PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Graduação em Engenharia de Software

Eduardo Bandeira de Melo Guimarães

RELATÓRIO DO TRABALHO PRÁTICO

Medição e Experimentação em Engenharia de Software

Belo Horizonte

09 de dezembro de 2024

Introdução

Este trabalho tem como objetivo avaliar algumas métricas dos repositórios de código aberto de três grandes empresas: Facebook, Microsoft e Google.

Para cada organização e seus repositórios de código aberto, são levantadas hipóteses a serem testadas por dados coletados referentes a esses repositórios.

Metodologia

Para cada uma das questões, foram formuladas uma hipótese nula e uma hipótese alternativa.
Então, foi escrito e executado um código em Python no ambiente de desenvolvimento "*Google Colab*" para obter dados dos diferentes repositórios de cada organização (o código de cada questão é disponibilizado através de link ao final da apresentação dos resultados de cada questão).
Após a extração dos dados e do cálculo das medidas, foi utilizado o teste estatístico "ANOVA" para testar cada uma das hipóteses.
O alfa considerado para essas análises foi de 0,10.
O motivo dessa escolha vem do fato de que este é um valor amplamente utilizado na área da Engenharia de Software.

Resultados

Questão 01: Na mediana, qual organização possui o top-10 repositórios mais antigos?

- Hipótese Nula (H_0): todas as empresas possuem a mesma mediana de data de criação de repositório, entre os 10 repositórios mais antigos.
- Hipótese Alternativa (H_a): nem todas as medianas das datas de criação dos 10 repositórios mais antigos de cada empresa são iguais.

Para responder a essa questão, foram realizadas as seguintes etapas, por meio de algoritmo:

1. Foram buscados todos os repositórios públicos disponíveis da organização usando a API do Github.
2. A lista de repositórios foi percorrida e, dela, foi criada uma lista com as datas de criação de cada repositório.

3. Essa lista de datas for ordenada em ordem crescente.
4. Então, foi determinada a mediana dos 10 primeiros elementos dessa lista, da seguinte forma:
    1. Foram selecionadas as datas centrais dos 10 primeiros elementos da lista de datas ordenada (posições 5 e 6).
    2. Foi calculado, usando o pacote “DateTime”, a data central entre essas duas datas. Essa data central se trata da quantidade de dias entre elas, dividido por 2, somado à data de posição 5.
5. Comparou-se o resultado para cada organização.

Os resultados encontrados para a mediana dos top-10 repositórios mais antigos, para cada organização, foram os seguintes:

- Facebook: **30/08/2012**
- Microsoft: **04/06/2013**
- Google: **03/08/2012**

Com esses dados em mãos, é possível rejeitar a hipótese nula, uma vez que a mediana encontrada para as 10 datas de criação mais antigas para os repositórios de cada organização são todos diferentes entre si.
Com esses mesmos dados, também é possível confirmar a hipótese alternativa. Assim, conclui-se que:

A mediana da data de criação dos 10 repositórios mais antigos de cada organização não é a mesma

E respondendo à pergunta inicial:

A organização que possui, na mediana, os 10 repositórios mais antigos é o Google.

Prints dos Resultados:

```
Analisando dados da organização: facebook
Quantidade de páginas a serem analisadas: 2
Consumindo dados da página 1
Consumindo dados da página 2
tamanho da lista: 143
['2010-01-02T01:17:06Z', '2010-03-16T18:45:15Z', '2010-05-10T17:17:33Z', '2010-06-24T22:11:
Mediana dos TOP-10 repositórios mais antigos da organização facebook: 2012-08-30 12:00:00
```

```
Consumindo dados da página 64
Consumindo dados da página 65
Consumindo dados da página 66
tamanho da lista: 6575
['2011-06-21T22:49:39Z', '2012-01-02T02:03:37Z', '2012-04-17T01:33:37Z', '2012-08-28T18:33:1
Mediana dos TOP-10 repositórios mais antigos da organização microsoft: 2013-06-14 12:00:00
```

```
Consumindo dados da página 21
Consumindo dados da página 22
Consumindo dados da página 23
Consumindo dados da página 24
Consumindo dados da página 25
Consumindo dados da página 26
Consumindo dados da página 27
tamanho da lista: 2696
['2011-06-22T18:55:12Z', '2012-01-23T17:09:03Z', '2012-01-23T17:13:56Z', '2012-04-09T20:0
Mediana dos TOP-10 repositórios mais antigos da organização google: 2012-08-03 12:00:00
```

Link do Colab:
https://colab.research.google.com/drive/1THX92QB4nlafsPHyoSPWzjWL4zoUj_dJ?usp=sharing

Questão 02: Na mediana, qual organização possui o top-10 repositórios com mais *issues* totais?

