Laboratório de Experimentação de Software

Relatório Final: GitHub Issues e Stack Overflow Q&As

16/11/2020 | Bruno Marini - 634883

Introdução

Com mais de 40 milhões de usuários e a criação de mais de 44 milhões novos repositórios em 2019 [1], o GitHub acaba sendo a plataforma de armazenamento remoto de código versionado mais popular dentre seus concorrentes, principalmente quando falamos de projetos *open source*. Se atentando a isso, a plataforma também disponibiliza diversas funcionalidades que fomentam o desenvolvimento de software de maneira colaborativa ao mesmo tempo que organizada. Através de *issues*, *pull requests*, *releases*, *forks* e *wikis*, milhares de usuários podem identificar, contribuir, evoluir e documentar o seu software, contanto que seja público.

Desde 2008 auxiliando neste processo de desenvolvimento através de um fórum de perguntas e respostas, o Stack Overflow fornece uma plataforma rica de informações a serem consultadas pelos desenvolvedores que encontrarem algum obstáculo nos seus projetos, seja este sintático, lógico, arquitetural ou de boa prática de programação. Portanto, não é raro encontrarmos um relacionamento entre as duas plataformas, principalmente de *issues* do GitHub sendo solucionadas no Stack Overflow.

Tido esse contexto, este trabalho - proposto na disciplina de Laboratório de Experimentação de Software do curso de Engenharia de Software da PUC Minas no 2º semestre de 2020 - tem o objetivo de, através de uma análise quantitativa, responder uma série de questões que expõem o impacto do Stack Overflow às *issues* do GitHub, considerando critérios como popularidade das *issues*, do repositório e da linguagem desenvolvida.

Questões de Pesquisa

As questões que fomentam este trabalho são as seguintes.

Pré definidas pelo professor:

1. Com que frequência *issues* do GitHub são discutidas no Stack Overflow?

2. Qual o impacto das discussões de *issues* do GitHub no Stack Overflow?
3. Existe alguma relação entre a popularidade dos repositórios e o *buzz* gerado?

Criadas pelo autor:

4. Existe alguma relação entre a popularidade da linguagem do repositório e o *buzz* gerado?
5. Existe alguma relação entre a popularidade da *issue* com a sua frequência de menções no StackOverflow?
6. Existe alguma relação entre a frequência de discussão no StackOverflow das *issues* de um repositório com a sua taxa de *issues* fechadas?

Hipóteses

Com o objetivo de conjecturar em cima das questões de pesquisa propostas, para cada uma, foi elaborado uma hipótese - e sua respectiva justificativa - a ser validada a partir da análise dos resultados obtidos. Dessa forma, podemos validar cada hipótese a partir da resposta quantitativa de cada cada questão.

Dito isso, as hipóteses são:

1. A *issues* do GitHub não são discutidas frequentemente no Stack Overflow.

   Justificativa: a ferramenta de *issues* do GitHub já disponibiliza um fórum próprio para que os colaboradores de um repositório discutam sobre aquela determinada *issue* atraveś de comentários, além de ser totalmente integrada com outras funcionalidades como links com *pull requests* e *commits*, que disponibilizam o fechamento automático de uma *issue*. Somado à isto, as discussões realizadas no Stack Overflow têm um cunho mais genérico de comportamento inesperado e não específico à uma *issue* do projeto no GitHub. Por este motivo, as *issues* do GitHub acabam sendo mais discutidas na própria plataforma e, por isso, menos frequentemente no Stack Overflow.

2. Não é possível identificar o impacto das discussões de *issues* do GitHub no Stack Overflow.

   Justificativa: por mais que possa ser encontrado alguns repositórios com *issues* sendo discutida no Stack Overflow, o impacto - ou seja, o total de perguntas e respostas relacionadas - será muito pequeno quando comparado com todos os outros repositórios que não realizam o mesmo. Além disso, as discussões do Stack Overflow não necessitam ter uma relação com uma *issue* do GitHub, portanto é muito provável

que exista diversas discussões que tratam sobre o mesmo assunto de uma *issue*, mas que não tenha nenhuma menção à ela. Portanto, não será possível identificar o impacto.

3. Existe uma relação entre a popularidade dos repositórios e o *buzz* gerado.

   Justificativa: devido a popularidade do repositório, é esperado que ele tenha uma maior ocorrência de perguntas e respostas relacionadas no Stack Overflow do que aqueles com menor popularidade devido ao seu renome.

4. Existe uma relação entre a popularidade da linguagem do repositório e o *buzz* gerado.

   Justificativa: semelhante à justificativa da hipótese anterior, é esperado que uma linguagem mais popular tenha uma maior ocorrência de perguntas e respostas relacionadas do que outra menos popular devido ao seu renome.

