SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DA FRONTEIRA SUL
DIRETORIA DE PESQUISA
Avenida Fernando Machado, 108-E, Centro, Chapecó-SC, CEP 89802-112, 49 2049-3748 49 20493743
dir.dpe@uffs.edu.br www.uffs.edu.br/pesquisa

"DPE – F05" - Formulário para entrega de Relatório de atividades[1"2]

01 – Dados do projeto "guarda-chuva" ou subprojeto

Número do Registro no sistema Prisma: PES-2020-0256
Orientador: Guilherme Dal Bianco
Coorientador: (para editais de Ensino Médio):
E-mail do coorientador:

02 – Dados [X] Bolsista [ ] Voluntário

Nome: MATHEUS VINICIUS TODESCATO
Campus: Chapecó
Período de execução (mês/ano): 29/09/2020 à 31/12/2020
Horas semanais (10h a 30h): 20
Horas totais (apenas para voluntários de projetos "guarda-chuva"):

03 – Participação em eventos e atividades do projeto

Nº de encontros com o orientador realizados no período 16
Nº de encontros do grupo de pesquisa de que participou no período 2
Participação em eventos no período:_1 Tipo de participação[X] ouvinte [ ] apresentador

4 – Detalhamento das atividades

4.1 Descrição das atividades e suas contribuições para a pesquisa

O projeto tem como objetivo explorar a geração de características para identificar trechos vulneráveis no código fonte e desenvolver abordagens para reduzir o esforço de rotulação do usuário para a execução do modelo de previsão. Para isto, a tarefa desenvolvida durante a execução da bolsa foi estudar a aplicabilidade dos principais métodos de identificação de documentos relevantes com menor esforço do usuário. Dessa forma, os seguintes métodos foram estudados:

- Autotar: é um dos métodos mais simples para o problema TAR e sua estratégia é explorada em diversos outros métodos. Seu diferencial é ser totalmente autônomo, não requerendo ajuste de parâmetros por conta de tópicos ou bases de dados específicas. No

[1] Este formulário deve ser entregue por: *BOLSISTAS:* na metade do período de vigência ou em caso de substituições para comprovação das atividades realizadas até a substituição > *VOLUNTÁRIOS:* como condição para certificação

[2] Para agilizar o fluxo de trabalho, ao inserir o documento no sistema Prisma, por favor informar à CAPPG do campus via e-mail.

entanto, o método tem como ponto fraco a alta demanda de documentos rotulados pelo usuário.

- SCAL: é proposto uma melhoria no AutoTAR para aprimorar a escalabilidade em grandes conjuntos de dados. O objetivo do S-CAL é oferecer uma alta revocação em grandes bases de dados com reduzido esforço do usuário. O trabalho tem como foco bases de dados voltadas para *Eletronic Discovery* que nada mais é do que a busca por documentos relevantes para processos judiciais. Para se atingir tal objetivo, essa nova versão traz mudanças no funcionamento do *AutoTAR*. A principal diferença é que são rotuladas apenas pequenas amostras de cada lote sucessivo de documentos e o processo se faz escalar até que se esgote a coleção. Com isso se reduz o esforço do usuário, não sendo mais necessário rotular todo o lote de documentos a cada rodada.
- Fast: foi proposto um método baseado em aprendizado ativo, explorando técnicas para a construção de semente. O diferencial deste trabalho é que apresenta novas formas de abordar os desafios relacionados à geração de treinamento inicial, o erro humano no processo de revisão e o ponto de parada para o método. O FAST2 traz uma abordagem baseada em aprendizado ativo com utilização de métodos para construção da semente inicial e semelhante ao AutoTar, utiliza uma abordagem incremental na qual a cada ciclo um lote de documentos é avaliado. Tal lote é treinado utilizando o algoritmo SVM, empregando uma estratégia de predição de erros humanos e um estimador de revocação.
- WEST: tem como foco a geração de semente inicial para a construção do treinamento. A abordagem tem como base dois módulos principais: (A) um gerador de pseudo-documento que leva em consideração a informação da semente para pré-treinar uma rede neural;(B) um módulo de autotreinamento com documentos reais não rotulados utilizando a rede treinada pelos pseudo-documentos. Nos experimentos realizados foi identificado que a abordagem tem uma performance significativamente melhor do que os métodos bases (TF-IDF, LDA e etc). No entanto, um problema é a falta de integração entre as informações diferentes das sementes. Se isso for resolvido pode impulsionar ainda mais os resultados positivos.

A partir do estudo dos métodos presentes no estado-da-arte, iniciou-se o processo de busca por uma coleção de dados de teste para avaliar o funcionamento dos métodos acima citados. A coleção selecionada contém 15 projetos do Github que estão disponíveis publicamente [1]. A base contém 2 tipos de arquivo: o código fonte puro e o código com seus atributos

extraídos. Para o nosso trabalho foi utilizado o arquivo com os atributos e dentro dele já se encontrava a informação de quantos bugs existia em cada código.

A base então foi pré-processada para ser possível a execução nos métodos citados e testes iniciais foram executados. No entanto, os resultados finais ainda dependem de execuções sistemáticas e padronização de parâmetros para ser possível a coleta de valores confiáveis que apresentem as diferenças entre os métodos.

[1] Tóth, Z. et al. "A Public Bug Database of GitHub Projects and Its Application in Bug Prediction." *ICCSA* (2016).

4.2 Resultados das atividades desenvolvidas

Devido ao curto período da bolsa, às atividades desenvolvidas envolveram um bibliográfico dos métodos KNEE, FAST, SCAL e Autotar. Além da seleção da base de dados para realização de testes iniciais. Atualmente, como voluntário do projeto, estou desenvolvendo a adequação dos métodos para execução com a base de dados e a realização dos experimentos iniciais.

4.3 Autoavaliação do estudante

O trabalho foi produtivo mesmo com o pequeno tempo de bolsa. Consegui investigar bibliograficamente os métodos que poderia aplicar no problema e lidar com uma base de dados de código fonte e suas características. Trabalhar neste projeto me agregou muito conhecimento e melhoria no desenvolvimento de uma pesquisa. Pretendo continuar trabalhando no tópico e produzir mais material científico do mesmo.

4.4 Avaliação do orientador sobre o desempenho do estudante

O estudante desenvolveu todas as atividades conforme descrito no cronograma do projeto. Além disso, o estudante se mostra proativo e motivado para o desenvolvimento das tarefas.

Local e data: Chapecó, 01/02/2021