Caracterizando a Atividade de Code Review no GitHub

Hugo Poletto Alacoque Gomes[1], **Matheus Nolasco**[1],
Maria Aryene Costa[1], **Lucas Santos Rosa**[1]

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

Resumo. *Este meta-artigo descreve o estilo a ser usado na confecção de artigos e resumos de artigos para publicação nos anais das conferências organizadas pela SBC. É solicitada a escrita de resumo e abstract apenas para os artigos escritos em português. Artigos em inglês deverão apresentar apenas abstract. Nos dois casos, o autor deve tomar cuidado para que o resumo (e o abstract) não ultrapassem 10 linhas cada, sendo que ambos devem estar na primeira página do artigo.*

1. Introdução

O Code Review é uma prática constante não só nos repositórios do GitHub, como também em outras plataformas de repositórios de Git. Ela consiste na interação dos desenvolvedores com os revisores, visando inspecionar o código antes de integrá-lo na branch principal. No contexto de repositírios abertos do GitHub, a prática do Code Review é feita por meio de Pull Requests (PR).

Neste projeto, pretende-se analisar a atividade de code review desenvolvida em repositórios populares do GitHub, identificando variáveis que influenciam no merge de um Pull Request, sob a perspectiva de desenvolvedores que submetem código aos repositórios selecionados.

Para isso, será utilizado uma matriz de correlação destas variáveis para identificar a relação entre as métricas e, assim, responder as seguintes perguntas de pesquisa:

A. Feedback Final das Revisões (Status do PR):

1. RQ: Qual a relação entre o tamanho dos PR's e o feedback final das revisões?
2. RQ: Qual a relação entre o tempo de análise dos PR's e o feedback final das revisões?
3. RQ: Qual a relação entre a descrição dos PR's e o feedback final das revisões?
4. RQ: Qual a relação entre as interações nos PR's e o feedback final das revisões?

B. Número de Revisões:

1. RQ: Qual a relação entre o tamanho dos PR's e o número de revisões realizadas?
2. RQ: Qual a relação entre o tempo de análise dos PR's e o número de revisões realizadas?
3. RQ: Qual a relação entre a descrição dos PR's e o número de revisões realizadas?
4. RQ: Qual a relação entre as interações nos PR's e o número de revisões realizadas?

1.1. Hipóteses Iniciais

A partir de cada uma dessas perguntas, hipóteses iniciais (HI) foram formuladas. Essas "hipóteses informais" são, respectivamente:

1. HI-RQ: PR's maiores, recebem feedbacks maiores, uma vez que para haver uma observação, é necessário avaliar uma quantidade maior de mudanças.
2. HI-RQ: PR's que tem um tempo de análise menor, geralmente são aceitos, ou seja, merged ou closed de forma positiva.
3. HI-RQ: PR's com descrições mais detalhadas geralmente são aceitos, ou seja, merged ou closed de forma positiva.
4. HI-RQ: PR's com mais interações tendem a alcançar uma resolução positiva, porque recebem mais colaboração.
5. HI-RQ: O tamanho dos PR's é proporcional ao número de revisões realizadas.
6. HI-RQ: PR's que demoram mais para serem analisados geralmente têm um maior número de revisões em comparação com PR's analisados rapidamente.
7. HI-RQ: PR's que tem descrições mais claras, tem um maior número de revisões.
8. HI-RQ: PR's com um número maior de interações (comentários, discussões) geralmente têm um maior número de revisões em comparação com PR's com menos interações.

2. Metodologia

2.1. Criação do Dataset

O dataset utilizado neste laboratório será composto por PR's de 200 repositórios que tenham passado pelo processo de Code Review nos repositórios mais populares do GitHub que possuam pelo menos 100 PR's mesclados e fechados (Merged + Closed), no mínimo 1 PR com revisão nos últimos 1000 PR's e que levou pelo menos uma hora de revisão para remover PR's revisadas de forma automática por meio de bots ou ferramentas de CI/CD.

