Uma Análise Comparativa de Repositórios Python - Relatório Final

André Sampaio Chacham
Brian Bruno Emanuel Brandao Silva

1 - Introdução:

Este trabalho foi realizado para a disciplina Laboratório de Experimentação de Software. Nele, iremos responder quatro questões de pesquisa. Para isso, foi necessário minerar e analisar os repositórios escritos na linguagem de programação Python pertencentes ao seu criador, Guido Van Rossum, e também os 1000 repositórios mais populares da língua disponíveis no Github.

Para cada questão de pesquisa, serão feitas hipóteses sobre os futuros resultados de cada uma.

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

Repositórios com bastante watchers, devido ao fato de ser do “pai do Python”. Com poucas releases, porém com bastante forks e poucas LOC. Com uma média de idade bem alta, devido ao fato do Python ser uma linguagem relativamente antiga.

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

Repositórios com muitas LOC, mas com poucas releases e com poucos forks. Com idade relativamente novas, pois python se popularizou recentemente.

RQ 03: Repositórios populares Python são de boa qualidade?

Não necessariamente. Pode acontecer de repositórios com muita popularidade mas com muitas issues, problemas e pull requests.

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

Sim, a popularidade pode influenciar a quantidade de releases, estrelas e forks de determinados repositórios. Podem possuir mais popularidade que os repositórios do “pai do python”. Esses repositórios tendem a ter mais popularidade devido ao fato de terem mais exposição na comunidade.

2 - Metodologia:

Para responder às questões, precisamos gerar arquivos CSV com todas as métricas necessárias para as repositórios do Guido (baseline) e também para os 1000 mais populares.

Para a baseline, escrevemos um script em python que consultou as métricas que queríamos por meio da API de GraphQL do Github. Porém, para a métrica de LOC foi necessário baixar os repositórios um por um, e realizar a contagem de LOC por meio da biblioteca Radon. Para os 1000 repositórios, utilizamos a mesma lógica, mas dessa vez foi realizado para os 1000 repositórios mais populares da linguagem python no Github.

O código dos dois scripts está disponível no arquivo zip da entrega.

3 - Resultados e Discussão:

RQ 01: Quais as características dos repositórios Python do Guido van Rossum?
Popularidade (medianas): Estrelas: 28.5, Watchers: 2.5, Forks: 3

Tamanho: 84

Atividade: Releases: 0, frequência de releases: 0

Maturidade (em anos): 3

É bastante divergente da hipótese. Não existem releases, a maturidade em anos é bastante pequena, 3 anos é um tempo bastante recente. O tamanho em linhas de código é pouco, conforme esperado.

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

Métrica	Mediana	Top (250)	Bottom (250)
Estrelas	780	1271	521
Watchers	52	74	37
Forks	184	308	125
Tamanho	3717	2918	3328
Releases	0	0	0
Frequência de R.	0	0	0
Maturidade (anos)	4	4	5

Maturidade relativamente nova e o tamanho relativamente alto, pois o menor tamanho é 2918 LOC. Está de acordo com o que era esperado: idade relativamente nova conforme o esperado pois o Python se popularizou recentemente. Além disso, os repositórios contam com nenhuma release.

RQ 03: Repositórios populares Python são de boa qualidade?

Sim. Eles possuem mais LOC, em média, do que os repositórios do Guido. Isso vai contra o que era esperado, que era “não necessariamente” os repositórios são de boa qualidade. Contudo, eles possuem boa qualidade sim, de acordo com as métricas analisadas.-

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

Sim, a popularidade influencia. A primeira diferença entre o TOP 250 para o BOTTOM 250 é o número de watchers, mas como os sistemas tem mais estrelas, é intuitivo. O mesmo vale para o número de forks, com uma diferença maior ainda. O número de LOC é razoavelmente similar, e o mesmo vale para os releases e sua frequência.
O que foi mais diferente do esperado foi a métrica de qualidade. Sistemas mais populares de forma intuitiva seriam aqueles que teriam mais tempo de idade, para conseguir acumular mais estrelas. Porém, como foi observado, o BOTTOM na verdade possui mediana de idade maior do que o TOP.