Pontifícia Universidade Católica de Minas Gerais

Praça da Liberdade – Noite

Engenharia de Software

Laboratório de experimentação de software

Professor: Jose Laerte Xavier

Alunos: Felipe Fantoni, Mateus Fonseca

Avaliando o Buzz de Issues do Github no StackOverflow

Introdução

Este relatório tem como objetivo realizar o agrupamento de hipóteses e resultados obtidos sobre as seis perguntas propostas na atividade denominada Avaliando o Buzz de Issues do Github no StackOverflow. Os dados contidos neste relatório foram obtidos no dia 14/05/2020.

Questões de pesquisa (RQS):

As três primeiras foram definidas pelo professor:

- Com que frequência issues do GitHub são discutidas no StackOverflow?
- Qual o impacto das discussões de issues do GitHub no StackOverflow?
- Existe alguma relação entre a popularidade dos repositórios e o buzz gerado?
- No contexto dos repositórios mais populares do GitHub, suas linguagens primárias são frequentes no corpo das questões mais populares do stackoverflow?
- As issues mais populares dos repositórios mais populares do GitHub possuem uma alta frequência de menções nas questões mais populares do stackoverflow?
- As issues menos populares dos repositórios mais populares do GitHub possuem uma alta frequência de menções nas questões mais populares do stackoverflow?

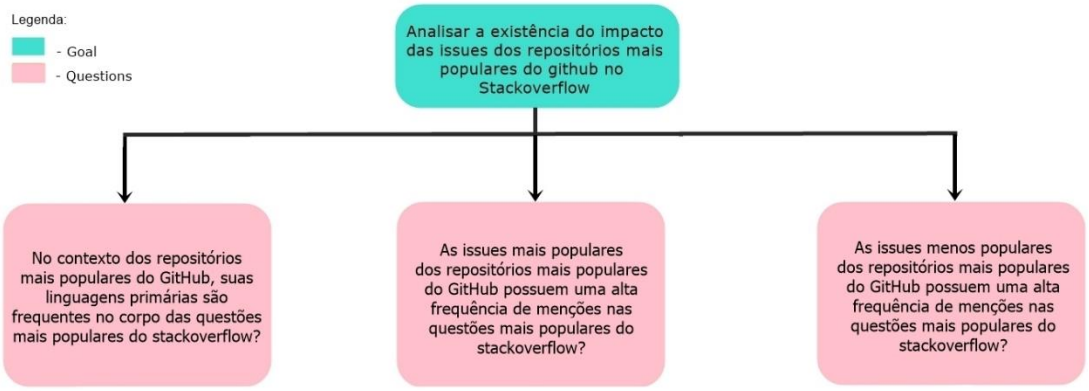
OBS.: Para verificar a menção de issues dos repositórios do GitHub nas questions do SO, estamos utilizando uma técnica nova: Verificamos se cada uma das palavras do título da issues está contida em qualquer tag da questão do stackoverflow. Se estiver, são relacionadas.

Métricas das nossas perguntas:

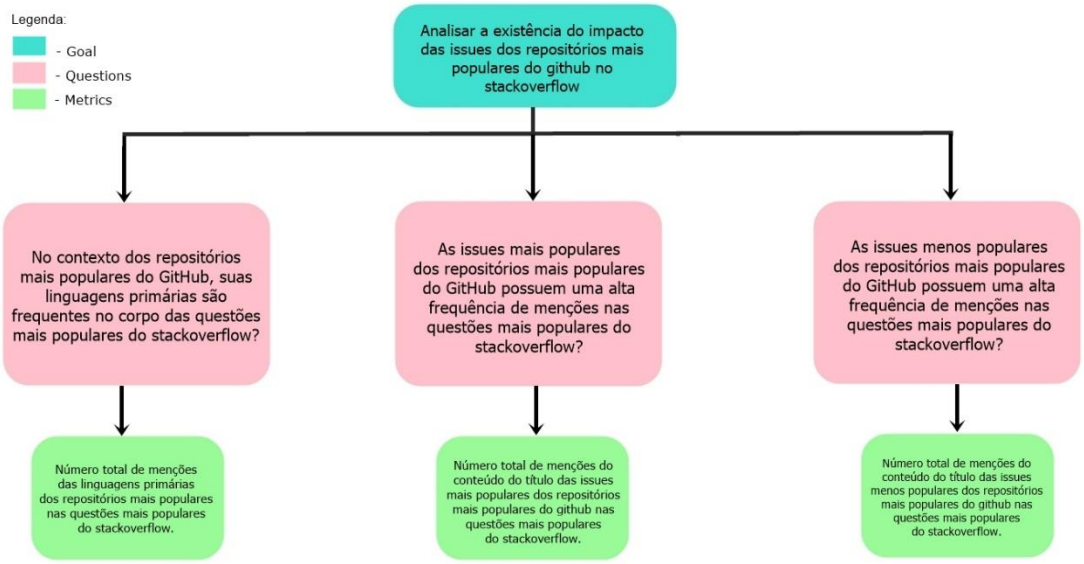
- Número total de menções das linguagens primárias dos repositórios mais populares nas questões mais populares do stackoverflow.
- Número total de menções do conteúdo do título das issues mais populares dos repositórios mais populares do github nas questões mais populares do stackoverflow.
- Número total de menções do conteúdo do título das issues menos populares dos repositórios mais populares do github nas questões mais populares do stackoverflow.

Representações gráficas do RQS e GQM:

RQS:



GQM:



Definição do conjunto de dados estudados (Datasets):

Iremos estudar as TOP 100 questões mais populares do StackOverflow (nossa definição de popularidade de uma questão do StackOverflow leva em consideração o score das questões, ordenadas descentemente) e iremos estudar os TOP 100 repositórios mais populares do GitHub. Além disso, para cada repositório popular do GitHub, iremos analisar as suas 50 issues mais populares e 50 menos populares (a referência utilizada para popularidade das issues é a sua quantidade total de comentários).

Metodologia

Para desenvolver os scripts necessários para concluir o objetivo do projeto, foi utilizada a linguagem Python. Além disso, tendo em vista a necessidade de obter informações do GitHub e do Stackoverflow, foram utilizadas a API GraphQL do GitHub e a API do Stackoverflow.

Seguindo a ordem de execução dos scripts, a primeira etapa foi, utilizando a API do SO, recuperar as top 100 Questions do SO, utilizando a própria API do SO. Estas questions, assim que recuperadas, são exportadas para um arquivo .csv com os campos "Question ID", "Quant. views", "Quant. answers", "Score", "Título da Question", "Tags" e "Link da Question". Logo após a recuperação das questions, recuperamos os top 100 Repositórios mais populares do GitHub (independentemente da linguagem), usando a API do próprio GitHub.

Finalmente, com todos os repositórios e questions retornados, para cada repositório, é buscado suas TOP 50 Issues e BOTTOM 50 Issues mais populares e suas informações são exportadas em 3 arquivos diferentes: repositorio_info.csv, top_issues_info.csv e bottom_issues_info.csv (lembrando que esses arquivos são exportados para cada repositório recuperado).

Finalmente, como já dito anteriormente, para cada repositório iterado, é verificado:

- Se a linguagem do repositório está contida nas tags das questions do SO (Existe uma coluna que conta valores repetidos e uma que descarta valores repetidos.)

- Se cada palavra obtida através do split do título das TOP issues possui menções nas tags das questions do SO (Existe uma coluna que conta valores repetidos e uma que descarta valores repetidos.)

- Se cada palavra obtida através do split do título das BOTTOM issues possui menções nas tags das questions do SO (Existe uma coluna que conta valores repetidos e uma que descarta valores repetidos.)