Para responder a essa questão, foram realizadas as seguintes etapas:

1. Foram buscados todos os repositórios públicos disponíveis da organização usando a API do Github.
2. A partir dessa lista, para cada repositório, foram buscados os dados detalhados de *issues*, usando dois *end-points* específicos para isso:
    1. Issues abertas e
    2. Issues fechadas.
3. A quantidade de *issues* total foi somada para cada repositório e adicionada à uma lista.
4. Ao final, a lista foi ordenada em ordem decrescente e foi calculada a mediana das 10 primeiras ocorrências desta lista.
5. Comparou-se o resultado para cada organização.

- Hipótese Nula (H_0): todas as organizações possuem a mesma mediana entre os 10 repositórios com as maiores quantidades de *issues* totais.
- Hipótese Alternativa (H_a): nem todas as organizações possuem valores iguais de mediana entre os 10 repositórios com as maiores quantidades de *issues* totais.

Os resultados encontrados para a mediana dos 10 repositórios com as maiores quantidades de *issues* totais, para cada organização, foram:
- Facebook:
- Microsoft:
- Google:

Com esses dados em mãos é possível rejeitar a hipótese nula e confirmar a hipótese alternativa. Assim, conclui-se:

A mediana do número total de *issues*, dentre os 10 repositórios com as maiores quantidades de *issues* totais, para cada organização, não é a mesma

E, respondendo à pergunta inicial:

A organização que possui, na mediana, os 10 repositórios com as maiores quantidades de *issues* totais é o …

Prints dos resultados:

Link do Colab:
https://colab.research.google.com/drive/1JwSEpDevYHJiPF_4lPDa3l6dpyyqEy27?usp=sharing

Questão 03: Na mediana, qual organização possui o top-10 repositórios com mais contribuições via *Pull-Requests*?

Para responder a essa questão, foram realizadas as seguintes etapas, por meio de algoritmo:
1. Foram buscados todos os repositórios públicos disponíveis da organização usando a API do Github.
2. A lista de repositórios foi percorrida e, dela, foi analisado o *end-point* específico para buscar as pull-requests fechadas de cada repositório.

3. Dentre as *pull-requests* fechadas, foram contabilizadas aquelas cujo valor do atributo "*merged_at*" era diferente de nulo, ou seja, aquela que foi aceita e mesclada à *branch* principal.
4. A quantidade de contribuições foi adicionada à uma lista, que depois foi ordenada em ordem decrescente.
5. Foi calculada a mediana dos 10 primeiros valores dessa lista.
6. Comparou-se o resultado para cada organização.

- Hipótese Nula (H_0): todas as empresas possuem a mesma mediana de contribuições entre os 10 repositórios com as maiores quantidades de contribuições
- Hipótese Alternativa (H_a): a mediana de contribuições entre os 10 repositórios com as maiores quantidades de contribuições, para cada organização, não é a mesma para todas elas.

Os resultados encontrados para as medianas de contribuições dos 10 repositórios com as maiores quantidades de contribuições, para cada organização, foram:
- Facebook: 2109,5
- Microsoft: 1699,5
- Google: 4412

Com esses resultados em mãos, foi utilizada a técnica estatística "ANOVA" para testar as hipóteses. Os resultados encontrados foram os seguintes:

```
from scipy.stats import f_oneway


facebook = [110993, 5188, 3417, 2847, 2285, 1934, 1197, 875, 603, 392]
microsoft = [9022, 6934, 6823, 6635, 5024, 3824, 3512, 3321, 2084, 1315]
google = [9078, 5905, 4796, 4762, 4584, 4240, 4170, 4143, 4012, 3961]

f_oneway(facebook, microsoft, google)
```

```
F_onewayResult(statistic=0.5437053063042034, pvalue=0.5868171821474453)
```

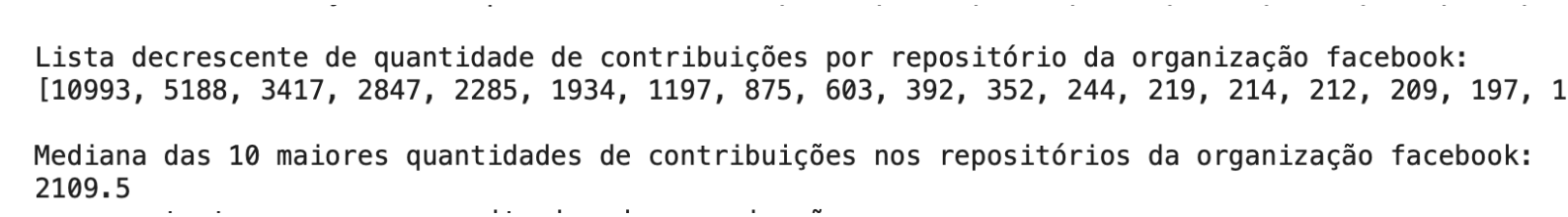
Assim, como o p-value é maior que o alfa, não é possível rejeitar a hipótese nula e aceitar a hipótese alternativa. Dessa forma, conclui-se:

