Classificação de Issues do Github com Relação a Segurança

1st Bruno Gonçalves de Oliveira
Universidade Federal do Paraná (UFPR)
Curitiba– PR – Brasil
bruno.mphx2@gmail.com

2nd Diogo Cezar Teixeira Batista
Universidade Federal do Paraná (UFPR)
Curitiba– PR – Brasil
diogocezar@utfpr.br

***Abstract*—Para a matéria de Aprendizagem de Máquina foi solicitado um problema que poderia ser resolvido através da técnica. Entre as características da aprendizagem de máquina, está a eficácia em classificação. O trabalho tem a meta de classificar issues descendentes de projetos do Github e classificalas se são referentes a segurança ou não. As issues serão analisadas por suas discussões, tirando vantagem de palavras frequentes em textos referentes à segurança. O trabalho utilizou da técnica de Bag-of-Words aliada com o classificador Naive Bayes para detenção da melhor eficácia em classificação.**

***Index Terms*—issues, github, segurança, classificação**

I. Introdução

O controle, gerenciamento e manutenção de arquivos, especialmente no âmbito do desenvolvimento de software, sempre foi um desafio. Problemas recorrentes como: backups não realizados, sobrescritas de arquivos, dificil manutenabilidade de projetos desenvolvidos em equipes, são motivações para a utilização de algum sistema de versionamento de arquivos. [1]

Dentre as várias ferramentas existenstes no mercado, como por exemplo: CVS, Subversion, TFS, Mercurial, o Git se destaca por ter sido amplamente utilizado pela comunidade de desenvolvimento, com o advento da popularização dos projetos OpenSource. Estes projetos se consolidaram em com ferramentas como o GitHub que é uma plataforma para versionamento, gerenciamento e colaboração de projetos, que utiliza o Git como base.

São várias as possibilidades que essas ferramentas proporcionam para verionamento dos projetos, mas para o contexto deste trabalho, destaca-se uma muito importante, a mensagem enviada ao realizar uma modificação na estrutrua de arquivos. Essa ação é denominada como commit, e cada commit possui um label atrelado a ele. É através desta mensagem que um desenvolvedor pode descrever quais foram as alterações que ele realizou.

Eventualmente, dentre os ajustes realizados neste projetos, são realizados ajustes relacionados a segurança. Esses ajustes são críticos e precisam ser analisados por outras pessoas, mesmo após já terem sidos enviados.

Mas como identificar quais são os ajustes de um projeto que estão relacionados com segurança? Como separar estes ajustes para que especialistas possam analisar o cógidos? Essas perguntas são a motivação para o desenvolvimento deste trabalho.

A proposta deste trabalho é a criação de uma ferramenta que utilize técnicas de aprendizagem de máquina para a criação de um oráculo classificador que consiga analisar as palavras contidas em uma mensagem de commit, e classificar se este commit está ou não relacionado à uma implementação de segurança.

II. Problema Proposto

Durante um estudo de requisitos não-funcionais se deu a necessidade de classificação de issues para o desenvolvimento do projeto.

Essa classificação contrariamente ao que é realizado normalmente, deveria ser realizado através dos textos das issues e sem levar em consideração o código-fonte. A classificação se dará por relacionados a segurança ou não relacionados.

III. Base de Dados

Para a base de dados de treinamento e testes, duzentas issues foram enumeradas e classificadas dos projetos OKHTTP, jgit e couchbase. Sendo cem issues relacionadas a segurança e cem não relacionadas. Estas mensagens foram manualmente classificadas e arquivadas como arquivo CSV, assim como apresentado na Figura 1.

A cada linha se tem a classificação e então o texto determinante da classificação recebida, security ou not.

IV. Experimentos

Para o início dos experimentos envolvendo os dados propostos, o uso da aprendizagem de máquina se faz necessário algumas etapas essenciais. Nas próximas seções, estas etapas serão detalhadas.

A. Conversão de Textos

Nesta parte do processo é utilizada a técnica conhecida como Bag-of-Words, as sentenças serão representadas através da identificação de suas palavras e a quantidade que aparecem. Por exemplo: My dog's name is Rex. Na Figura 2 é feita a identificação e relação de cada palavra da frase.

Na Figura 3, é possível observar a relação de identificação da palavra e a quantidade.

Após identificadas todas as palavras das sentenças, se tem a quantidade que as palavras se apresentam nos textos, o que será útil para a classificação.

B. Sanitização da Entrada

Para identificação de palavras-chave no contexto descoberto, é necessário remover palavras que não representam o contexto das frases ou ainda palavras semelhantes que se representam da mesma forma. Algumas das técnicas aplicadas:

- Transformar todo texto em minúsculo, assim não terá variações por casos de maiúsculo e minúsculo;
- Ignorar pontuações;
- Corrigir palavras com ortografia incorreta;
- Remover as chamadas stop words que não acrescentam informação aos textos, por exemplo: of, a, in, on, etc;

C. Importância da Palavra

Com os vetores de palavras formados com a identificação e a quantidade, é possível aplicar a técnica de TF-IDF (term frequency-inverce document frequency). A técnica pretende refletir quão importante uma palavra é ao documento. O valor de tf-idf aumenta proporcionalmente conforme o número de vezes a palavra aparece no texto e é compensado pelo número de textos na base de dados, ajustando o número da frequência das palavras.

V. Resultados

VI. Conclusão

[**?**].

References

[1] S. Chacon and B. Straub, *Pro Git*. Apress, 2020.