Relação entre issues do GitHub e perguntas do Stackoverflow

Isabela E. Oliveira[1], **Marco Túlio R. Zuquim**[2]

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

isabela.edilene@sga.pucminas.br, marco.zuquim@sga.pucminas.br.

Resumo. *O GitHub e o StackOverflow são ferramentas amplamente conhecidas e utilizadas pelos desenvolvedores na atualidade. O GitHub possui um serviço de gerenciamento de issues que contribui para a solução e discussão de problemas dos projetos ali armazenados. O objetivo deste trabalho, é descobrir se o StackOverflow também é utilizado para solução e discussão dos problemas provenientes do GitHub. A intenção é descobrir qual é o impacto das issues do GitHub no StackOverflow. Para isso, foram analisdos os 1.000 repositórios Python mais populares com no máximo 4 anos de idade do GitHub. Em seguida, foi coletado um total de 222.808 issues para, em seguida, tentar coletar perguntas do StackOverflow relacionadas a essas issues. Infelizmente, o tempo disponível não foi suficiente para a coleta das perguntas do StackOverflow relacionadas às issues do GitHub.*
Palavras-chave: *GitHub, StackOverflow, Issues, Questions*

Abstract. *GitHub and StackOverflow are tools widely known and used by developers today. GitHub has a issue management service that contributes to the solution and discussion of project issue hosted on the platform. The goal of this work is to find out if StackOverflow is also used to discuss and solve issues related to GitHub projects. The intent is to uncover what is the impact of GitHub issues on StackOverflow. In order to do it, the 1.000 most popular Python repositories, no older than 4 years old, from GitHub were analyzed. A total of 222.808 issues were collected to try to obtain StackOverflow questions related to them. Unfortunately, the available time was not enough to query the StackOverflow questions related to the GitHub issues.*
Keywords: *GitHub, StackOverflow, Issues, Questions*

1. Introdução

O GitHub e o StackOverflow são ferramentas populares na área de desenvolvimento de software. São amplamente utilizadas para apoiar as diversas situações geradas pelo desenvolvimento de software, como a necessidade de tirar dúvidas sobre programação (StackOverflow) e a hospedagem, versionamento e compartilhamento de código aberto (GitHub).

Considerando a popularidade de tais ferramentas, e o fato de que a grande maioria dos usuários de uma ferramenta é também usuário da outra, é possível que as duas ferramentas estejam relacionadas e contribuam na resolução de problemas encontrados durante o ciclo de desenvolvimento. Uma forma de se relacionar as duas ferramentas é avaliar se a discussão de *issues* de repositórios do GitHub se reflete em perguntas e repostas relacionadas no StackOverflow.

O objetivo e as questões de pesquisa propostas no presente trabalho podem ser definidos utilizando-se o método Goal Question Metrics (GQM) da seguinte forma:

1.1. Objetivo (Goal)

Analisar o impacto de *issues* de repositórios do GitHub nas discussões do StackOverflow.

1.2. Perguntas (Questions)

RQ01 Com que frequência *issues* do GitHub são discutidas no StackOverflow?
RQ02 Qual o impacto (popularidade) das discussões de *issues* do GitHub no StackOverflow?
RQ03 Existe alguma relação entre a popularidade dos repositórios e o impacto gerado?
RQ04 Os 100 primeiros repositórios com mais *issues* são também os repositórios com mais discussões no StackOverflow?
RQ05 As discussões relacionadas à *issues* do GitHub que demoram para serem fechadas, também demoram para serem encerradas no StackOverflow? O tempo que *issues* e discussões relacionadas levam para serem fechadas é similar?
RQ06 Existe uma sintonia entre os *issues* do GitHub e as discussões do StackOverflow com relação à resolução das mesmas?

1.3. Métricas (Metrics)

- Número de perguntas por *issue* (RQ01 e RQ03)
- Média do número de respostas por discussões (RQ02)
- Número de estrelas do repositório (RQ03)
- Quantidade de *issues* do repositório (RQ04)
- Quantidade de discussões associadas às *issues* do repositório (RQ04)
- Tempo que a *issue* permaneceu aberta (RQ05)
- Tempo que a discussão permaneceu aberta (RQ05)
- Relação *issues* abertas / discussões relacionadas abertas (RQ06)
- Relação *issues* abertas / discussões relacionadas fechadas (RQ06)
- Relação *issues* fechadas / discussões relacionadas abertas (RQ06)
- Relação *issues* fechadas / discussões relacionadas fechadas (RQ06)

2. Metodologia

O *dataset* definido para este trabalho é composto por 1.000 repositórios mais populares, baseando-se no número de estrelas, de projetos *open-source* escritos em Python hospedados no GitHub. Com base em uma pesquisa anterior, concluiu-se que os repositórios mais populares possuem uma média de 4 anos de idade. Assim, optou-se por restringir os repositórios, e suas *issues* e perguntas do StackOverflow consequentemente, somente naqueles que possuem até 4 anos de idade. Para as *issues*, foi definido o limite máximo de até 1000 issues por repositório.

Considerando que há muitos perguntas no StackOverflow, um filtro era necessário para reduzir o número de perguntas que seriam coletadas. Os filtros usados na busca por perguntas relacionadas às *issues* coletadas foram: idade (menor ou igual a 4 anos), *tags* (”python” e o nome do repositório do GitHub), pontuação (maior ou igual a zero), máximo de perguntas (10), ordenação (decrescente pela pontuação), e conteúdo do título ou corpo da pergunta.

2.1. Seleção de Issues do GitHub

Para a coleta dos repositórios e das suas respectivas *issues*, foi desenvolvido um script em Python que faz chamadas HTTP, usando a biblioteca "requests", à API do GitHub através da linguagem de consulta GraphQL. Essa consulta retorna as métricas, já definidas na seção 1.3, que são armazenadas em um arquivo CSV.

2.2. Identificação de Perguntas do StackOverflow

Para a coleta das perguntas do StackOverflow, foi desenvolvivo um script em Python que faz chamadas HTTP, usando a biblioteca "StackAPI".

Porém, foi identificado que existe uma cota diária de 300 consultas por endereço de IPv4 válido para as consultas à API do StackExchange. Isso se mostoru um grande obstáculo à coleta de resultados, visto que é necessário realizar uma consulta para cada uma das *issues* coletadas, ou seja 222.808 consultas. O que significa que, respeitando a definição de cotas, seriam necessários mais de 742 dias para realizar 100% das consultas de todas as *issues*.

Na tentativa de filtrar aquelas *issues* que muito provavelmente trariam resultados falsos positivos nas consultas, foi definido que só seriam consultadas aquelas *issues* com títulos que tenham pelo menos 4 palavras. Após um breve teste, foi estimado que aproximadamente 30% das *issues* estariam descartadas apenas por este filtro, mas ainda restaram muitas a serem consultadas.

Um outro artifício usado para tentar contornar o problema foi o uso de conexões privadas de VPN, para, caso o endereço IPv4 válido fosse bloqueado por estourar sua cota diária, a saída por outro proxy pudesse ajudar a resumir as consultas. Porém, a estratégia se mostrou ineficaz, visto que não havia um número suficiente de servidores disponíveis para realizar a façanha, e, no cenário em questão, não seria possível automatizar a reconexão das VPNs, tornado isso um processo manual e trabalhoso.

O script desenvolvido para coleta e análise dos projetos e geração dos arquivos CSV estão disponíveis publicamente no GitHub em github/LabEx.

3. Conclusões

Pelas razões supracitadas, não foi possível realizar a coleta de dados suficientes para a realização do estudo proposto.