Finalmente, um arquivo .csv chamado resultado_analise_final.csv é gerado, contendo, para cada repositório, as informações anteriormente obtidas: “Nome repositório”, "Nome repositório", "Linguagem primária", "Qt. menções da LP nas quest. do SO", "Qt. menções da LP nas quest. do SO s/ repetição", "Qt. menções do título das top issues nas quest. do SO", "Qt. respostas da questions relacionadas as top issues", "Score das questions relacionadas as top issues", "Qt. menções do título das bottom issues nas quest. do SO", "Qt. respostas da questions relacionadas as bottom issues", "Score das questions relacionadas as bottom issues".

Hipóteses informais

RQ 01. Com que frequência issues do GitHub são discutidas no StackOverflow?

- Acreditamos que apenas as issues menos populares dos respositórios mais populares do GitHub irão possuir uma alta frequência de discussão no StackOverflow.

RQ 02. Qual o impacto das discussões de issues do GitHub no StackOverflow?

- Supomos que o impacto das discussões de issues do GitHub no StackOverflow não será grande em sua totalidade.

RQ 03. Existe alguma relação entre a popularidade dos repositórios e o buzz gerado?

- Acreditamos que o buzz gerado estará diretamente ligado a popularidade dos repositórios do GitHub.

RQ 04. No contexto dos repositórios mais populares do GitHub, suas linguagens primárias são frequentes no corpo das questões mais populares do StackOverflow?

- Supomos que as linguagens primárias dos respositórios mais populares do GitHub serão frequentemente encontradas no corpo das questões mais populares do StackOverflow.

RQ 05. As issues mais populares dos repositórios mais populares do GitHub possuem uma alta frequência de menções nas questões mais populares do StackOverflow?

- Acreditamos que as issues mais populares dos repositórios mais populares do GitHub não irão possuir uma frequência alta de menções nas questões mais populares do StackOverflow.

RQ 06. As issues menos populares dos respositórios mais populares do GitHub possuem uma alta frequência de menções nas questões mais populares do StackOverflow?

- Acreditamos que as issues menos populares dos repositórios mais populares do GitHub não irão possuir uma frequência alta de menções nas questões mais populares do StackOverflow, porém, uma menção maior que as issues mais populares.

Respostas e Discussão dos resultados obtidos

RQ 01. Com que frequência issues do GitHub são discutidas no StackOverflow?

Repositório	Menções do título das top issues nas quest. do SO	Menções do título das bottom issues nas quest. do SO	Menções totais
freeCodeCamp	71	37	108
996.ICU	0	0	0
vue	24	32	56
. . .			
Totais:	3340	2885	6225

- Utilizamos a tabela acima para, em uma menor escala, demonstrar como foi feito o cálculo para essa resposta. Somamos a quantidade de menções do título das top issues nas questões do StackOverflow com a quantidade de menções do título das bottom issues nas questões do StackOverflow, e realizamos um somatório dos valores da Coluna Menções Totais, para chegarmos ao valor de 6.225 perguntas relacionadas no total.

- A hipótese se mostrou falsa, pois as issues mais populares possuem um maior número de menções.

RQ 02. Qual o impacto das discussões de issues do GitHub no StackOverflow?

Repositório	Respostas da questions relacionadas as top issues	Respostas da questions relacionadas as bottom issues	Respostas relacionadas totais
freeCodeCamp	3361	1657	5018
996.ICU	0	0	0
vue	1053	1548	2601
. . .			
Totais:	139298	117839	257137

- Utilizamos a tabela acima para, em uma menor escala, demonstrar como foi feito o cálculo para essa resposta. Somamos a quantidade de respostas da questions relacionadas as top issues com a quantidade de respostas da questions relacionadas as bottom issues, e realizamos um somatório dos valores da Respostas relacionadas totais, para chegarmos ao valor de 257.137 perguntas relacionadas no total. Por fim, dividimos as menções totais pelas respostas relacionadas totais e alcançamos o valor de 0,024.