5. Existe uma relação entre a popularidade da *issue* com a sua frequência de menções no StackOverflow.

   Justificativa: pela popularidade de uma *issue* ser considerada pelo seu nº de comentários, uma *issue* mais popular possui uma maior gama de discussão e difusão entre os desenvolvedores e, com isso, possuir maior probabilidade de ter uma menção no Stack Overflow devido ao seu renome.

6. Não existe nenhuma relação entre a frequência de discussão no StackOverflow das *issues* de um repositório com a sua taxa de *issues* fechadas.

   Justificativa: como já foi mencionado, pelo fórum que o próprio GitHub disponibiliza na sua ferramenta de *issues* através dos comentários, não é necessário que qualquer discussão aconteça no Stack Overflow para que uma *issue* seja fechada. Portanto, existirão muitos repositórios sem nenhuma *issue* mencionada no Stack Overflow que terão uma taxa alta de *issues* fechadas.

Métricas

Visando testar a validade das hipóteses e, com isso também, responder as questões de pesquisa, a análise será baseada nas seguintes métricas:

1. Nº de perguntas relacionadas;
2. Nº de respostas / nº de perguntas relacionadas;
3. Nº de estrelas do repositório vs. total de perguntas relacionadas;
4. Principal linguagem do repositório vs. nº de perguntas relacionadas;
5. Nº de comentários da *issue* vs. nº de perguntas relacionadas;

6. Nº de *issues* / nº de issues fechadas de um repositório e o seu nº de perguntas relacionadas.

Dataset

A análise será feita a partir das 10 issues (5 abertas e 5 fechadas) mais populares (pelo nº de comentários) dos 100 repositórios mais populares (pelo nº de estrelas) que tiverem pelo menos 1 *issue* no GitHub, totalizando então 1.000 issues a serem analisadas.

Metodologia

Foi desenvolvido um *script* em Node.js que, a partir de um *token* da API do GitHub, realiza uma busca paginada da *query* GraphQL a seguir enquanto são coletados algumas métricas necessárias à API do Stack Overflow e, paralelamente, os resultados são salvos nos respectivos arquivos CSV:

- ***repos.csv*:** métricas dos repositórios;
- ***open_issues.csv*:** métricas das issues abertas de cada repositório;
- ***closed_issues.csv*:** métricas das issues fechadas de cada repositório.

```
{
  search(type: REPOSITORY, query: "stars:>100 sort:stars", first: 100) {
    repositoryCount
    pageInfo {
      endCursor
    }
    nodes {
      ... on Repository {
        nameWithOwner
        stargazerCount
        primaryLanguage {
          name
        }
        totalIssues: issues {
          totalCount
        }
        openIssues: issues(states: [OPEN], orderBy: {field: COMMENTS, direction: DESC}, first: 5) {
          totalCount
          nodes {
            id
            number
            title
            createdAt
            updatedAt
            closedAt
            comments {
              totalCount
            }
          }
        }
```

```
        closedIssues: issues(states: [CLOSED], orderBy: {field: COMMENTS, direction: DESC}, first: 5) {
          totalCount
          nodes {
            id
            number
            title
            createdAt
            updatedAt
            closedAt
            comments {
              totalCount
            }
          }
        }
      }
    }
  }
}
```

Ao longo da busca, os repositórios que não possuírem *issues* são descartados, contabilizados e informados ao usuário para que seja possível identificar o déficit faltante até o nº de resultados almejado. Por exemplo, em uma execução realizada no dia 28/10/2020 às 23h00, 7 dos 100 repositórios retornados foram descartados, por isto a quantidade de páginas a serem retornadas foi aumentada de 25 para 30.

Tanto o código quanto a sua documentação podem ser encontrados em: https://github.com/TheMarini/python-vs-java-metrics/tree/v0.3.0.

Métricas do Stack Overflow

Conforme os resultados das páginas são retornados, para cada issue é realizado uma busca para a API do Stack Overflow na rota "/search/advanced" com o termo "<usuário>/<repositório> <número da issue>". Além disso, o filtro "!4(EH5lWNG)R3L7Drl" é utilizado para que só seja retornado as métricas necessárias.

Exemplo de busca realizada:
https://api.stackexchange.com/docs/advanced-search#page=1&order=desc&sort=votes&q=vue%2Fvuejs%202873&filter=!4(EH5lWNG)R3L7Drl&site=stackoverflow&run=true

Problemas encontrados e modificação do *dataset*

Ao executar a busca, alguns problemas foram identificados com a API do Stack Overflow:

- **Limite de requisições**
  A API possui um limite de 300 requisições diárias à clientes anônimos. Isto quer dizer que, a não ser que a aplicação seja cadastrada no Stack Apps (serviço do próprio Stack Exchange), não é possível ultrapassar este limite, caso contrário, há um bloqueio do IP por 20h. É possível obter um limite de 10.000 requisições diárias através desse

cadastro, porém é necessário algumas informações inviáveis à execuções locais (semelhante ao *script* deste trabalho), como *OAuth* e *website*.

