A influência da documentação README no engajamento de repositórios no GitHub

Beatriz de Oliveira Silveira[1], **Joaquim de Moura Thomaz Neto**[1], **Maria Eduarda Chrispim Santana**[1], **Matheus Pereira**[1], **Vitor Fernandes de Souza**[1], **Vitória Ye Miao**[1]

[1]Pontifícia Universidade Católica de Minas Gerais (PUC Minas) – Belo Horizonte, MG

{beatriz.silveira, joaquim.thomaz, maria.santana, matheus.pereira, vitor.souza, vitor

Resumo. *Os arquivos README de repositórios no GitHub servem para introduzir e guiar um projeto, trazendo informações essenciais, como sua finalidade e instruções de uso. Repositórios que são bem documentados melhoram a experiência dos desenvolvedores e podem estar associados à popularidade de um repositório. Entretanto, essa popularidade é frequentemente relacionada apenas à qualidade de código, negligenciando a importância da documentação. Dessa forma, para investigar essa hipótese, o presente estudo coleta dados (README e histórico de commits) de 933 repositórios públicos no GitHub — desconsiderando repositórios de livros e exemplos — e analisa a estrutura e a qualidade da documentação em relação ao tempo do primeiro e do último commit no README, a fim de investigar a relação entre a qualidade da documentação de um repositório e sua popularidade. Os resultados revelam que a completude de uma documentação e atualizações no README podem impactar significativamente na popularidade.*
Palavras-chave: documentação, popularidade, qualidade, métricas.

1. Introdução

Durante os últimos anos, o modelo open source tem crescido cada vez mais. Diferente de projetos de software em que há apenas um proprietário desenvolvendo um código, o open source consiste em um código-fonte aberto, em que qualquer pessoa pode ver e modificar de acordo com suas necessidades. Isso permite que as comunidades possam desenvolver projetos de forma mais prática e colaborativa. Muitos desses projetos são hospedados no GitHub, no qual é possível ter acesso a repositórios e participar de comunidades.

Dentro de um repositório no GitHub, é possível armazenar códigos, acompanhar histórico de versões e aprovar modificações de um projeto, além de permitir realizar documentações de código e informações adicionais, como instruções de utilização. Uma das formas de documentar um projeto é utilizando o README, um arquivo que introduz o projeto e apresenta informações, como finalidade, funcionalidades, instruções de instalação e uso. Assim sendo, ele é importante para facilitar a compreensão para outros usuários e ajudar outras pessoas a encontrar o projeto e se interessarem em contribuir, promovendo, assim, a colaboração.

No entanto, poucos desenvolvedores realizam documentações em projetos open source, e, consequentemente, menos pessoas desejam contribuir para o projeto, já que,

sem documentação, é reduzido o entendimento sobre a finalidade e funcionalidade deste. Logo, uma documentação pobre pode reduzir a participação de outros contribuidores, e reduzir o potencial de crescimento e manutenção do projeto. Ademais, percebe-se que a popularidade de um repositório no GitHub é frequentemente associada à qualidade do código, o que pode levar os desenvolvedores a negligenciarem a documentação, que também é importante em um projeto.

Desse modo, deve-se considerar que o código e a documentação formam um pacote, e que mesmo um código bem escrito necessita de uma documentação clara para ser entendido, de tal forma que investir na qualidade do README é essencial para garantir o sucesso de um projeto.

Os arquivos README se categorizam em diferentes seções, como introdução, instruções de instalação e exemplos de uso, que contribuem para a organização da documentação [Prana et al. 2018]. Portanto, este estudo visa analisar a qualidade e as características dos arquivos README em projetos open source hospedados no GitHub, relacionando-as a métricas de popularidade e engajamento, como contagem de issues, pull requests e forks. Além disso, o projeto tem os seguintes objetivos específicos:

- Classificar as seções mais comuns nos READMEs;
- Avaliar a relação entre a qualidade da documentação e a popularidade do projeto;
- Propor métricas para medir a qualidade dos READMEs.

Assim sendo, para realizar essa análise, são respondidas as seguintes questões de pesquisa:

- Q1: Como a qualidade do README impacta o engajamento em repositórios?
- Q2: Como mudanças na documentação ao longo do tempo estão associadas ao aumento de engajamento?
- Q3: Qual a relação entre a completude da documentação e o tempo até o primeiro engajamento?

As Q1 e Q3 buscam insights sobre a qualidade da documentação, enquanto a Q2 busca analisar se a popularidade de um repositório se deu por causa da documentação, ou se já era popular sem a presença dela. Para responder essas questões, é realizada uma análise quantitativa e qualitativa de uma amostra de projetos no GitHub, no qual é avaliado o conteúdo dos arquivos README, relacionando-os a métricas.

