PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Bacharelado em Engenharia de Software - Instituto de Ciências Exatas e Informática Unidade Educacional Praça da Liberdade

Daniel Henrique Vargas
Davi Brandão Saldanha

Relatório Final
Análise de Issues do GitHub e Perguntas do Stack Overflow

Belo Horizonte
2020

Sumário

1. Introdução

Duas grandes ferramentas utilizadas pelos desenvolvedores atualmente são o GitHub e o Stack Overflow, esse estudo busca saber se é possível relacionar o conteudo dessas duas plataformas.

Para isso, recolhemos todas as issues dos top 100 repositórios do GitHub com a linguagem principal sendo C#, escolhemos essa linguagem por ser uma linguagem popular nos dias atuais e pelos membros da equipe estarem familiarizados com ela.

Em seguida, buscamos então as top 100 perguntas no Stack Overflow que tivessem alguma relação com os repositórios do GitHub dos quais colhemos as issues. Após essa busca tentamos fazer a ligação entre as issues e perguntas coletadas.

Para alcançar este objetivo, esse trabalho visa responder às algumas questões de pesquisas, baseado na seguinte GQM:

Objetivo	Questões	Métricas
Analisar se as issues do top-100 repositórios de C# mais populares do GitHub são discutidas no Stack Overflow	Com que frequência issues do GitHub são discutidas no StackOverflow?	Total de perguntas relacionadas
	Qual o impacto das discussões de issues do GitHub no StackOverflow?	Total de repostas / total de perguntas relacionadas
	Existe alguma relação entre a popularidade dos repositórios e o buzz gerado?	Número de estrelas vs total de perguntas relacionadas
	Usuários com muita reputação costumam responder muitas questões do StackOverflow sobre issues de repositórios C# populares do GitHub?	Relação da popularidade de usuários que respondem perguntas sobre as issues dos repositórios
	As perguntas mais populares no StackOverflow são sobre as issues dos repositórios C# mais populares do GitHub?	Relação das perguntas mais populares do StackOverflow com as issues dos repositórios
	Respostas aceitas para perguntas relacionadas às issues possuem um alto número de comentários no StackOverflow?	Relação do número de comentários das respostas aceitas para perguntas relacionadas às issues no StackOverflow

2. Metodologia

O processo de coleta e análise comparativa das issues dos 100 repositórios C# mais populares com as perguntas do StackOverflow relacionadas as essas issues foi realizado em três etapas.

2.1.Coleta das Issues

A mineração dos dados foi feita por meio de uma requisição em forma de query realizada em uma API de GraphQL, disponibilizada pelo próprio GitHub, para obter informações sobre as 10 primeiras issues dos repositórios C# com mais de 100 estrelas da plataforma.

Entretanto, para poder obter os dados necessários de todos os 100 repositórios, visto que o volume de informações era muito alto para a API obter de uma vez só, foi preciso fazer o uso de recursos como paginação e um sistema para repetir a requisição caso está falhasse.

Logo após realizar a consulta das issues, os dados foram enviados para um arquivo CSV, onde foram organizados, tratados e analisados.

2.2.Coleta das Perguntas

A mineração dos dados foi feita por meio de uma requisição realizada em uma API em Python disponibilizada pelo Stack Exchange, chamada StackAPI, para obter informações sobre um total de 100 perguntas para cada um dos repositórios C# coletados na primeira etapa.

Além disso, foram coletadas as top-1000 perguntas do StackOverflow em geral para podermos avaliar a relevância das perguntas especificas coletadas anteriormente.

Entretanto, visto que o volume de informações era muito alto para a API obter de uma vez só, foi preciso fazer o uso de um sistema para repetir a requisição caso está falhasse, além de nos atentarmos ao limite de requisições da API.

Logo após realizar a consulta das perguntas, os dados foram enviados para um arquivo CSV, onde foram organizados, tratados e analisados.

2.3.Coleta das Respostas

A mineração dos dados foi feita por meio de uma requisição realizada em uma API em Python disponibilizada pelo Stack Exchange, chamada StackAPI, para obter informações sobre todas as respostas aceitas das perguntas relacionadas as issues dos repositórios C# coletadas na primeira etapa.