- **Bloqueio de IP**
  A 1ª execução da busca resultou no bloqueio do IP por 20h, gerando um tempo ocioso que poderia atrasar a próxima entrega deste trabalho se ocorresse novamente.

Consequente a esse limite e pela necessidade da realização de 1 requisição por *issue*, não foi possível realizar o *dataset* almejado com total 1.000 *issues* analisadas por ultrapassar mais de 3 vezes o limite diário. Por isto, o *dataset* foi alterado para o máximo possível, com um intervalo de segurança de 80 requisições, para não ocorrer o bloqueio de IP novamente. Portanto, o *dataset* foi modificado para ser adequado à 210 requisições:

*A análise será feita a partir das 10 issues (5 abertas e 5 fechadas) mais populares (pelo nº de comentários) dos **21 repositórios** mais populares (pelo nº de estrelas) que tiverem pelo menos 1 issue no GitHub, totalizando então **210 issues** a serem analisadas.*

Com isto definido, **a busca foi realizada às 21h25 do dia 16/11/2020**.

Resultados Obtidos

Como proposto pelo próprio enunciado deste trabalho e para ser justo nos casos que há um repositório com o valor de uma métrica muito superior do que os demais na mesma categoria - como pelo nº de estrelas, nº de comentários, nº de perguntas relacionadas e etc. -, tantos os gráficos quanto as análises foram baseadas na **mediana** [2] dos resultados obtidos para questão de pesquisa e hipótese.

Dito isso, segue os resultados obtidos.

Questão 1: "Com que frequência *issues* do GitHub são discutidas no Stack Overflow?

Como é possível observar na tabela 1.1, existem repositórios com perguntas e respostas no Stack Overflow relacionadas às suas *issues*. Porém, como demonstrado tanto pela mediana quanto pelo desvio padrão, este paradigma só é encontrado em poucos repositórios, e não representa a maioria analisada.

Tabela 1.1

Frequência de Discussão no Stack Overflow		Métricas		
		Nº de perguntas no S.O.	Nº de respostas no S.O.	Nº de visualizações
Medidas	Total	7.064	6.879	10.223.072
	Máximo	2.926	2.982	4.049.478
	Mediana	0	0	0
	Mínimo	0	0	0
	Média	34	33	48.914
	Desvio padrão	258	254	363.268

Inclusive, ao filtrar os resultados pelos que contêm pelo menos 1 como valor de qualq uer uma das métricas acima, somente 8 dos 21 repositórios analisados (38%) foram retornados. Não obstante a isto, os valores se concentram em apenas 2 destes repositórios, como demonstrado no gráfico 1.1 e 1.2.

