Laboratório 03

Caracterizando a Atividade de Code Review no GitHub

Pedro Franco e Gabriel Pongelupe

Engenharia de Software - 6º Período

Laboratório de Experimentação de Software

Outubro de 2025

Professor: Danilo

Sumário

1 Introdução

A prática de *code review* é fundamental no desenvolvimento de software moderno, especialmente em projetos *open source* hospedados no GitHub. Por meio de *Pull Requests* (PRs), desenvolvedores submetem contribuições que são avaliadas por revisores antes de serem integradas à base principal do código. Este processo garante a qualidade do código e reduz a inclusão de defeitos no projeto.

Este trabalho tem como objetivo analisar a atividade de *code review* em repositórios populares do GitHub, identificando variáveis que influenciam no *merge* de um PR sob a perspectiva dos desenvolvedores que submetem código.

1.1 Hipóteses Informais

Com base na literatura e na experiência prática com desenvolvimento de software, formulamos as seguintes hipóteses para guiar nossa análise:

1.1.1 Dimensão A: Feedback Final das Revisões (Status do PR)

H1 - Tamanho dos PRs:

- **Hipótese:** PRs menores (com menos arquivos e menos linhas modificadas) têm maior probabilidade de serem aceitos (MERGED).
- **Justificativa:** PRs menores são mais fáceis de revisar, compreender e validar, reduzindo a carga cognitiva do revisor e facilitando a identificação de problemas.

H2 - Tempo de Análise:

- **Hipótese:** PRs que são mergeados tendem a ter tempo de análise menor do que PRs rejeitados (CLOSED).
- **Justificativa:** PRs bem estruturados e alinhados com os objetivos do projeto são revisados e aprovados mais rapidamente, enquanto PRs problemáticos geram mais discussões e acabam sendo rejeitados após longos períodos.

H3 - Descrição dos PRs:

- **Hipótese:** PRs com descrições mais detalhadas (maior número de caracteres) têm maior taxa de aceitação.
- **Justificativa:** Descrições completas facilitam o entendimento do revisor sobre o contexto, motivação e implementação das mudanças, agilizando o processo de revisão.

H4 - Interações nos PRs:

- **Hipótese:** PRs com mais interações (participantes e comentários) têm menor taxa de aceitação.
- **Justificativa:** Mais interações podem indicar controvérsias, problemas no código ou necessidade de ajustes significativos, o que pode levar à rejeição.

1.1.2 Dimensão B: Número de Revisões

H5 - Tamanho dos PRs:

- **Hipótese:** PRs maiores exigem mais revisões antes da decisão final.
- **Justificativa:** Alterações extensas requerem múltiplas rodadas de revisão para garantir a qualidade e identificar todos os problemas potenciais.

H6 - Tempo de Análise:

- **Hipótese:** Existe correlação positiva entre o número de revisões e o tempo de análise.
- **Justificativa:** Mais revisões naturalmente estendem o tempo total necessário para finalizar o processo de *code review*.

H7 - Descrição dos PRs:

- **Hipótese:** PRs com descrições mais detalhadas requerem menos revisões.
- **Justificativa:** Descrições claras reduzem dúvidas e a necessidade de esclarecimentos adicionais durante o processo de revisão.

H8 - Interações nos PRs:

- **Hipótese:** Existe correlação positiva forte entre o número de interações e o número de revisões.
- **Justificativa:** Cada revisão tende a gerar discussões e comentários, aumentando o número total de interações.

2 Metodologia

2.1 Coleta de Dados

O dataset foi construído a partir de *Pull Requests* de repositórios populares do GitHub, seguindo os critérios estabelecidos:

2.1.1 Critérios de Seleção de Repositórios

- Repositórios entre os 200 mais populares do GitHub
- Repositórios com pelo menos 100 PRs (MERGED + CLOSED)

2.1.2 Critérios de Seleção de Pull Requests

- Status: MERGED ou CLOSED
- Pelo menos uma revisão registrada (total count > 0)
- Tempo de revisão superior a 1 hora (para filtrar revisões automáticas por bots/CI-CD)

2.2 Definição de Métricas

Para responder às questões de pesquisa, coletamos as seguintes métricas para cada PR:

Métricas de Tamanho:

- Número de arquivos modificados
- Total de linhas adicionadas
- Total de linhas removidas

Métricas de Tempo:

- Data/hora de criação do PR
- Data/hora de fechamento ou *merge*
- Tempo de análise (diferença entre criação e fechamento/*merge*)

Métricas de Descrição:

- Número de caracteres no corpo da descrição do PR (formato markdown)

Métricas de Interação:

- Número de participantes únicos
- Número total de comentários

Métricas de Revisão:

- Número total de revisões realizadas
- Status final (MERGED ou CLOSED)

2.3 Ferramentas e Tecnologias

- **Linguagem de Programação:** Python 3.x
- **Biblioteca para API do GitHub:** PyGithub
- **Análise de Dados:** pandas, numpy
- **Análise Estatística:** scipy.stats
- **Visualização:** matplotlib, seaborn

2.4 Análise Estatística

Para avaliar as correlações entre as variáveis, utilizaremos o **Teste de Correlação de Spearman**:

- Escolhido por ser não-paramétrico, adequado para dados que podem não seguir distribuição normal
- Avalia correlações monotônicas (não apenas lineares)
- Robusto a *outliers*, comuns em dados de repositórios de software
- Nível de significância: $\alpha = 0.05$

Justificativa: Dados de repositórios de software frequentemente apresentam distribuições assimétricas e *outliers* (por exemplo, PRs excepcionalmente grandes ou com tempo de revisão muito longo). O teste de Spearman é mais apropriado para este cenário do que o teste de Pearson, que assume normalidade dos dados.

3 Questões de Pesquisa

3.1 Dimensão A: Feedback Final das Revisões (Status do PR)

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

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

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

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

3.2 Dimensão B: Número de Revisões

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

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

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

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

4 Processo de Coleta e Análise

4.1 Script de Coleta (Lab03S01)

O script de coleta foi desenvolvido em Python utilizando a biblioteca PyGithub para acessar a API do GitHub. O processo de coleta segue os seguintes passos:

1. Identificação dos 200 repositórios mais populares (por número de *stars*)
2. Filtragem de repositórios com pelo menos 100 PRs (MERGED + CLOSED)
3. Para cada repositório selecionado:
    - Coleta de todos os PRs com status MERGED ou CLOSED
    - Filtragem de PRs com pelo menos uma revisão
    - Cálculo do tempo de revisão e filtragem (> 1 hora)
    - Extração de todas as métricas definidas
4. Consolidação dos dados em um dataset único

4.2 Estrutura do Dataset

O dataset final contém as seguintes colunas apresentadas na Tabela **??**.

Tabela 1: Descrição das colunas do arquivo `prs_clean.csv`

Coluna	Descrição	Tipo
`id`	Identificador único do PR	Integer
`number`	Número do Pull Request	Integer
`title`	Título do Pull Request	String
`user`	Autor do Pull Request	String
`created_at`	Data e hora de criação do PR	Datetime
`closed_at`	Data e hora de fechamento do PR	Datetime
`merged_at`	Data e hora de merge (se aplicável)	Datetime
`comments`	Número de comentários gerais	Integer
`review_comments`	Número de comentários de revisão	Integer
`changed_files`	Quantidade de arquivos modificados	Integer
`additions`	Linhas adicionadas	Integer
`deletions`	Linhas removidas	Integer
`state`	Estado final do PR (ex: closed)	String
`merged`	Indica se foi mergeado (True/False)	Boolean
`body_length`	Número de caracteres na descrição	Integer
`end_date`	Data final (merge ou fechamento)	Datetime
`review_time_h`	Tempo total de revisão (horas)	Float

5 Resultados Preliminares

Esta seção será preenchida após a coleta completa dos dados e análise estatística.

5.1 Estatísticas Descritivas do Dataset

Aguardando coleta de dados.

5.2 Respostas às Questões de Pesquisa

Aguardando análise estatística.

6 Discussão

Esta seção será preenchida após a análise dos resultados, comparando as hipóteses formuladas com os dados obtidos.

7 Conclusões

Esta seção será preenchida na versão final do relatório.