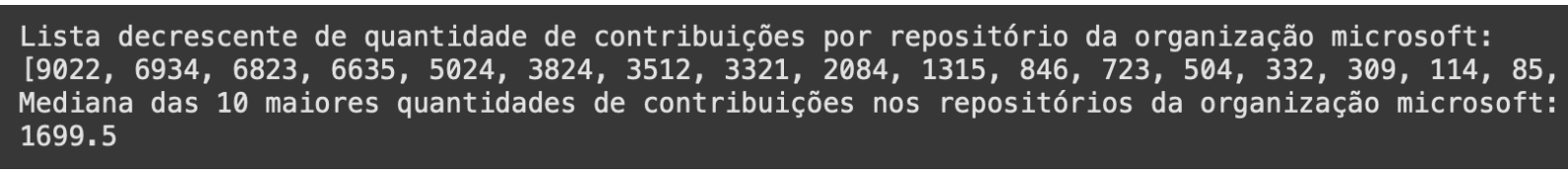
Para cada organização, a mediana da quantidade de contribuições dos 10 repositórios com as maiores quantidades de contribuições não é a mesma entre elas

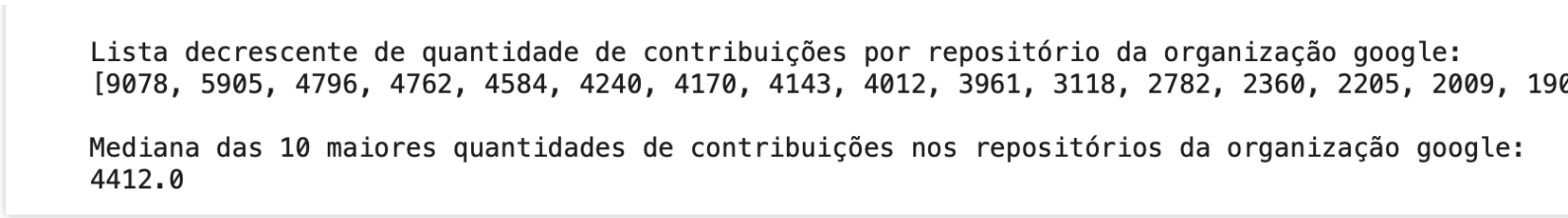
E, respondendo à pergunta inicial:

A organização que apresentou, na mediana, o top-10 repositórios com mais contribuições via Pull-Requests foi o Google

Prints dos Resultados:







Link do Colab:
https://colab.research.google.com/drive/1Vyvtct1himDemHFX4LHW_7qpFSdLm2Ge?usp=sharing

Questão 04: Na mediana, qual organização possui o top-10 repositórios que lança releases com mais frequência?

Para responder a essa questão, foram realizadas as seguintes etapas, por meio de algoritmo:

1. Foram buscados todos os repositórios públicos disponíveis da organização usando a API do Github.
2. A lista de repositórios foi percorrida e, dela, foi analisado o *end-point* específico para buscar as *releases* de cada repositório.
3. Então, foram percorridas todas as páginas de detalhamento das releases, contando a quantidade de ocorrências da lista de cada página.

4. Os números de releases de cada repositório foram adicionados a uma lista, que foi então ordenada decrescentemente.
5. Foi calculada a mediana entre as dez primeiras ocorrências desta lista.
6. Comparou-se o resultado para cada organização.

- Hipótese Nula (H_0): todas as empresas possuem a mesma mediana de *releases* entre os 10 repositórios com as maiores quantidades de *releases*
- Hipótese Alternativa (H_a): nem todas as empresas analisadas possuem medianas iguais para os 10 repositórios com as maiores quantidades de *releases.*

Os resultados encontrados para as medianas de *releases* dos 10 repositórios com as maiores quantidades de *releases*, para cada organização, foram:
- Facebook: **220**
- Microsoft: **601**
- Google: **216**

Com esses resultados em mãos, foi utilizada a técnica “ANOVA” para testar as hipóteses. Os resultados encontrados foram:

```
from scipy.stats import f_oneway


facebook = [379, 355, 332, 221, 220, 220, 218, 214, 198, 148]
microsoft = [117503, 1148, 773, 683, 610, 592, 591, 554, 544, 524]
google = [12291, 2045, 256, 232, 226, 206, 193, 174, 165, 154]

f_oneway(facebook, microsoft, google)
```

```
F_onewayResult(statistic=0.9568183627634427, pvalue=0.39675653938959765)
```

Com esses resultados, não é possível rejeitar a hipótese nula e aceitar a hipótese alternativa.