Entretanto, visto que o volume de informações era muito alto para a API obter de uma vez só, foi preciso fazer o uso de um sistema para repetir a requisição caso está falhasse, além de nos atentarmos ao limite de requisições da API.

Logo após realizar a consulta das respostas, os dados foram enviados para um arquivo CSV, onde foram organizados, tratados e analisados.

2.4.Hipóteses

As seguintes hipóteses foram formuladas priori a análise dos resultados:

- **Hipótese 1 –** Visto que ambas as plataformas são amplamente utilizadas por praticamente todos os desenvolvedores atualmente, esperasse que haja uma relação de no mínimo 30% dos casos estarem ligados.
- **Hipótese 2 –** Esperasse também que tenham mais respostas as perguntas do que perguntas relacionadas, visto que podem existir perguntas sem respostas, esperamos que essa situação não aconteça, é esperado que cada pergunta relacionada possua ao menos uma resposta.
- **Hipótese 3 –** Devido à existência das estrelas no Github, esperasse que os repositórios mais populares, ou seja, os que possuem mais estrelas, tenham mais perguntas relacionadas a ele. Assim ele seria mais discutido no Stack Overflow.
- **Hipótese 4 –** Como a discussão de um repositório em específico costuma ser algo mais complexo, é esperado que os usuários do Stack Overflow que responderem as perguntas, tenham uma reputação acima da média, visto se tratar de algo complexo, esperasse que alguém com experiência de soluções.
- **Hipótese 5 –** Já que ambas as ferramentas são muito usadas atualmente, é esperado que as top perguntas do Stack Overflow sejam relacionadas as issues do GitHub.
- **Hipótese 6 –** Esperasse que tenha sim um grande número de comentários nas respostas aceitas as perguntas relacionadas as issues do GitHub, visto que são problemas “populares” e acabam gerando uma discussão.

3. Apresentação e Discussão dos Resultados

Após a organização, tratamento e análise dos dados, os seguintes resultados foram obtidos e comparados com as hipóteses originalmente formuladas.

3.1.Com que frequência issues do GitHub são discutidas no StackOverflow?

Com base nos resultados obtidos, podemos perceber que cerca de 48% de um total de 3130 perguntas obtidas na coleta são válidas e possuem alguma relação com as issues coletadas dos repositórios do GitHub, representando um total de 1500 perguntas relacionadas e ultrapassado o mínimo estipulado pela nossa hipótese inicial.

3.2.Qual o impacto das discussões de issues do GitHub no StackOverflow?

Com base nos resultados obtidos, podemos perceber que, diferente do que imaginamos inicial, nem todas as perguntas relacionadas possuem resposta, contudo observamos uma relação de 1,62 respostas por pergunta relacionada. O gráfico abaixo é baseado em um total de 1500 perguntas relacionadas.

3.3.Existe alguma relação entre a popularidade dos repositórios e o buzz gerado?

Com base nos resultados obtidos, podemos perceber que a popularidade de um repositório não influencia no buzz gerado, diferentemente do que consideramos em nossa hipótese inicial, com repositórios mais populares muitas vezes tendo poucas ou nenhuma pergunta relacionada.

3.4.Usuários com muita reputação costumam responder muitas questões do StackOverflow sobre issues de repositórios C# populares do GitHub?

Com base nos resultados obtidos, podemos perceber que, diferente do que imaginamos inicial, a média de reputação dos usuários que responderam perguntas relacionadas está abaixo da média do top-100 usuários do StackOverflow em geral.

Usuário	Média
Geral	516419,3
Relacionado	26910,27

3.5.As perguntas mais populares no StackOverflow são sobre as issues dos repositórios C# mais populares do GitHub?

Com base nos resultados obtidos, podemos perceber que, diferente do que imaginamos inicial, apenas 2 perguntas relacionadas estão entre o top-1000 perguntas do StackOverflow em geral.

3.6.Respostas aceitas para perguntas relacionadas às issues possuem um alto número de comentários no StackOverflow?

Com base nos resultados obtidos, podemos perceber que, conforme a hipótese inicial, as respostas aceitas das perguntas relacionadas possuem um alto número de comentários, com

um total de 1555 comentários em 790 perguntas relacionadas que possuíam resposta aceita, chegando a uma relação de aproximadamente 1,97 comentários por resposta aceita.