Análise das Caractéristicas dos Sistemas Populares em Repositórios do Github

Ana Flavia de Souza Ribeiro, Elian Eliezer Fialho de Castro
Miguel Martins Fonseca da Cruz, Rafael Augusto Vieira De Almeida

[1]Instituto de Ciências Exatas e Informática
Pontif˜ıcia Universidade de Minas Gerais (PUC Minas)
Belo Horizonte – MG – Brasil

2

{afsribeiro, eefcastro, miguel.cruz, rafael.almeida.1098763}@sga.pucminas.br

Resumo. *Este projeto tem como escopo a mineração de dados dos 1000 repositórios mais avaliados do Github. A finalidade dessa mineração e´ analisar os dados, de acordo com as RQs propostas.*

1. Introdução

Neste trabalho foi realizada a mineração de repositórios do Github com a fim de responder perguntas pré-definidas. O trabalho foi feito a partir da utilização de Python e GraphQL. Ao se destacar a sprint 1, leva-se em conta que 1000 repositórios foram analisados a fim de se responder à s seguintes perguntas:

- **RQ 01.** Sistemas populares sã o maduros/antigos?
- **RQ 02.** Sistemas populares recebem muita contribuição externa?
- **RQ 03.** Sistemas populares lançam releases com frequência?
- **RQ 04.** Sistemas populares sã o atualizados com frequência?
- **RQ 05.** Sistemas populares sã o escritos nas linguagens mais populares?
- **RQ 06.** Sistemas populares possuem um alto percentual de issues fechadas?

Ao se fazer uma análise inicial das requisições, foi-se capaz de fazer as seguintes hipóteses:

- **RQ 01:** Sistemas populares sã o geralmente mais antigos devido ao tempo necessário para conquistar popularidade na comunidade.
- **RQ 02:** Sistemas populares provavelmente recebem uma quantidade significativa de contribuições externas devido a` sua ampla base de usuários e interesse cont´ınuo.
- **RQ 03:** Sistemas populares tendem a lançar releases com frequência devido à demanda por atualizações frequentes em projetos amplamente utilizados.
- **RQ 04:** Sistemas populares sã o atualizados com frequência devido a` maior visibilidade e escrut´ınio pela comunidade, o que facilita a identificação e correção de problemas.
- **RQ 05:** Sistemas populares podem ser predominantemente escritos em linguagens como JavaScript devido a` sua popularidade e versatilidade em uma variedade de projetos.
- **RQ 06:** Sistemas populares provavelmente tê m uma alta proporção de issues fechadas devido ao amplo envolvimento da comunidade na resolução de problemas.

2. Metodologia

A metodologia aplicada neste estudo tem como o objetivo extrair os dados dos 1000 repositórios mais avaliados do Github. A fim de responder os questionamentos feitos e aplicar suas respectivas métricas, foi realizada a análise de dados diante de todos os resultados obtidos. Houve uma abordagem quantitativa para a pesquisa, fundamentada em literatura acadêmica especializada que reconhece a relevância e eficácia dessa técnica em projetos focados em coleta e análise de dados para o desenvolvimento de pesquisas acadêmicas, como evidenciado no livro "Métodos e Técnicas de Pesquisa Social" por Antônio Carlos Gil.

Como caracter´ısticas da pesquisa quantitativa analisa-se:

Objetividade e neutralidade: Utilização de métodos e técnicas para coleta dados de forma imparcial.

Análise estat´ıstica: Identificar padrões, relações entre variáveis e realizar inferências sobre a população.

2.1. Tecnologias Utilizadas

As tecnologias utilizadas para mineração dos dados foram o Graphql e o Python com as bibliotecas pamdas, requests, csv, matplotlib, datetime e statistics.

2.2. Sprint 1

A primeira fase do projeto focou na mineração de dados iniciais seguindo as RQs e sua respectivas métricas.

2.3. Sprint 2

A segunda fase foi continuada a mineração da Sprint 1, com adição de salvamento dos dados em arquivos .csv e a elaboração inicial do relatório.

2.4. Sprint 3

A terceira fase foi marcada pela elaboração de gráficos através do Graphql, análise de dados e finalização do relatório.

3. Resultados

Realizou-se consultas a` API do GitHub para coletar informações sobre repositórios que atendem a certos critérios de seleção. Cada consulta retornou uma lista de repositórios, e iteramos sobre essas listas para extrair os dados relevantes, como nome do repositório, idioma principal, data de criação, data de última atualização e contagem de problemas fechados. Esses dados foram armazenados em arquivos csv para posterior análise.