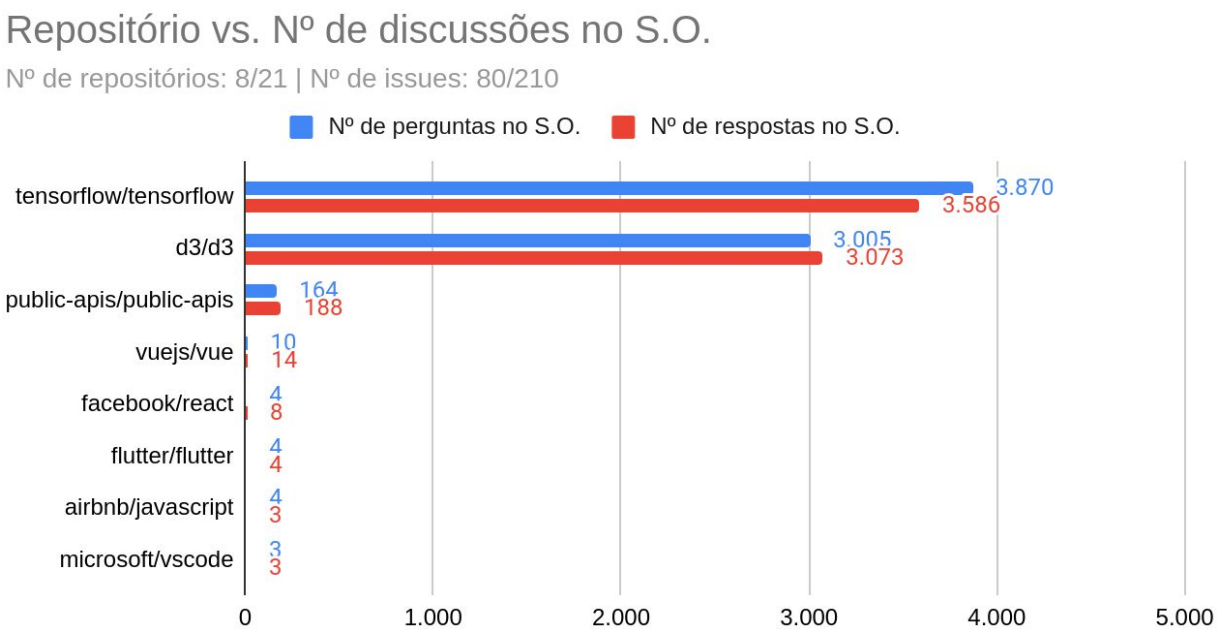

Gráfico 1.1

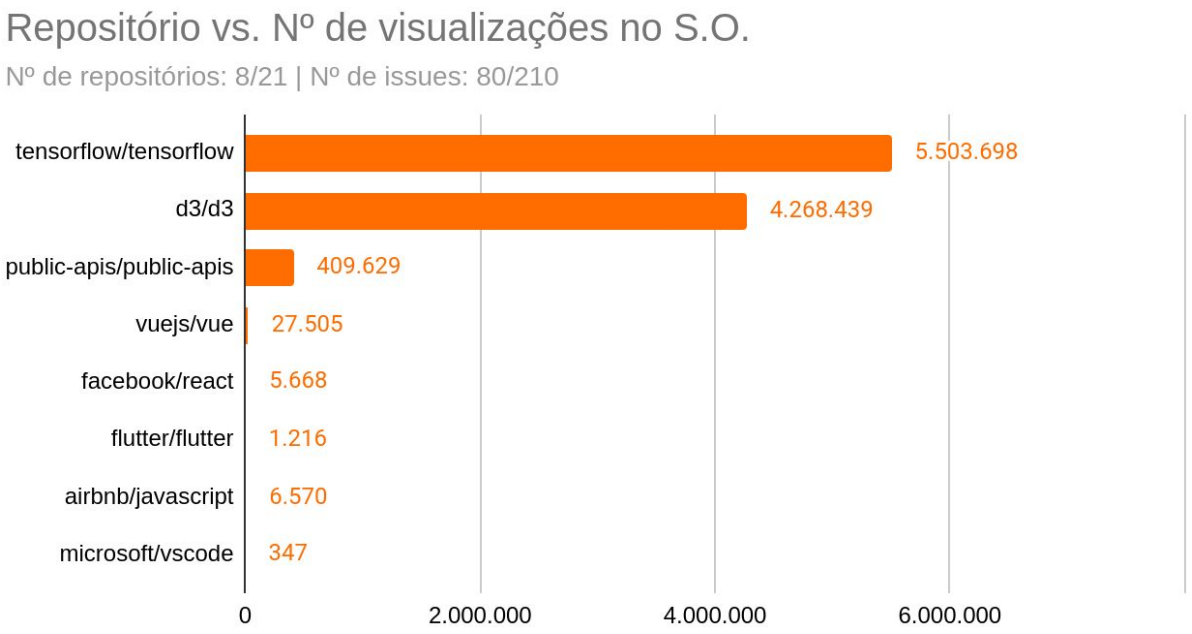

Gráfico 1.2.

Isto demonstra que, por mais que exista uma alta frequência em alguns dos repositórios, isto só está presente em uma parcela muito pequena (9,5%) dos repositórios analisados.

Hipótese 1: A *issues* do GitHub não são discutidas frequentemente no Stack Overflow."

Verdadeira. Como explicado no tópico anterior, por mais que exista uma frequência em uma ínfima parcela dos repositórios analisados, esta não é uma realidade para os outros 81,5% dos repositórios. Ou seja, não há nenhuma generalização para o atributo de frequência da discussão das *issues* no Stack Overflow, portanto não são discutidos frequentementes.

Questão 2: "Qual o impacto das discussões de *issues* do GitHub no Stack Overflow?"

Muito baixo. Considerando que 210 *issues* foram analisadas, mas somente 80 delas (38%) surtiram algum impacto no Stack Overflow e grande parte das discussões se encontrarem em apenas 20 delas (9,5%), como demonstrado anteriormente no gráfico 1.1, o impacto das discussões é baixo. Porém, não obstante a isto, as perguntas e respostas impactadas às *issues* possuem um número de visualizações considerável, como demonstrado no gráfico 1.2.

Hipótese 2: "Não é possível identificar o impacto das discussões de *issues* do GitHub no Stack Overflow."

Se considerando "impacto" como o número de perguntas e respostas com relacionamento às *issues*: falso, pois mesmo que tenha sido identificado um impacto muito baixo, ainda foi possível identificar algum.

Se considerando "impacto" como o número de visualizações à perguntas e respostas com relacionamento às *issues*: falso, pois foi possível identificar um valor total considerável de 10.223.072 visualizações, como demonstrado no gráfico 1.2.

Questão 3: "Existe alguma relação entre a popularidade dos repositórios e o *buzz* gerado?"