- A hipótese se mostrou verdadeira.

RQ 03. Existe alguma relação entre a popularidade dos repositórios e o buzz gerado?

Repositório	Score das	Score das	Score total

	questions relacionadas as top issues	questions relacionadas as bottom issues	
freeCodeCamp	373910	176779	550689
996.ICU	0	0	0
vue	156971	159003	315974
. . .			
Totais:	18951104	16156095	35107199

- Utilizamos a tabela acima para, em uma menor escala, demonstrar como foi feito o cálculo para essa resposta. Somamos a quantidade de score das questions relacionadas as top issues com a quantidade de score das questions relacionadas as bottom issues, e realizamos um somatório dos valores do Score total, para chegarmos ao valor de 35.107.199 de score total. Por fim, multiplicamos o score total pelas respostas relacionadas totais e alcançamos o valor de 842.573.

- A hipótese se mostrou verdadeira.

RQ 04. No contexto dos repositórios mais populares do GitHub, suas linguagens primárias são frequentes no corpo das questões mais populares do StackOverflow?

Linguagem primária	Quantidade de menções da linguagem nas questões do SO
JavaScript	25
Rust	0
C++	3
Shell	2
Phyton	11

- Não necessáriamente. Algumas linguagens que são mais populares, são encontradas com frequencia consideramente alta nas questões, porém, os repositórios com linguagens não tão populares não possuem alto índice de citação.

- A hipótese se mostrou parcialmente falsa, pois não necessariamente as linguagens do repositórios mais populares são frequentemente citadas no corpo das questões mais populares.

RQ 05. As issues mais populares dos repositórios mais populares do GitHub possuem uma alta frequência de menções nas questões mais populares do StackOverflow?

Repositório	Quantidade de menções do título das top issues nas quest. do SO
freeCodeCamp	71
996.ICU	0
vue	24
react	17
freeCodeCamp	37
. . .	
Mediana	22

- O valor da mediana de menções do título das top issues nas questões do StackOver é 22, logo, consideramos que as menções não são tão frequentes assim. Vale citar que os repositórios menos citados possuem zero citações, e o mais citado possui 207 citações.

- A hipótese se mostrou verdadeira.

RQ 06. As issues menos populares dos respositórios mais populares do GitHub possuem uma alta frequência de menções nas questões mais populares do StackOverflow?

Repositório	Quantidade de menções do título das bottom issues nas quest. do SO
freeCodeCamp	37
996.ICU	0
vue	32
react	5
freeCodeCamp	50
. . .	
Mediana	22

- O valor da mediana de menções do título das bottom issues nas questões do StackOver é 22, logo, consideramos que as menções não são tão frequentes assim. Vale citar que os repositórios menos citados possuem zero citações, e o mais citado possui 232 citações.

- A hipótese se mostrou parcialmente falsa, pois o valor da mediana das botton issues coincidiu com o valor da mediana das top issues, ao invés de ser maior como previmos.

Ameaças a validade identificados

A principal e única ameaça a validade identificada foi de **construção**: O método utilizado para verificar se uma issue estava relacionada com uma pergunta utilizado em nossa pesquisa se dá na verificação se qualquer palavra do título splitado da issue está contido nas tags da question. Tendo em vista este método, temos como ameaça o fato de que não é possível afirmar, com certeza, que essa palavra realmente é determinante para, caso a mesma seja identificada na tag da question, relacionar a question com a issue.

Ex.: No título da issue contém a palavra “JavaScript”. Tendo em vista que é comum se colocar a

linguagem sobre o que a question aborda nas tags do stackoverflow, este caso haverá um match, caso a question seja sobre JavaScript. Porém, embora estejam falando da mesma linguagem, o conteúdo desta question não é, necessáriamente, relacionada a issue comparada.