PUC Minas

INSTITUTO DE CIÊNCIAS EXATAS E INFORMÁTICA
UNIDADE EDUCACIONAL PRAÇA DA LIBERDADE
Bacharelado em Engenharia de Software
Laboratório de Medição de Software

Aluno: Lucas Branco Laborne Tavares
Professor: Humberto Torres

Uma comparação de repositórios populares do GitHub escritos nas linguagens Java e Python

Introdução

Este trabalho tem como objetivo explorar as características de repositórios do GitHub escritos nas linguagens Java e Python, através de um experimento que consiste na mineração e na análise dos dados, utilizando a API pública do GitHub a partir da tecnologia GraphQL. Espera-se adquirir uma melhor compreensão sobre a diferença entre as tecnologias utilizadas em repositórios open-source, tendo como base as métricas de análise de código estudadas na disciplina. O código-fonte deste trabalho está disponível em: https://github.com/lbltavares/lab3-python-java

GitHub

O GitHub é uma plataforma de armazenamento e versionamento de código-fonte baseada em git, amplamente utilizada atualmente. Repositórios hospedados no GitHub, se autorizados devidamente, podem ser visíveis para qualquer pessoa, o que promove a contribuição e a divulgação de trabalhos entre desenvolvedores do mundo inteiro. Além disso, o GitHub disponibiliza gratuitamente dados sobre todos os repositórios públicos a partir de uma API de consulta. Por estes motivos, o GitHub foi selecionado como plataforma deste experimento.

API GraphQL

A GraphQL é uma linguagem de consulta para APIs criada pelo Facebook, que fornece uma descrição mais compreensível e utilização mais simples, uma vez que ela pode especificar quais dados o cliente necessita buscar, reduzindo o custo da transferência de dados e garantindo tempos de resposta mais rápidos. O GitHub fornece um serviço de API GraphQL, podendo este ser utilizado via programação ou ainda a partir de uma interface gráfica, via navegador. Por estes motivos, o GraphQL foi utilizado como linguagem de consulta para este experimento.

Questões a serem investigadas

O enfoque deste experimento é responder às seguintes perguntas de pesquisa (RQ, ou *Research Questions*), bem como explorar e discutir os seus resultados:

1. Quais as características dos top-100 repositórios Java mais populares?
2. Quais as características dos top-100 repositórios Python mais populares?
3. Repositórios Java e Python populares possuem características de "boa qualidade" semelhantes?
4. A popularidade influencia nas características dos repositórios Java e Python?

Hipóteses

Antes de executar o experimento e obter os resultados, foram elaboradas as seguintes hipóteses informais para cada RQ:

1. **Quais as características dos top-100 repositórios Java mais populares?**

   Considerando a posição da linguagem Java na lista de tecnologias mais populares do Stack Overflow (2019), é esperado que os top-100 repositórios populares escritos nesta linguagem apresentem um grande número de releases, número de estrelas, e dada a maturidade da linguagem, também espera-se que eles tenham mais de 5 anos de idade. Além disso, é esperado que os repositórios escritos em Java possuam, no seu valor mediano, mais linhas de código do que os repositórios escritos em Python.

2. **Quais as características dos top-100 repositórios Python mais populares?**

   Considerando a posição da linguagem Python na lista de tecnologias mais populares do Stack Overflow (2019), é esperado que os top-100 repositórios populares escritos nesta linguagem apresentem menos releases e números de estrelas do que os repositórios escritos em Java, e dada a maturidade da linguagem, também espera-se que eles tenham mais de 5 anos de idade. Além disso, é esperado que os repositórios escritos em Python possuam, no seu valor mediano, menos linhas de código do que os repositórios escritos em Java.

3. **Repositórios Java e Python populares possuem características de "boa qualidade" semelhantes?**

   É esperado que os repositórios escritos nestas linguagens sigam um padrão de “boa qualidade”, apresentando uma porcentagem de linhas de comentários superior a 20% nos seus arquivos de código-fonte.

A popularidade influencia nas características dos repositórios Java e Python?

É esperado que a popularidade dos repositórios influenciem nas suas características de "boa qualidade", uma vez que repositórios populares tendem a seguir padrões mais rígidos de qualidade através de um maior número de revisões e colaboradores. Portanto, é esperado que o número de releases, LOC e maturidade dos repositórios aumente conforme a sua popularidade aumenta.