- **RQ01.**
  - Média: 7.11
  - Moda: 8
  - Mediana: 7.0

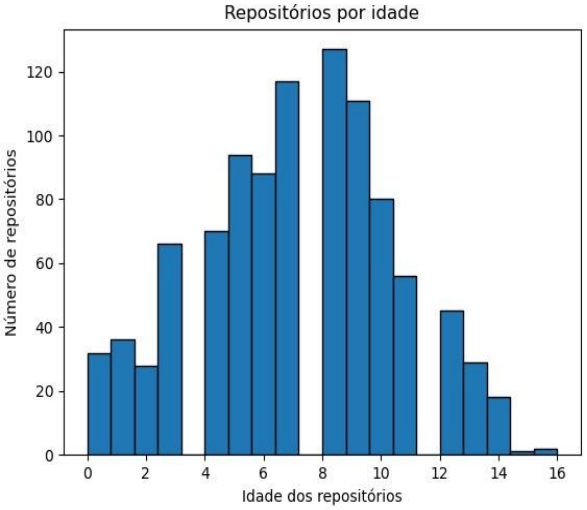

Figure 1. Gráfico RQ01

- **RQ02.**
  - Média: 1908.085
  - Moda: 0
  - Mediana: 323.0

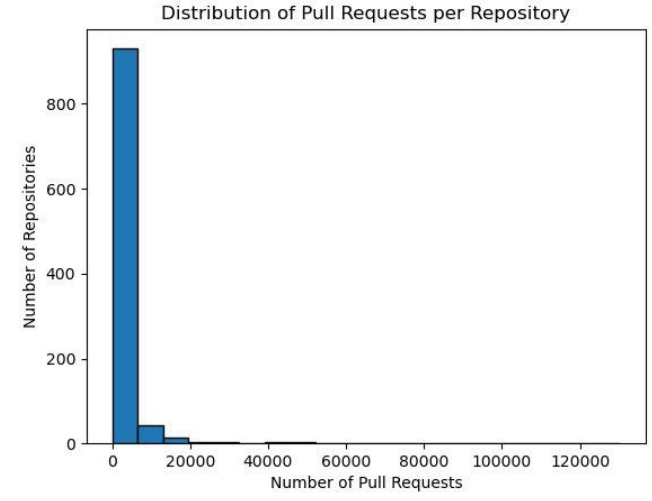

Figure 2. Gráfico RQ02

- **RQ03.**
- **RQ04.**
  - Média: 2024-03-13 10:55:28.449000+00:00
  - Moda: 2024-03-13 13:19:29+00:00
  - Mediana: 2024-03-13 15:17:50+00:00

