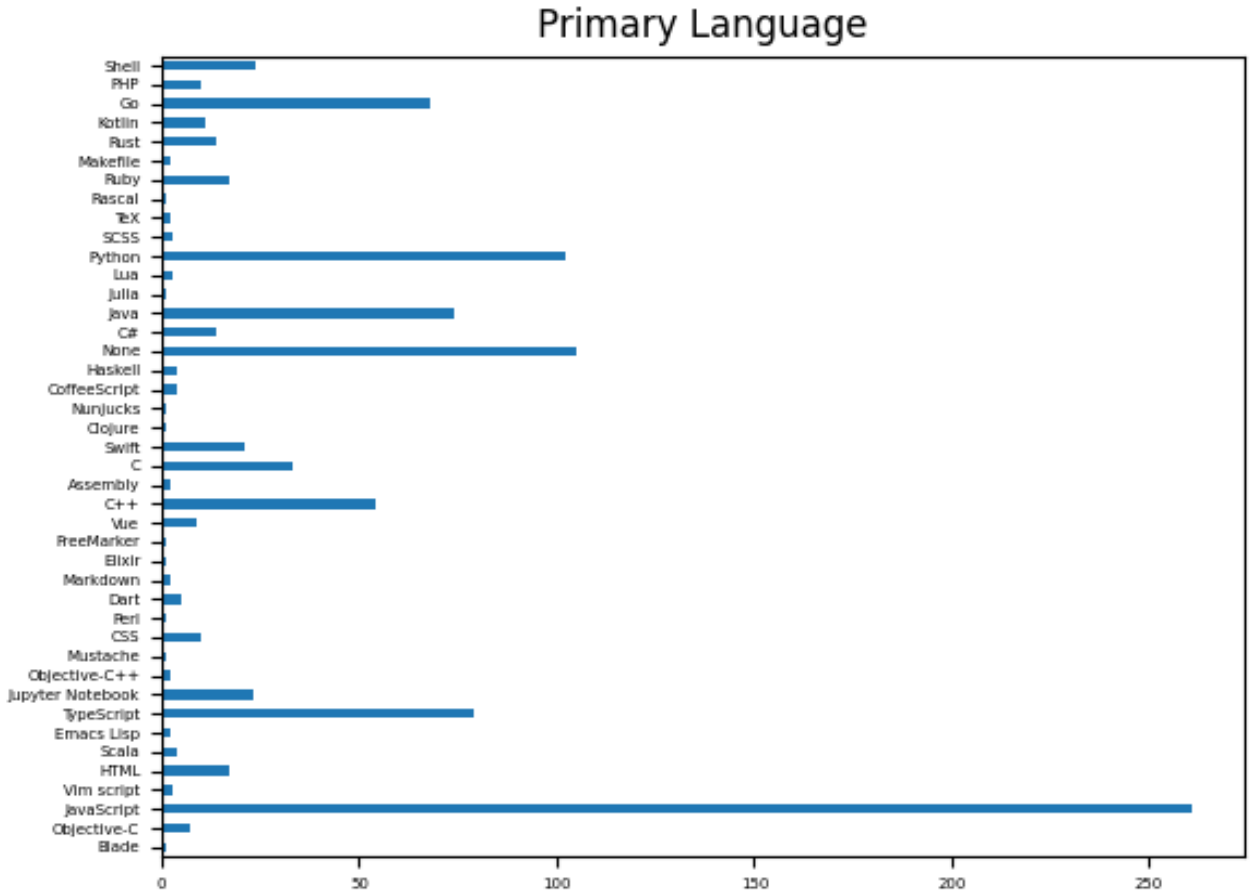
RQ 4:

Número médio de dias desde última atualização: 0.944



Pelos dados apresentados, podemos concluir que existe uma grande quantidade de atualizações nos repositórios mais populares analisados na pesquisa, chegando a apresentar uma média de menos de um dia, algo que está de acordo com a previsão proposta inicialmente.

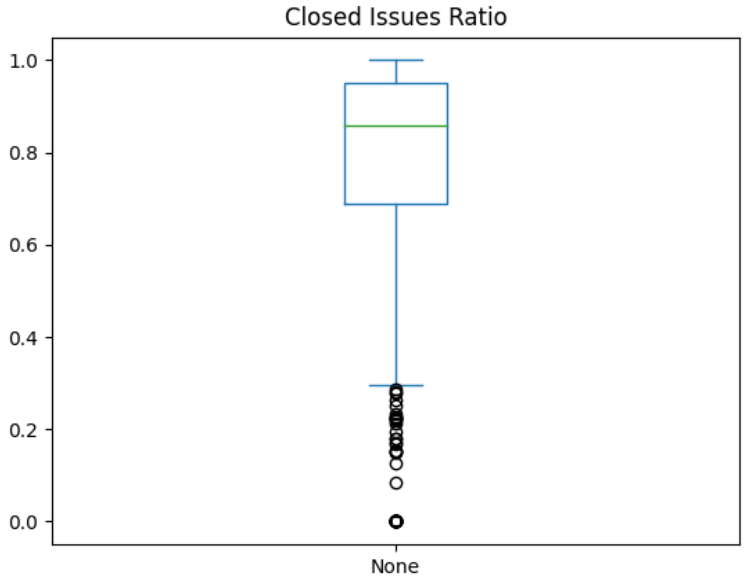
RQ 5:



A partir da análise dos dados obtidos, é possível concluir que as previsões de que as linguagens mais populares dentre os repositórios seriam Javascript, Python e Java estavam corretas. Apesar disso, é interessante notar como a segunda linguagem mais utilizada entre os repositórios na verdade não é uma linguagem. Isso porque muitos repositórios não apresentam linguagens primárias e muitas vezes não apresentam nenhum código sequer. É o caso de repositórios que existem apenas para o compartilhamento de livros de desenvolvimento grátis.

RQ 6:

Média da razão de issues fechadas para issues totais: 0.7750381750616498



Ao observar o gráfico de Issues fechadas, podemos inferir que os repositórios avaliados possuem grande atividade em suas issues, uma vez que a média se aproxima de 80%. Algo que representa grande diferença da quantidade inicialmente prevista pela pesquisa de 50% de issues fechadas.