PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS EXATAS E INFORMÁTICA

UNIDADE EDUCACIONAL PRAÇA DA LIBERDADE

Bacharelado em Engenharia de Software

Lucas Rotsen

Rafael Araújo Badaró

Relatório Laboratório de Medição 2

Belo Horizonte

2019

Sumário

1. Introdução
2. Métricas
3. Metodologia
4. Baseline
5. Questões e hipóteses
6. Resultados obtidos
7. Discussão sobre os resultados

1.Introdução

Este trabalho é um estudo sobre repositórios minerados do Github. Utilizando a API Graphql do Github foram mineradas métricas dos repositórios em Python do Guido van Rossum, criador da linguagem de programação Python, e dos 1000 repositórios mais populares da linguagem Python. A métricas coletadas foram analisadas em cima de uma comparação baseado em 4 perguntas sobre os repositórios.

2.Métricas

Popularidade: número de estrelas, número de watchers, número de forks

Tamanho: linhas de código (LOC)

Atividade: número de releases, frequência de releases (número de releases / dias)

Maturidade: idade (em anos)

3.Metodologia

Para alcançar os resultados obtidos, foram elaboradas consultas em GraphQL que buscam os repositórios em Python do Guido Van Rossum e os 1000 repositórios mais populares em Python. A consulta é executada através de um script em Python e o seu resultado gera um arquivo .csv com as métricas buscadas. Após essa busca, um script em shell utiliza o radon para realizar a contagem de linhas de código de cada um destes repositórios, esse script clona os repositórios, executa o radon, armazena as métricas obtidas pelo radon em um .csv e deleta o repositório clonado. As questões de pesquisa 1 e 2 serão respondidas a partir da análise quantitativa de cada uma métricas (através dos valores medianos). Para a RQ 03, os valores obtidos nas RQs 01 e 02 devem ser comparados e discutidos individualmente. Por fim, na RQ 04, os resultados apresentados na RQ 02 devem ser separados em dois grupos (*top*, com os 250 mais populares do dataset; e *bottom*, com os 250 menos populares). Em seguida, a diferença da mediana dos valores de cada uma das métricas deve ser discutidas para ambos os grupos.

4.Baseline

Características dos repositórios Python do Guido van Rossum:

Medianas das métricas

1. Número de estrelas: 35.0
2. Número de watchers: 4.5
3. Número de forks: 2.5
4. Linhas de código (LOC): 2858
5. Número de releases: 13
6. Frequência de releases (número de releases / dias): 0.0
7. Idade (em anos): 3

5.Perguntas analisadas e Hipóteses

RQ 01: Quais as características dos repositórios Python do Guido van Rossum?

Métricas: todas
Hipótese: Para essa pergunta serão listados as medianas das métricas selecionadas.

RQ 02: Quais as características dos top-1000 repositórios Python mais populares?

Métricas: todas
Hipótese: Para essa pergunta serão listados as medianas das métricas selecionadas.

RQ 03: Repositórios populares Python são de boa qualidade?

Métricas: Todas

Hipótese: Os repositórios Python de boa qualidade têm que possuir as medianas das métricas parecidas (valores próximos), iguais ou acima dos repositórios avaliados na baseline.

RQ 04: A popularidade influência nas características de repositórios Python?

Métricas: Todas

Hipótese: Os repositórios Python com mais influência devem possuir as medianas das métricas parecidas (valores próximos), iguais ou acima dos repositórios avaliados na baseline.

6.Resultados obtidos

RQ 01: Quais as características dos repositórios Python do Guido van Rossum?

Resultado: Mediana de todas as métricas

Número de estrelas	Número de watchers	Número de forks	Linhas de código (LOC)	Número de releases	Frequência de releases (número de releases / dias)	Idade (em anos)
35	4.5	2.5	2858	13	0	3

RQ 02: Quais as características dos top-1000 repositórios Python mais populares?

Resultado: Mediana de todas as métricas

Número de estrelas	Número de watchers	Número de forks	Linhas de código (LOC)	Número de releases	Frequência de releases (número de releases / dias)	Idade (em anos)
4296.5	201.0	818.5	6851.0	1	-1654.0	4

RQ 03: Repositórios populares Python são de boa qualidade?

Embora tenhamos feito a análise de proximidade proposta em nossa hipótese (entre os repositórios mais populares e a baseline), chegamos à conclusão que essa não era uma boa métrica para qualidade. Por exemplo, a mediana da nossa baseline é 35 e a dos repositórios mais populares é 4295. Dos repositórios mais populares, o mais próximo da baseline em relação ao número de estrelas também é um com um dos menores números de forks (considerado por nós uma das métricas mais significativas).

RQ 04: A popularidade influência nas características de repositórios Python?

Não conseguimos encontrar nenhuma correlação entre a popularidade de um repositório e as características deste. Foi complicado definir o que seriam “características” uma vez que as métricas coletadas não diziam tanto sobre os repositórios.

7.Discussão sobre os resultados

RQ 01: No início deste segundo trabalho da disciplina de Laboratório de Experimentação em Engenharia de Software nós acreditamos que os repositórios do Guido van Rossum fossem realmente representar os melhores estandartes de qualidade de repositórios GitHub (com relação às métricas mineradas), porém, após minerar os 1000 repositórios mais populares de Python percebemos, através das medianas das métricas calculadas, que os repositórios do Guido não representam realmente esses estandartes.

RQ 02: Como suspeitamos desde o início, o número de releases acabou não sendo muito relevante para esse trabalho, uma vez que os chamados commits tagueados ainda não são muito populares entre os desenvolvedores desses repositórios.

RQ 03: De modo geral, tivemos impressão que as métricas coletadas desses repositórios não proveram informações o suficiente para que pudéssemos fazer análises mais aprofundadas acerca da qualidade dos repositórios, porém, foi possível observar que os repositórios com o maior número de estrelas sempre coincidiram com aqueles com o maior número de watchers e forks (sendo o último considerado por nós uma das métricas mais representativas em relação ao engajamento da comunidade e qualidade).

RQ 04: Não conseguimos encontrar nenhuma correlação entre a popularidade de um repositório e as características deste. Foi complicado definir o que seriam “características” uma vez que as métricas coletadas não diziam tanto sobre os repositórios.