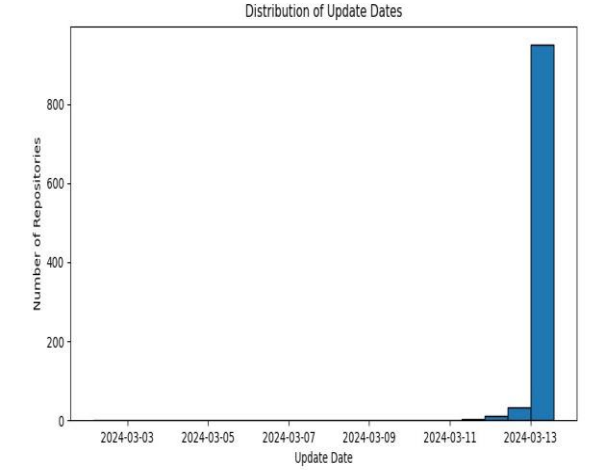

Figure 3. Gráfico RQ04

- **RQ05.**
  - Média: 22.22222222222222
  - Moda: 1
  - Mediana: 4

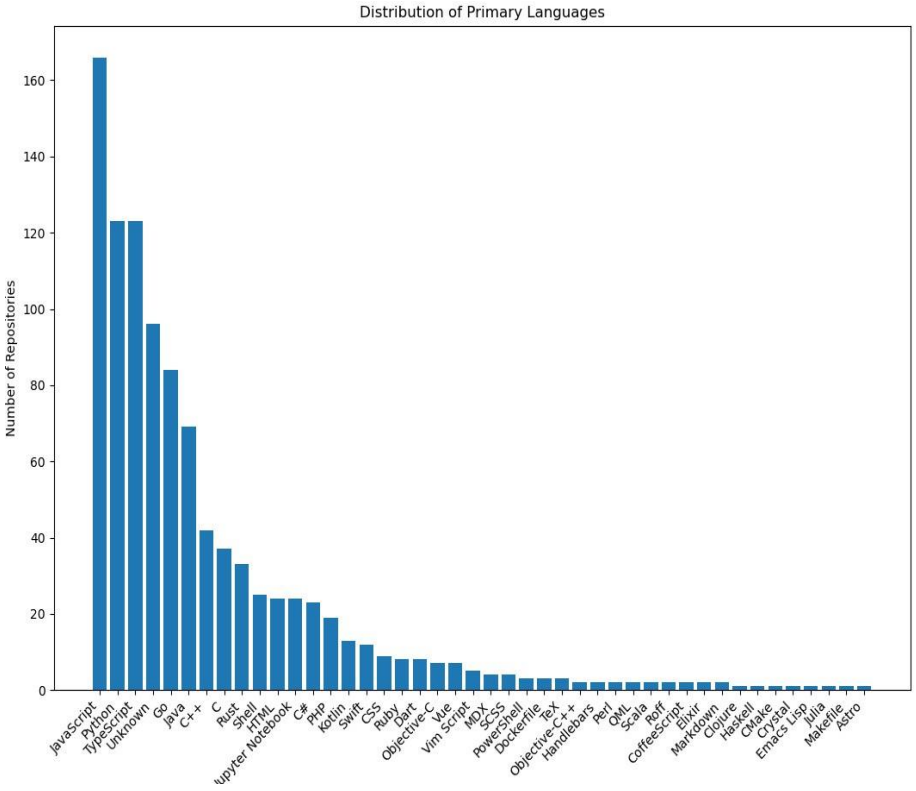

Figure 4. Gráfico RQ05

- **RQ06.**
  - Média: 2793.626
  - Moda: 0
  - Mediana: 748.5

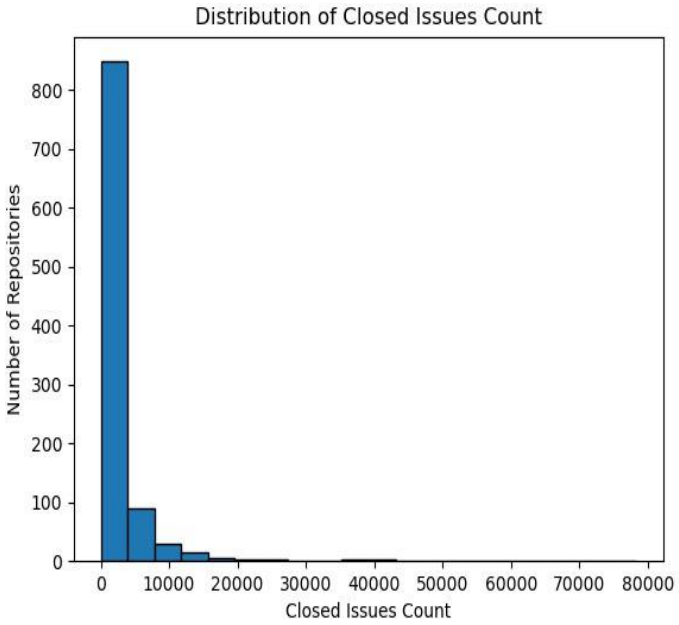

Figure 5. Gráfico RQ06

4. Discussão

Diante dos resultados obtidos com as hipóteses formuladas, podemos determinar se as suposições feitas são verídicas pelos dados analisados.

- **RQ01** A média das respostas é 7.11, indicando que há uma tendência de que sistemas populares sejam mais antigos. Isso pode apoiar a hipótese, sugerindo que a maturidade está correlacionada com a antiguidade.

- **RQ02** A média das contribuições externas e´ 1908.085, o que sugere que, em média, ha´ uma quantidade substancial de contribuições. Isso apoia a hipótese, indicando que a popularidade esta´ relacionada a` participação ativa da comunidade.

- **RQ03** Não há dados fornecidos para esta questão.

- **RQ04** A média das datas das ultimas alterações e´ Média: 2024-03-13 10:55:28.449000+00:00, indicando que os sistemas mais populares são atualizados com mais frequência.

- **RQ05** Diante do resultado obtido, observa-se que a linguagem de programação JavaScript e´ predominante nos 1000 repositórios mais populares do Github, corroborando com a hipótese definida.

- **RQ06** A média e a mediana indicam que ha´ uma alta proporção de issues fechadas. Isso confirma a hipótese, sugerindo que a popularidade esta´ correlacionada com a resolução eficaz de problemas.