

Relatório Sprint 2 - Laboratório 03 - Caracterizando a Atividade de Code Review no GitHub

Integrantes do grupo:

- Bernardo Cruz Rohlf
- Leonardo Augusto Pereira do Carmo
- Matheus Belo Santos Mello
- Tarcísio Ney Martins Filho

Professora: Aline Norberta de Brito

Introdução

Resumo:

Este artigo aborda a prática de code review, fundamental nos processos de desenvolvimento ágeis, especialmente no contexto de sistemas open source hospedados no GitHub. A prática de code review envolve a revisão do código produzido por desenvolvedores antes de sua integração na base principal, visando garantir a qualidade e evitar a inclusão de defeitos. Neste estudo, o foco está na análise das atividades de code review em repositórios populares do GitHub, com o objetivo de identificar as variáveis que influenciam o merge de um Pull Request (PR) sob a perspectiva dos desenvolvedores que submetem código aos repositórios selecionados.

Introdução:

A prática de code review é essencial para garantir a qualidade do código integrado em projetos de desenvolvimento de software. No contexto de sistemas open source hospedados no GitHub, as atividades de code review ocorrem principalmente através da avaliação de Pull Requests (PRs). Um PR é uma solicitação de integração de código em uma branch principal do repositório, que é revisada e

discutida por colaboradores do projeto antes de ser aprovada ou rejeitada para merge.

Objetivo:

O objetivo deste trabalho é analisar a atividade de code review em repositórios populares do GitHub, com o intuito de identificar as variáveis que influenciam o merge de um PR, sob a perspectiva dos desenvolvedores que submetem código aos repositórios selecionados. Algumas das variáveis a serem consideradas incluem o tamanho dos PRs, o tempo de análise, a qualidade da descrição do PR, as interações durante o processo de revisão, entre outras.

Perguntas a serem respondidas:

RQ 01. Qual a relação entre o **tamanho** dos PRs e o feedback final das revisões?

Hipótese Informal: PRs menores tendem a receber feedback mais rápido e com menos problemas identificados, levando a um processo de revisão mais rápido e a um merge mais suave.

RQ 02. Qual a relação entre o **tempo de análise** dos PRs e o feedback final das revisões?

Hipótese Informal: PRs que são analisados mais rapidamente tendem a ter um feedback mais positivo, já que os problemas são identificados e resolvidos de maneira mais eficiente.

RQ 03. Qual a relação entre a **descrição** dos PRs e o feedback final das revisões?

Hipótese Informal: PRs com descrições claras e detalhadas têm maior probabilidade de receber um feedback positivo, pois facilitam a compreensão do que foi implementado e quais são as expectativas de revisão.

RQ 04. Qual a relação entre as **interações** nos PRs e o feedback final das revisões?

Hipótese Informal: Um maior número de interações (comentários, revisões, sugestões) nos PRs pode indicar uma maior colaboração e engajamento da equipe, resultando em um feedback mais completo e construtivo.

RQ 05. Qual a relação entre o **tamanho** dos PRs e o número de revisões realizadas?

Hipótese Informal: PRs maiores tendem a exigir mais revisões devido à complexidade e ao número de mudanças, enquanto PRs menores podem ser revisados mais rapidamente e com menos iterações.

RQ 06. Qual a relação entre o **tempo de análise** dos PRs e o número de revisões realizadas?

Hipótese Informal: PRs que são analisados mais rapidamente geralmente passam por menos revisões, pois os problemas são identificados e corrigidos mais cedo no processo.

RQ 07. Qual a relação entre a **descrição** dos PRs e o número de revisões realizadas?

Hipótese Informal: PRs com descrições completas e claras podem reduzir a necessidade de revisões, pois os revisores têm uma compreensão mais abrangente do que está sendo proposto.

RQ 08. Qual a relação entre as **interações** nos PRs e o número de revisões realizadas?

Hipótese Informal: Um maior número de interações nos PRs pode indicar um processo de revisão mais detalhado e abrangente, levando a um número maior de revisões antes do merge.

Metodologia

O dataset utilizado nesta pesquisa será composto por Pull Requests (PRs) submetidos a repositórios populares do GitHub. A seleção dos PRs seguirá critérios específicos:

Repositórios Populares: Serão considerados os 200 repositórios mais populares do GitHub, com base em métricas como estrelas, forks e atividade.

Quantidade Mínima de PRs: Serão selecionados apenas os repositórios que possuam pelo menos 100 PRs, incluindo tanto PRs MERGED quanto CLOSED.

PRs com Processo de Code Review: A inclusão ocorrerá somente para PRs com status MERGED ou CLOSED e que possuam ao menos uma revisão (conforme o campo `review_total_count`).

Exclusão de PRs Revisados Automaticamente: PRs revisados automaticamente por bots ou ferramentas de CI/CD serão excluídos do dataset. Serão considerados apenas os PRs cuja revisão tenha levado pelo menos uma hora, ou seja, a diferença entre a data de criação e a data de merge (ou close) seja maior que uma hora.

Com base no dataset criado, a pesquisa visa responder às seguintes questões de pesquisa, divididas em duas dimensões:

Feedback Final das Revisões (Status do PR):

Qual a relação entre o tamanho dos PRs e o feedback final das revisões?

Qual a relação entre o tempo de análise dos PRs e o feedback final das revisões?

Qual a relação entre a descrição dos PRs e o feedback final das revisões?

Qual a relação entre as interações nos PRs e o feedback final das revisões?

Número de Revisões:

Qual a relação entre o tamanho dos PRs e o número de revisões realizadas?

Qual a relação entre o tempo de análise dos PRs e o número de revisões realizadas?

Qual a relação entre a descrição dos PRs e o número de revisões realizadas?

Qual a relação entre as interações nos PRs e o número de revisões realizadas?

Para cada dimensão e questão de pesquisa, serão utilizadas as seguintes métricas para realizar as análises e correlações:

Tamanho: Número de arquivos modificados e total de linhas adicionadas e removidas nos PRs.

Tempo de Análise: Intervalo de tempo entre a criação do PR e a última atividade registrada (merge ou close).

Descrição: Tamanho do corpo de descrição dos PRs em caracteres, utilizando a versão markdown para essa contagem.

Interações: Número de participantes envolvidos nos PRs e número de comentários realizados durante o processo de revisão.

Essas métricas permitirão explorar e compreender as relações entre as variáveis estudadas e os resultados obtidos nos processos de code review em repositórios populares do GitHub.

Resultados

RQ 01.

Contextualizar:

Gráfico:

Conclusão (Hipótese Informal x Resultados Encontrados):

RQ 02.

Contextualizar:

Gráfico:

Conclusão (Hipótese Informal x Resultados Encontrados):

RQ 03.

Contextualizar:

Gráfico:

Conclusão (Hipótese Informal x Resultados Encontrados):

RQ 04.

Contextualizar:

Gráfico:

Conclusão (Hipótese Informal x Resultados Encontrados):

RQ 05.

Contextualizar:

Gráfico:

Conclusão (Hipótese Informal x Resultados Encontrados):

RQ 06.

Contextualizar:

Gráfico:

Conclusão (Hipótese Informal x Resultados Encontrados):

RQ 07.

Contextualizar:

Gráfico:

Conclusão (Hipótese Informal x Resultados Encontrados):

RQ 08.

Contextualizar:

Gráfico:

Conclusão (Hipótese Informal x Resultados Encontrados):