UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

Bruno Antônio Vieira

Como os Arquivos README Evoluem ao Longo do Tempo em Projetos do Github?

Proposta de projeto científico de Monografia em Sistemas de Informação do bacharelado em Sistemas de Informação da Universidade Federal de Minas Gerais.

Orientador:
Prof. Dr. André Cavalcante Hora

Belo Horizonte
2019

INTRODUÇÃO

O Github é uma plataforma de hospedagem de código-fonte e outros arquivos importantes para um projeto de desenvolvimento, que se tornou popular entre empresas e a comunidade de desenvolvedores, em geral. Atualmente, existem mais de 100 milhões de repositórios de software hospedados no serviço [1], dentre repositórios públicos e privados.

Com a popularização do serviço, muitos projetos de código aberto passaram a ser hospedados, incentivando a colaboração entre desenvolvedores de todo o mundo. Dentro destes repositórios existe a possibilidade de se utilizar o README, um arquivo de texto que contém informações sobre o projeto, tais como, um resumo e como utilizá-lo. Esse arquivo é importante por se tratar de uma forma de documentar atualizações no código para usuários e para desenvolvedores interessados [3].

Existem indícios de que há uma correlação entre a documentação de um projeto e a sua popularidade no Github [5]. Sendo o README uma forma de documentar, é relevante entender como esse arquivo é construído em diferentes repositórios e como ele evolui conforme um projeto é atualizado.

Neste trabalho são analisados os arquivos README de 50 projetos do Github. Foram selecionados 10 projetos com mais de 10.000 estrelas, ordenados de forma decrescente, para cada uma das 5 linguagens de programação mais utilizadas. O objetivo é verificar o que muda no arquivo ao longo do tempo de sua existência no repositório, analisando por meio dos *commits* o que foi alterado e onde a alteração foi feita, além do levantamento de estatísticas básicas para verificar se existe um padrão entre os diferentes projetos de software do Github.

TRABALHOS RELACIONADOS

Alguns trabalhos sobre os arquivos README e outras formas de documentação de projetos do Github já foram conduzidos, seguindo diferentes perspectivas. Ikeda et al. [3], por exemplo, investigaram o conteúdo dos READMES de projetos escritos em Javascript, analisando apenas documentos escritos em inglês, com o intuito de saber (i) o que os desenvolvedores costumam escrever nos arquivos README e (ii) Se o tipo de projeto afeta como os desenvolvedores escrevem os arquivos README.

Prana et al. [4] seguiram um caminho diferente no estudo do README, realizando uma pesquisa qualitativa com o intuito de classificar as seções do arquivo em diferentes categorias, automaticamente. Estas categorias geradas foram então mostradas para vinte profissionais da área de software que avaliaram a qualidade da categorização. Esse estudo se mostrou importante para melhorar a qualidade da documentação e entendimento dos arquivos README do Github.

Com isso, este trabalho se apoia nos estudos anteriores, realizados sobre o mesmo tema, e tem como foco a exploração da evolução dos arquivos README ao longo do tempo, mostrando como o arquivo costuma ser alterado. A partir dessa exploração, espera-se adquirir mais conhecimento que contribua para o entendimento dos arquivos README, assim como os outros trabalhos da área.

METODOLOGIA

Inicialmente, é feita uma **leitura de artigos relacionados** com o tema proposto por esse trabalho, visando referências para o seu desenvolvimento. Além disso, é feita uma familiarização com a interface do Github e os indicadores presentes na plataforma, como o número de estrelas (stars), forks, contribuidores, watches, issues, etc. Também, são exploradas as APIs do Github que podem ser úteis para a coleta e tratamento de dados, posteriormente.

Em seguida, é realizada a **montagem do dataset**. Para isso, são selecionados os projetos que possuem mais de 10.000 estrelas (medida que pode indicar a popularidade do repositório) e são escolhidos 10 projetos, ordenados pela quantidade de estrelas, de cada uma das 5 linguagens mais utilizadas no Github. São elas: Javascript, Python, Java, Go e C++.

Na **análise dos dados**, para cada um dos arquivos README dos repositórios presentes no dataset será verificado, desde a sua criação até o último commit, onde houve alguma alteração. Pretende-se, então, apurar os tópicos que costumam mais se alterar e a semelhança das mudanças nos diferentes projetos, gerando estatísticas indicativas.

Feito isso, planeja-se a **apresentação dos resultados**, onde será exposto um mapa de calor de uma dimensão indicando os locais do arquivo README que costumam sofrer mais alterações. Além disso, serão gerados gráficos pertinentes que caracterizem padrões descobertos nestes arquivos.

Por fim, todo o trabalho e os resultados gerados serão compilados na **elaboração do texto final e apresentação do trabalho**. Um relatório bem detalhado com todos os processos da pesquisa, bem como um pôster resumido, serão produzidos para apresentação futura.

RESULTADOS ESPERADOS

Ao fim do semestre, espera-se obter informações e padrões sobre a construção dos arquivos README que serão tratados e exibidos na forma de gráficos. Estes resultados são importantes para entender como este tipo de documentação é tratado em projetos bem populares do Github, assim como desenvolvedores devem estruturar os arquivos README de seus projetos para deixá-los mais inteligíveis.

Como mostrado nos trabalhos relacionados, o README, dentre outras formas de documentar, está correlacionado com a popularidade do repositório. Assim, os resultados gerados neste trabalho podem ser aplicados em trabalhos futuros que busquem o maior engajamento de desenvolvedores em projetos do Github.

ETAPAS E CRONOGRAMA

Atividades	**Ago**	**Set**	**Out**	**Nov**	**Dez**
Leitura de artigos relacionados	■				
Montagem do dataset		■			
Análise dos dados		■	■		
Apresentação dos resultados			■	■	
Revisão do trabalho				■	
Entrega final					■

REFERÊNCIAS

[1] Thank you for 100 million repositories - The GitHub Blog. Disponível em: <https://github.blog/2018-11-08-100m-repos/> Acesso em: 4 Set. 2019

[2] GitHub's products.
Disponível em: <https://help.github.com/en/articles/githubs-products> Acesso em: 4 Set. 2019

[3] IKEDA, Shohei et al. An Empirical Study of README contents for JavaScript Packages. IEICE TRANSACTIONS on Information and Systems, v. 102, n. 2, p. 280-288, 2019.

[4] PRANA, Gede Artha Azriadi et al. Categorizing the content of GitHub README files. Empirical Software Engineering, v. 24, n. 3, p. 1296-1327, 2019.

[5] AGGARWAL, Karan; HINDLE, Abram; STROULIA, Eleni. Co-evolution of project documentation and popularity within GitHub. In: Proceedings of the 11th Working Conference on Mining Software Repositories. ACM, 2014. p. 360-363.

[6] BORGES, Hudson; HORA, Andre; VALENTE, Marco Tulio. Understanding the factors that impact the popularity of GitHub repositories. In: 2016 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE, 2016. p. 334-344.

[7] HASSAN, Foyzul; WANG, Xiaoyin. Mining readme files to support automatic building of java projects in software repositories: Poster. In: Proceedings of the 39th International Conference on Software Engineering Companion. IEEE Press, 2017. p. 277-279.