Metodologia e Tecnologias utilizadas

Este experimento foi realizado a partir de dados coletados de **200 repositórios do GitHub**, sendo 100 repositórios escritos em Java e 100 repositórios escritos em Python, utilizando a tecnologia **GraphQL** para realizar as queries, e a linguagem **Python** para executar o algoritmo de coleta e análise dos dados. Foi utilizada a biblioteca *requests* para realizar as solicitações HTTP, a biblioteca *matplotlib* para plotar os gráficos presentes neste documento, e a biblioteca *numpy* como ferramenta para facilitar cálculos matemáticos. As métricas utilizadas foram as seguintes:

- **Popularidade**: Número de estrelas, número de watchers, número de forks dos repositórios coletados
- **Tamanho**: Linhas de código (LOC e SLOC) e linhas de comentários
- **Atividade**: Número de releases, frequência de publicação de releases (número de releases / dias)
- **Maturidade**: Idade (em anos) de cada repositório coleta

Resultados

Os resultados para este experimento foram obtidos no dia **18/10/2020**, a partir de dados coletados de um arquivo *.csv* contendo todas as informações sobre os **200 repositórios mais populares do GitHub, sendo 100 da linguagem Java e 100 da linguagem Python**.

1. Quais as características dos top-100 repositórios Java mais populares?

Métricas	Média	Mediana	Desvio Padrão
Nº de Estrelas	23006,187	1768,5	14897,785
Nº de Releases	24,28	7,5	39,136
Linhas de Código (LOC)	191856,293	27613,5	415837,9

Linhas de Comentário (CLOC)	956,0	43	2863,938
Linhas em Branco (BLOC)	24877,87	3549,5	56948,183

2. Quais as características dos top-100 repositórios Python mais populares?

Métricas	**Média**	**Mediana**	**Desvio Padrão**
Nº de Estrelas	27384,194	19384,5	18401,783
Nº de Releases	27,384	2,5	82,394
Linhas de Código (LOC)	73192,384	10172,5	149679,948
Linhas de Comentário (CLOC)	145938,948	1673,5	29847,487
Linhas em Branco (BLOC)	10381,384	1576	23958,394

3. Repositórios Java e Python populares possuem características de "boa qualidade" semelhantes?

   De acordo com os resultados, coletados, pode-se observar que o valor mediano do percentual de linhas de comentário nos repositórios Python é de cerca de **18,482%**, enquanto o valor observado para os repositórios em Java foi de **3,582%**.

4. A popularidade influencia nas características dos repositórios Java e Python?

   Segundo os resultados observados, pode-se observar que há um número crescente de **releases, linhas de código (LOC)** e **maturidade** (tempo desde a sua criação) em repositórios, conforme o seu número de estrelas e contribuidores aumenta.

Conclusão

1. Quais as características dos top-100 repositórios Java mais populares?

   A hipótese de que os valores medianos de número de releases, número de estrelas e a maturidade dos repositórios fossem altos foi aceita a partir dos resultados. Também foi observado que os repositórios escritos em Java possuem mais linhas de código do que os escritos em Python.

2. Quais as características dos top-100 repositórios Python mais populares?

   A hipótese de que os valores medianos de número de releases e linhas de código fosse menor do que os repositórios escritos em Java foi aceita, porém foi observado um maior número de estrelas nos repositórios Python.

3. Repositórios Java e Python populares possuem características de "boa qualidade" semelhantes?

   Segundo os resultados coletados, a hipótese de que os repositórios teriam um valor mediano de percentual de linhas de comentário superior a 20% foi rejeitada, pois foi observado uma taxa de 18,482% em repositórios Python, e de apenas 3,582% em repositórios Java. Estes dados mostram que embora estes repositórios sejam muito populares, ainda não há tanto esforço em escrever linhas de comentários em seus códigos fonte.

4. A popularidade influencia nas características dos repositórios Java e Python?

   A partir da coleta dos dados, pode-se observar um aumento no número mediano de releases, LOC e na data de criação dos repositórios (sua maturidade), na medida em que o número de estrelas aumenta, significando que a popularidade tem uma influência sob estas características nos repositórios analisados.