E, respondendo à pergunta inicial:

Na mediana, a Microsoft é a organização que possui o top-10 repositórios que lança *releases* com maior frequência

Prints dos resultados:

```
Lista decrescente da quantidade de releases para cada repositório da organização facebook
[379, 355, 332, 221, 220, 220, 218, 214, 198, 148, 132, 129, 103, 86, 83, 67, 57, 57, 54, 54,
Mediana da quantidade de releases dos top-10 repositórios com maior frequência de releases:
220.0
```

```
Lista decrescente da quantidade de releases para cada repositório da organização microsoft
[17503, 1148, 773, 683, 610, 592, 591, 554, 544, 524, 513, 505, 472, 471, 446, 439, 425, 409,
Mediana da quantidade de releases dos top-10 repositórios com maior frequência de releases:
601.0
```

```
Lista decrescente da quantidade de releases para cada repositório da organização google
[2291, 2045, 256, 232, 226, 206, 193, 174, 165, 154, 148, 140, 131, 120, 116, 113, 110, 107,
Mediana da quantidade de releases dos top-10 repositórios com maior frequência de releases:
216.0
```

Link do Colab:
https://colab.research.google.com/drive/17UpNufAs0wrgD9Gqpk2eh8zKYk7oPQ9a?usp=sharing

Questão 05: Na mediana, qual organização possui o top-10 repositórios atualizados com mais frequência?

Para responder a essa questão, foram realizadas as seguintes etapas, por meio de algoritmo:

1. Foram buscados todos os repositórios públicos disponíveis da organização usando a API do Github.
2. A lista de repositórios foi percorrida e, a partir dela, foi analisado o *end-point* específico para buscar os *commits* de cada repositório.
3. Então, como o *end-point* fornece o histórico de commits em ordem crescente de proximidade de datas (mais recentes primeiro), bastou analisar a data do primeiro valor da lista.
4. Então, foi calculado o tempo percorrido entre a última atualização do repositório e o horário de início da execução do algoritmo.
5. Os valores desses tempos foram colocados em uma lista, que foi então ordenada crescentemente.
6. Foi calculada, finalmente, a mediana entre os 10 primeiros valores de tempo dessa lista.
7. Comparou-se o resultado para cada organização.

- Hipótese Nula (H_0): todas as empresas possuem a mesma mediana de tempos entre os 10 repositórios com os menores tempos entre a última atualização e a data/hora do momento da verificação.
- Hipótese Alternativa (H_a): as medianas de tempos entre a última atualização de um repositório e a data/hora de verificação, entre os 10

repositórios mais recentemente atualizados não é a mesma para todas as organizações.

Os resultados encontrados para as medianas de tempo entre a última atualização e o momento da verificação, dos 10 repositórios com os menores tempos entre a última atualização e o momento da verificação, para cada organização, foram:

- Facebook: **3872 s**
- Microsoft: **1401 s**
- Google: **2085 s**

Com as listas dos top-10 repositórios atualizados mais frequentemente (menores tempos entre momento de análise e última atualização), foi utilizada a técnica “ANOVA”. Os resultados obtidos foram os seguintes:

```
from scipy.stats import f_oneway


facebook = [876, 1915, 1957, 2043, 3317, 4427, 4737, 5284, 5707, 5848]
microsoft = [78, 91, 642, 1154, 1328, 1474, 1928, 2234, 2782, 2923]
google = [151, 535, 683, 883, 1828, 2342, 3584, 5722, 6104, 8782]

print(f_oneway(facebook, microsoft, google))
```

```
F_onewayResult(statistic=2.905568575768201, pvalue=0.07196536069073116)
```

Com esses resultados em mãos, é possível rejeitar a hipótese nula e aceitar a hipótese alternativa (p-value < alfa). Dessa forma, conclui-se que:

As empresas apresentam medianas diferentes dentre os 10 repositórios com os menores intervalos de tempo entre a última atualização e o momento da verificação

E, respondendo à pergunta inicial:

Na mediana, a Microsoft é a organização que possui o top-10 repositórios atualizados com maior frequência

Link do Colab:

https://colab.research.google.com/drive/1JTDQupG_NwwLLSBRf_gKRhSbkfKqDg3S?usp=sharing

Link do Colab de cálculo do “ANOVA”:

https://colab.research.google.com/drive/1idXbXttNrruuFh6axhvspXhgzxtDq9In?usp=sharing