Este artigo está estruturado da seguinte forma: a Seção 2 apresenta o referencial teórico sobre software open source, GitHub e documentação; a Seção 3 detalha a metodologia adotada; a Seção 4 traz os resultados da análise; e, por fim, a Seção 5 apresenta as conclusões e sugestões para trabalhos futuros.

2. Trabalhos Relacionados

O estudo de Mockus et al. (2002), *Two Case Studies of Open Source Software Development: Apache and Mozilla*, investiga o processo de desenvolvimento de dois grandes projetos open source: Apache e Mozilla. Os autores analisam como a comunidade colabora de forma distribuída e quais práticas são adotadas para garantir a evolução contínua do software. Os resultados revelam que, mesmo com a ausência de uma estrutura formal

de desenvolvimento, os projetos open source conseguem manter a qualidade e a produtividade por meio de processos bem definidos e participação ativa dos colaboradores. Este estudo se relaciona com o trabalho do grupo por mostrar a importância da organização e da documentação em projetos abertos, reforçando o papel do *README* como ponto de entrada para novos contribuidores.

Prana et al. (2019), no artigo *Categorizing the Content of GitHub README Files*, analisam mais de 4.000 arquivos *README* de repositórios no GitHub com o objetivo de entender quais categorias de informações são mais frequentemente incluídas nesses documentos. Os autores identificam oito categorias principais, como instruções de instalação, uso, informações sobre contribuição e status do projeto. A pesquisa conclui que repositórios mais completos em termos de *README* tendem a facilitar a adoção e a colaboração. Este estudo é diretamente relacionado ao trabalho do grupo, que foca na qualidade e organização do *README* como fator fundamental para o sucesso e popularidade de projetos hospedados no GitHub.

No artigo *An Empirical Study On Correlation Between Readme Content and Project Popularity*, os autores investigam se há uma relação entre o conteúdo dos arquivos *README* e a popularidade dos repositórios, medida por estrelas, *forks* e contribuidores. A análise empírica revela que repositórios com *README* mais completos e bem estruturados tendem a ser mais populares. Isso evidencia que a documentação inicial é um fator importante de engajamento. Este achado apoia a proposta do grupo de analisar e melhorar a documentação de projetos visando sua maior aceitação e engajamento na comunidade open source.

O artigo *Beyond Accuracy: Assessing Software Documentation Quality* propõe uma abordagem para avaliar a qualidade de documentação de software com base em critérios como completude, clareza, organização e atualidade, indo além da simples presença de informações. O estudo mostra que a percepção de qualidade está fortemente ligada à estrutura e à utilidade da documentação para diferentes perfis de usuários. Este trabalho contribui para o projeto do grupo ao oferecer parâmetros objetivos que podem ser usados na análise e aprimoramento de *READMEs*, focando não apenas no conteúdo, mas também na experiência do usuário com a documentação.

Laerte Xavier et al., no artigo *Um Estudo Empírico sobre Critérios de Seleção de Repositórios GitHub*, analisam como usuários escolhem quais repositórios utilizar ou contribuir. Os resultados mostram que aspectos como número de estrelas, qualidade da documentação, frequência de *commits* e presença de *issues* ativas são considerados pelos usuários. O estudo é relevante para o trabalho do grupo, pois reforça a ideia de que a documentação — especialmente o *README* — é um dos principais fatores de decisão para engajamento em repositórios.

Por fim, o estudo *Identificação de Padrões de Comportamentos e Atividades de Usuários no GitHub* investiga o comportamento de usuários na plataforma, identificando padrões de navegação, interação com repositórios e tomada de decisão. Os autores destacam que a forma como um projeto se apresenta (nome, *README*, organização dos arquivos) influencia diretamente o interesse dos usuários. Este trabalho se conecta à proposta do grupo ao evidenciar que a documentação não é apenas técnica, mas também estratégica na comunicação com potenciais colaboradores.

3. References

Bibliographic references must be unambiguous and uniform. We recommend giving the author names references in brackets, e.g. [Knuth 1984], [Boulic and Renault 1991], and [Smith and Jones 1999].

The references must be listed using 12 point font size, with 6 points of space before each reference. The first line of each reference should not be indented, while the subsequent should be indented by 0.5 cm.

Referências

Boulic, R. and Renault, O. (1991). 3d hierarchies for animation. In Magnenat-Thalmann, N. and Thalmann, D., editors, *New Trends in Animation and Visualization*. John Wiley & Sons ltd.

Knuth, D. E. (1984). *The TEX Book*. Addison-Wesley, 15th edition.

Prana, G. K., Treude, C., Storey, M.-A., and Adams, B. (2018). Categorizing the content of github readme files. *Empirical Software Engineering*, 23(4):1816–1857.

Smith, A. and Jones, B. (1999). On the complexity of computing. In *Advances in Computer Science*, pages 555–566. Publishing Press.