2.2. Definição de Métricas

Para cada dimensão, as correlações com as métricas serão feitas da seguinte forma:

- **Tamanho**: número de arquivos; total de linhas adicionadas e removidas.
- **Tempo de Análise**: intervalo entre a criação do PR e a última atividade (fechamento ou mesclagem).
- **Descrição**: número de caracteres do corpo de descrição do PR.
- **Interações**: número de participantes; número de comentários.

2.3. Processo de Desenvolvimento

- **Sprint 01**: Lista de repositórios selecionados + Criação do script de coleta dos PR's e das métricas definidas.
- **Sprint 02**: Dataset completo, com os valores de todas as métricas necessárias + Primeira versão do relatório final, com as hipóteses iniciais.
- **Sprint 03**: Análise e visualização de dados + Elaboração do relatório final.

3. Resultados

Os resultados podem ser vistos na seguinte matriz de correlação de Pearson

References

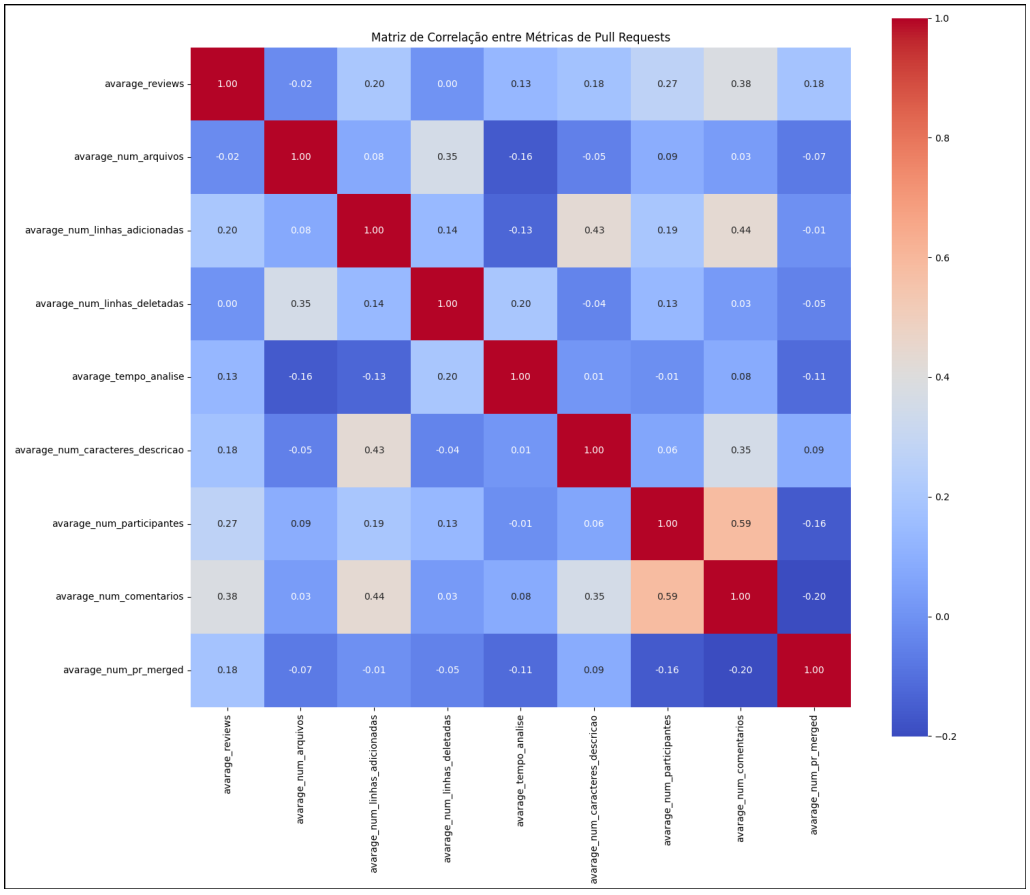

Figure 1. Esta figura é um matriz de correlação entre as métricas .