Não. Como demonstrado no gráfico 3.1, a seguir, não existe nenhuma correlação entre o número de estrelas do repositório no GitHub e o número de perguntas no Stack Overflow relacionadas às suas *issues*. Esta falta de relação se torna nítida quando observamos repositórios com muitos mais estrelas, ou até mesmo com quantidade próxima, mas sem qualquer pergunta relacionada à qualquer uma das suas *issues* mais populares.

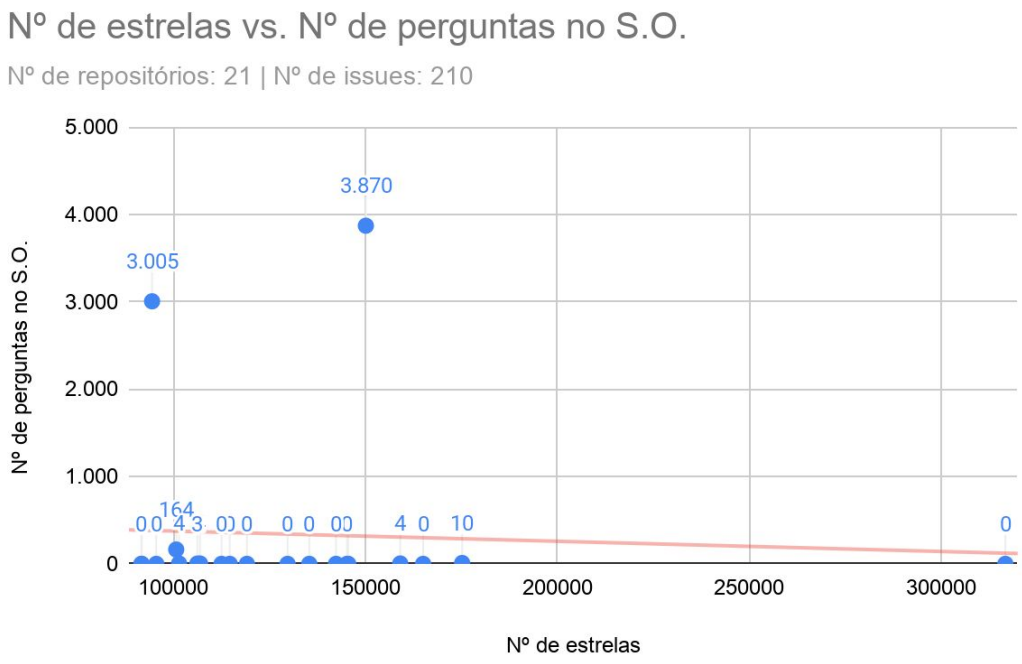

Gráfico 3.1

Hipótese 3: "Existe uma relação entre a popularidade dos repositórios e o buzz gerado."

Falsa. Por mais que seja esperado essa relação devido a popularidade do repositório, não existe nenhum indício dela existir, conforme demonstrado no gráfico 3.1.

Questão 4: "Existe alguma relação entre a popularidade da linguagem do repositório e o *buzz* gerado?"

Não. Baseado no ranking de linguagens populares, realizado pelo Stack Overflow em 2020 [3], foi realizado um *crossing* das métricas de cada linguagem ordenado pela sua popularidade (de acordo com a % de votos da pesquisa do Stack Overflow), encontrado na tabela 4.1. Então foi criado um gráfico de dispersão, encontrado no gráfico 4.1.

A partir dos dois, disponíveis a seguir, nenhuma correlação foi encontrada entre a popularidade da linguagem do repositório e o *buzz* gerado. Isto se torna claro quando comparamos as métricas de JavaScript e Python - as linguagens mais populares presentes entre os 17 repositórios analisados, com 67,7% e 44,1% de popularidade respectivamente - com C++, o qual possui valores superiores mesmos sendo uma linguagem menos popular (com 23,9% dos votos).

Tabela 4.1

		Métricas			
		Popularidade (% de votos)	**Nº de perguntas no S.O.**	**Nº de respostas no S.O.**	**Nº de visualizações**
Linguagem	**JavaScript**	67,7%	3.023	3.098	4.308.182
	Python	44,1%	164	188	409.629
	Java	40,2%	0	0	0
	Shell	33,1%	0	0	0
	TypeScript	25,4%	3	3	347
	C++	23,9%	3.870	3.586	5.503.698
	Dart	4,0%	4	4	1.216
	N/A	-	0	0	0

