Laboratório 4 – GitHub Issues e Stack Overflow Q&As

- Uso do GitHub Issue Tracker para analisar Questão e Repostas do Stack Overflow
- Prática de Laboratório

Introdução (Prof. Laerte Xavier, 2020)

- GitHub e Stack Overflow são plataformas amplamente utilizadas no desenvolvimento de software moderno.
- Apesar de possuírem objetivos distintos (a primeira preocupa-se com hospedagem, compartilhamento de código aberto e colaboração, enquanto a segunda destina-se à discussão de dúvidas de programação), ambas são frequentemente adotadas de maneira complementar pelos desenvolvedores.
- Entre os serviços mais populares disponibilizados pelo GitHub, o gerenciamento de *issues* através do seu *issue tracker* apresenta-se como ferramenta eficaz na solução e discussão de problemas mais comuns.

Objetivo

- O objetivo deste laboratório é avaliar se a discussão de *issues* nos repositórios mais populares do GitHub se refletem em perguntas e respostas relacionadas no Stack Overflow (Prof. Laerte Xavier, 2020).

“Buzz” de Issues do GitHub no Stack Overflow

- Material do Prof. Laerte Xavier (PUC Minas)
- Disponível em: https://bit.ly/3lt1bGn (último acesso em 09/10/2020)

- State-of-art: text similarity computing: https://dl.acm.org/doi/10.1145/3290420.3290473

Stack Overflow API

- Material elaborado por Aline Brito e André Hora do (ASERG – DCC/UFMG)
- Disponível em: https://bit.ly/34G6jQH (último acesso em 09/10/2020)

Metodologia (1/4) – Seleção de GitHub Issues

- Para formação do *dataset* de *issues* a serem avaliadas neste estudo, você deverá, inicialmente, definir o conjunto de repositórios que pretende estudar, segundo algum critério que achar mais interessante. Por exemplo: os mil repositórios mais populares do GitHub, ou os top-10 mais populares de algumas linguagens de programação específicas.
- Lembre-se que a definição de um *dataset* deve amenizar ameaças à validade relacionadas à generalização dos seus resultados. Neste caso, escolha um subconjunto de repositórios que seja interessante para o estudo e minere as suas *issues* a partir da API GraphQL do GitHub.

Metodologia (2/4) – Identificação de SO Posts

- A partir do conjunto de *issues* selecionadas, é necessário que identifiquemos a ocorrência da discussão dessas *issues* em posts do Stack Overflow. Para tanto, você pode escolher entre duas técnicas sugeridas (ou propor uma nova com base no cálculo de similaridade de textos curtos):
  1. Identificar no título das *issues* palavras chave específicas, que indiquem qual parte do código está sendo discutido (por exemplo, a partir de expressões regulares que incluam o padrão *camelCase*). Para cada módulo identificado nas *issues*, buscar no *dump* do Stack Overflow perguntas que possuam referência a eles, numa janela de tempo próxima à criação da *issue*.

Metodologia (2/4) – Identificação de SO Posts

(continuação)

2. Buscar pelo código das *issues* (string: "Issue #CODIGO") no *dump* do Stack Overflow. Neste caso, adicione também o nome e o *onwer* do repositório na sua consulta, para garantir que a *issue* pesquisada é referente ao repositório analisado.

Atenção: No relatório final, você deve indicar e justificar a abordagem escolhida. Apresente os resultados obtidos com esta mineração.

Metodologia (3/4) – Questões de Pesquisa

Inicialmente, este laboratório tem o objetivo de responder:

- **RQ 01:** Com que frequência *issues* do GitHub são discutidas no Stack Overflow?
- **RQ 02:** Qual o impacto das discussões de *issues* do GitHub no Stack Overflow?
- **RQ 03:** Existe alguma relação entre a popularidade dos repositórios e o *buzz* gerado?

Atenção: três questões de pesquisa adicionais devem ser adicionadas pelo grupo (Sprint 01).

Metodologia (4/4) – Definição de Métricas

Para cada questão de pesquisa, as seguintes métricas serão calculadas:

- **RQ 01:** total de perguntas relacionadas *(PostType: Question)*
- **RQ 02:** total de repostas / total de perguntas relacionadas *(PostTypes: Answer / Question)*
- **RQ 03:** número de estrelas vs total de perguntas relacionadas

Atenção: defina as métricas necessárias para responder às questões de pesquisa definidas pelo grupo e discuta no relatório final.

Ameaças à Validade

- De acordo com a metodologia escolhida e a estratégia adotada para selecionar os Posts no SO, discuta as ameaças a validade do seu trabalho, dividindo-as em validade de construção, interna e externa. Discuta cada uma delas e aponte abordagens que você adotou (ou poderia ter adotado) para mitigar essas ameaças.

Relatório Final
(proposta do Prof. Laerte Xavier)

- Elabore um documento que apresente:
  i. uma introdução simples com hipóteses informais;
  ii. as questões de pesquisa investigadas;
  iii. a metodologia que você utilizou para respondê-las;
  iv. os resultados obtidos;
  v. a discussão sobre o que você esperava como resultado (suas hipóteses) e os valores obtidos;
  vi. as ameaças à validade do seu estudo.

Sprints de Trabalho em Dupla (1/4)

- **Sprint 01**: *criação das três questões de pesquisa adicionais e definição do dataset (documento de texto descrevendo as RQs escolhidas, as métricas associadas e o dataset a ser analisado).*
  - Valor: 3 pontos
  - Entrega em 21/10/2020 até às 23:59 no Canvas e no SGA

Sprints de Trabalho em Dupla (2/4)

- **Sprint 02**: *Coleta de issues (arquivo .csv contendo as issues do dataset definido)*
  - Valor: 4 pontos
  - Entrega em 28/10/2020 até às 23:59 no Canvas e no SGA

Sprints de Trabalho em Dupla (3/4)

- **Sprint 03**: *Mineração do Stack Overflow (consulta dos posts no Stack Overflow + valores das métricas do estudo)*
  - Valor: 8 pontos
  - Entrega em 04/11/2020 até às 23:59 no Canvas e no SGA

Sprints de Trabalho em Dupla (4/4)

- **Sprint Final**: *Análise de dados + elaboração do relatório final*
  - Valor: 10 pontos
  - Entrega em 11/11/2020 até às 23:59 no Canvas e no SGA

OBS.: Desconto de 0,5 por dia de atraso de entrega