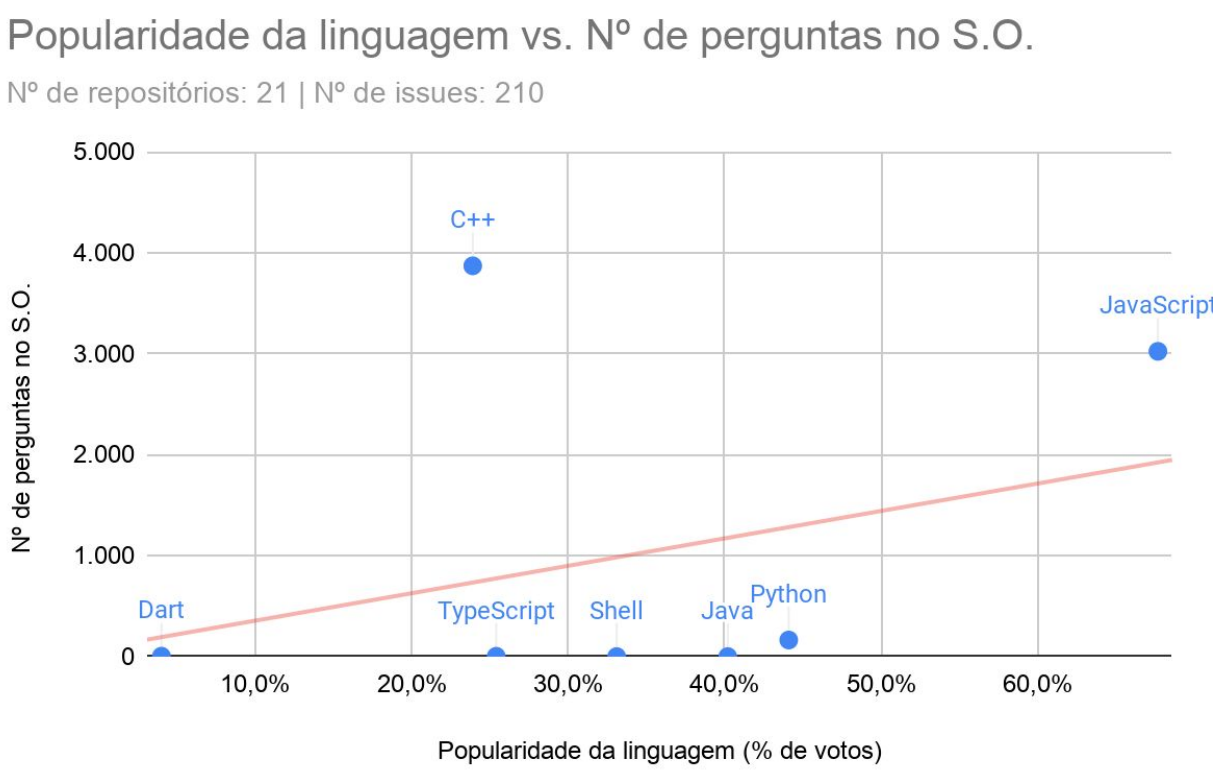

Gráfico 4.1

Hipótese 4: "Existe uma relação entre a popularidade da linguagem do repositório e o *buzz* gerado."

Falsa. Conforme descrito na resposta da questão 4, associada à esta hipótese, não existe nenhuma correlação entre a popularidade de um repositório e o *buzz* gerado.

Questão 5: "Existe alguma relação entre a popularidade da *issue* com a sua frequência de menções no StackOverflow?"

Não. Como demonstrado a seguir nos gráficos 5.1 e 5.2, não há nenhuma correlação entre a popularidade de uma *issue* (pelo nº de comentários) e a quantidade de perguntas relacionadas no Stack Overflow, seja aberta ou fechada.

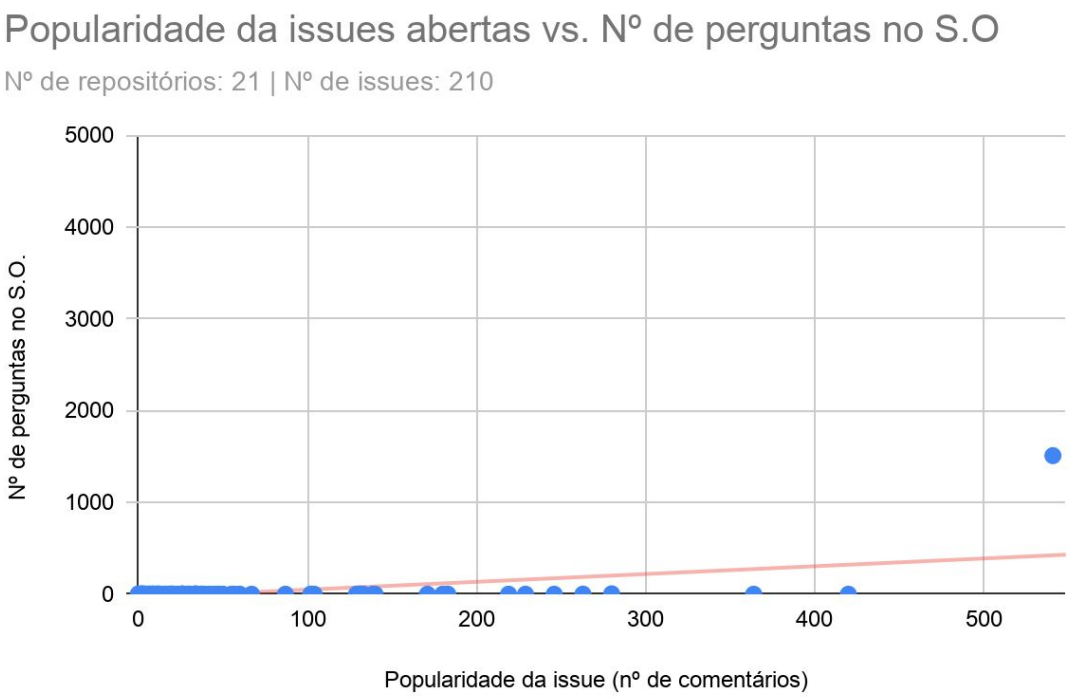

Gráfico 5.1

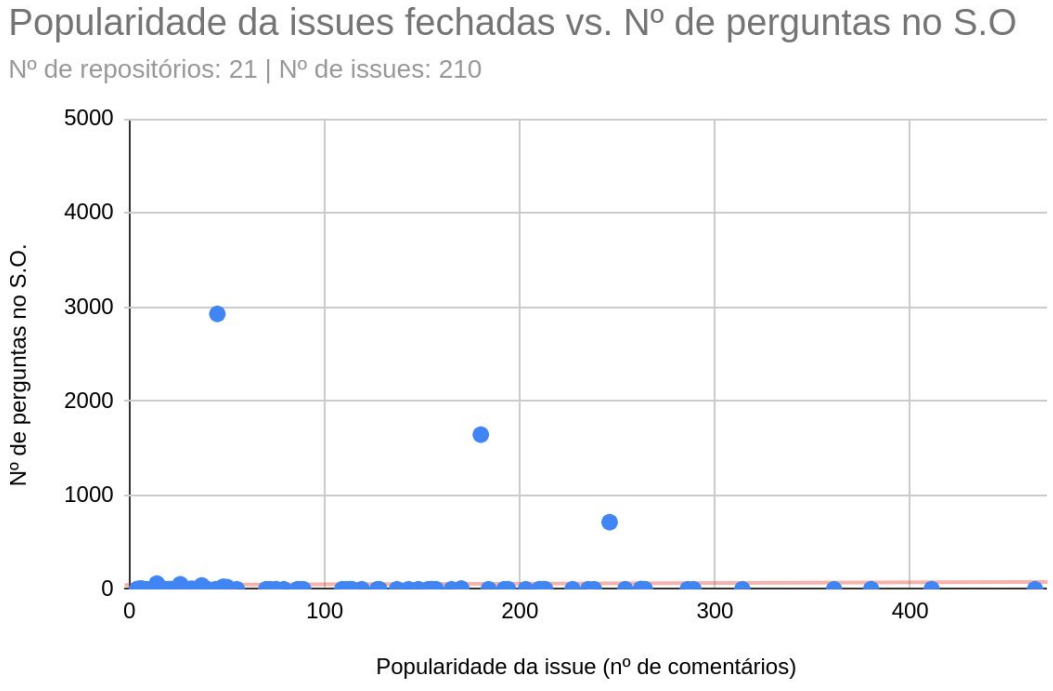

Gráfico 5.2

Hipótese 5: "Existe uma relação entre a popularidade da *issue* com a sua frequência de menções no StackOverflow."

Falsa. Conforme descrito na resposta da questão 5, associada à esta hipótese, não existe nenhuma correlação entre a popularidade de uma *issue* e a sua frequência de menções (tanto em perguntas quanto em respostas) no Stack Overflow.

Questão 6: "Existe alguma relação entre a frequência de discussão no StackOverflow das *issues* de um repositório com a sua taxa de *issues* fechadas?"

Não. Como demonstrado no gráfico abaixo, por mais que exista alguns repositórios com alta taxa de *issues* fechadas - como d3/d3 e tensorflow/tensorflow -, isto não é um padrão encontrado em outros repositórios que possuem a mesma taxa. Isto se torna muito evidente quando encontramos muitos repositórios com taxa semelhante (entre 80% e 100%), mas sem qualquer pergunta relacionada à suas *issues* mais populares.

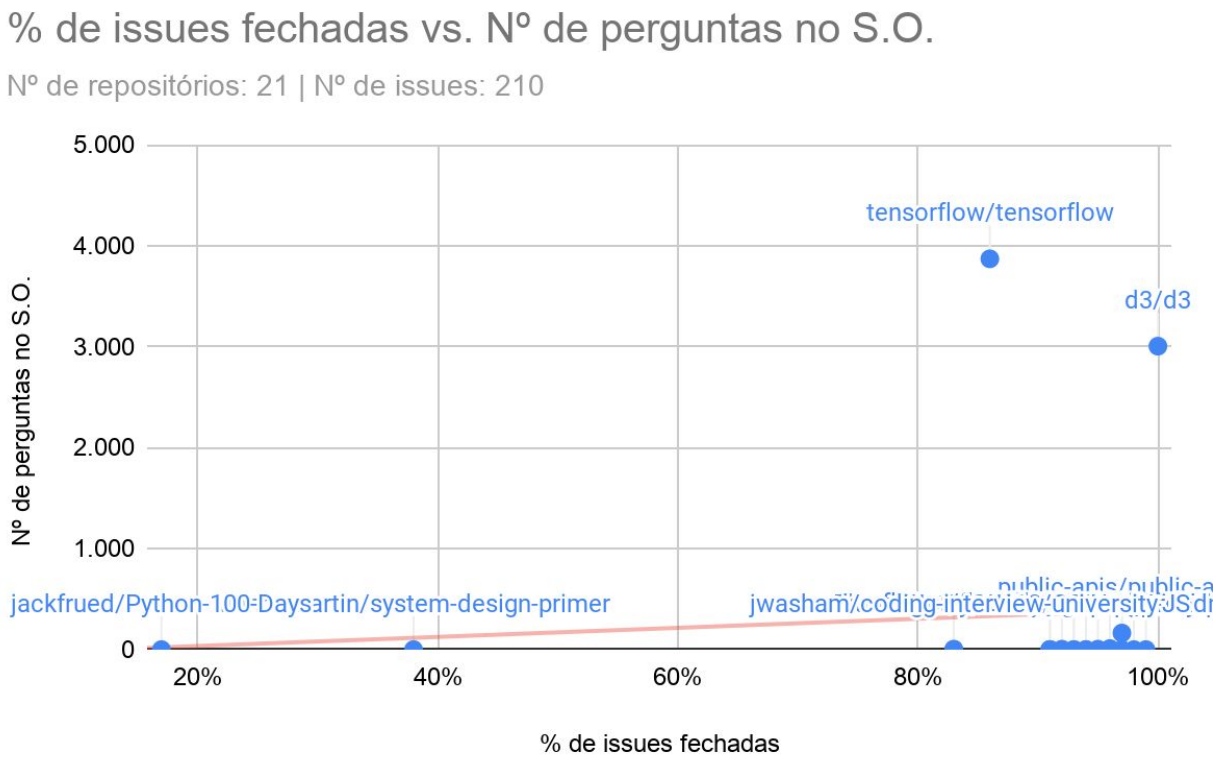

Gráfico 6.1

Hipótese 6: "Não existe nenhuma relação entre a frequência de discussão no StackOverflow das *issues* de um repositório com a sua taxa de *issues* fechadas."

Verdadeira. Conforme descrito na resposta da questão 6, associada à esta hipótese, não existe nenhuma correlação entre a frequência de discussão das *issues* de um repositório e a sua taxa de *issues* fechadas.

Ameaças a validade

- **Generalização dos resultados**

  A definição de um *dataset* que englobe tanto *issues* fechadas quanto para *issues* abertas teve o objetivo de permitir a generalização para os dois tipos de estado, não se atrelando a somente a 1 deles. Além disso, 10 *issues* para cada repositório pode não ser o suficiente para representar todas as suas *issues*, por isto foi pego as mais populares de acordo com o número de comentários.

  Portanto, estas ações tiveram o objetivo de mitigar essa ameaça a validade. Sua eficiência poderá ser comprovada em experimentos futuros, com mais repositórios e *issues* analisadas.

- **Método de identificação de questões relacionadas às *issues***

  O método utilizado para identificar questões relacionadas às *issues* foi feito através de da funcionalidade de busca avançada disponibilizada pela própria API do Stack Overflow. Porém, a *query* utilizada pode ter sido insuficiente para encontrar as perguntas relacionadas, já que necessita que tenha tanto a ocorrência do "<usuário>/<repositório>" quando do número da *issue*.

  Isto pode ser melhorado em um trabalho futuro através da utilização de outras rotas.

Conclusão

Com a associação de algumas métricas do GitHub - como estrelas, *issues* e *comentários* - com algumas métricas do Stack Overflow - como questões, respostas e visualizaçoes - dos 17 repositórios públicos mais populares, foi possível responder todas as questões propostas e validar algumas hipóteses que respondam constantes dúvidas no ambiente de projetos *open source*, como o impacto do Stack Overflow no processo de desenvolvimento com o GitHub.

Referências

1. The expanding Octoverse. Disponível em: https://octoverse.github.com (acessado em 23/09/2020);
2. Measures of Central Tendency: Mean, Median, and Mode. Disponível em: https://www.sciencebuddies.org/science-fair-projects/science-fair/summarizing-your-data#meanmedianandmode (acessado em 25/09/2020);
1. Most Popular Technologies: Programming, Scripting, and Markup Languages. Disponível em: https://insights.stackoverflow.com/survey/2020/#most-popular-technologies (acessado em 23/09/2020);