Classificação de *issues* do Github relacionadas a Segurança

Aprendizagem de Máquina

Bruno Gonçalves de Oliveira

bruno.mphx2@gmail.com

Diogo Cezar Teixeira Batista

diogocezar@utfpr.br

Universidade Federal do Paraná - UFPR

Curitiba - 2020

Agenda

1. Introdução
2. Obtenção dos Dados
3. Pré-processamento
4. Extração de Características
5. Resultados
6. Conclusão

Introdução

- Gerenciamento e manutenção de arquivos: desafio;
  - backups não realizados;
  - sobrescrita de arquivos;
  - difícil manutenabilidade em times;
- Diferentes soluções no mercado: *CVS*, *Subversion*, *TFS*, *Mercurial*.

GitHub

- *Git* + *GitHub* = *OpenSource*.
- *GitHub* que é uma plataforma para versionamento, gerenciamento e colaboração de projetos, que utiliza o *Git* como base. [CS20]

Controle de Issues

- Dentre outras ferramentas temos o controle de *Issues:*.
  - documentar possíveis *bugs*, melhorias, ou novas *features* para os projetos

Exemplo Issues

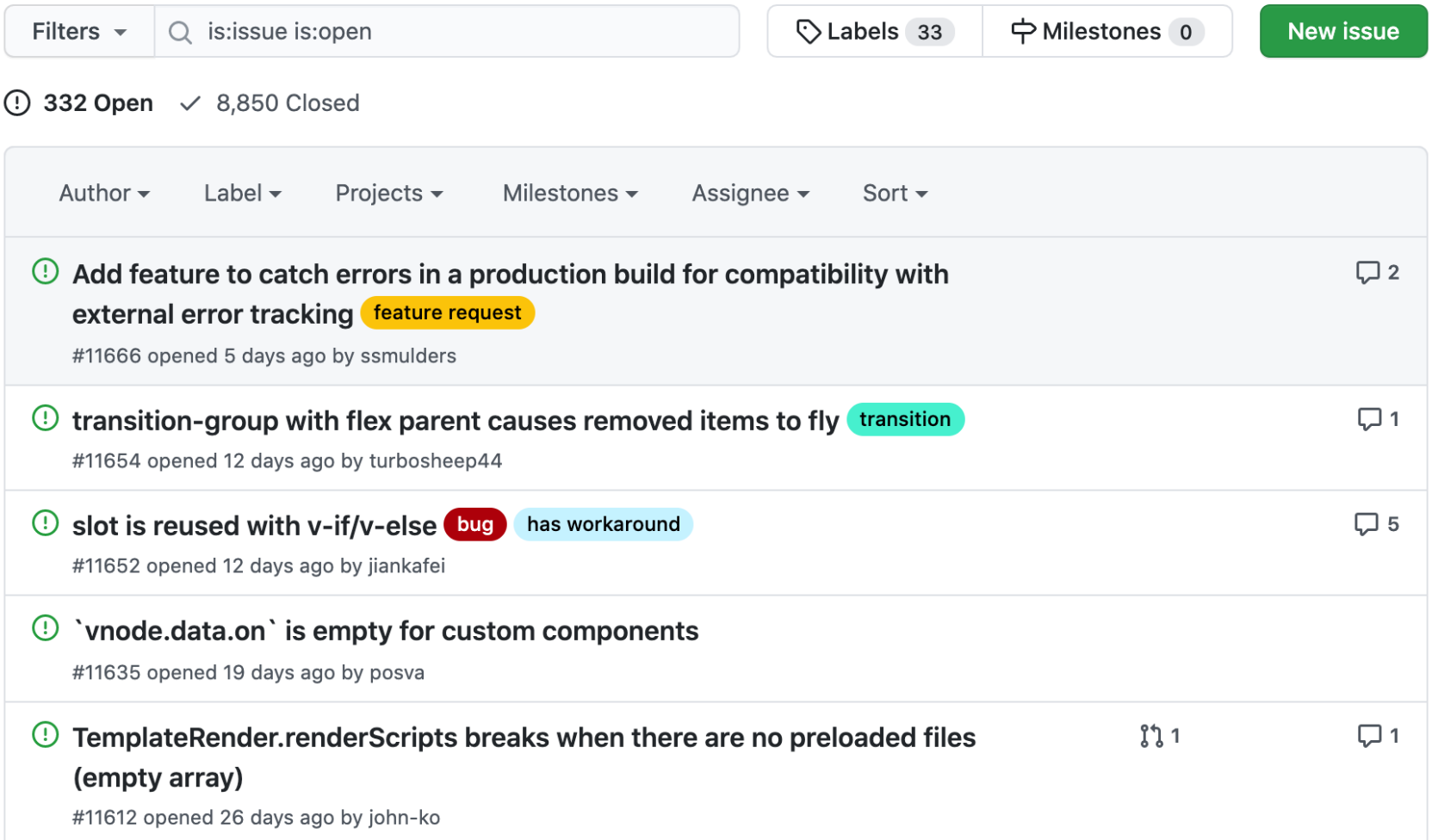

Figura: Exemplos de Issues do Projeto Vue.js

Issues sobre Segurança

- Eventualmente, as *issues* podem estar relacionadas a tópicos de segurança.
- Quando consideradas críticas, podem ser analisadas por outros especialistas;
- Como identificar quais *issues* que são relacionadas com segurança?
- Como classificar estas *issues* para que especialistas possam analisar os cógidos?

Proposta do Trabalho

- Criação de uma ferramenta que utilize técnicas de **aprendizagem de máquina** para o desenvolvimento de um classificador que consiga analisar as palavras contidas nas mensagens das *issues* de um dado projeto, e classificar se esta *issue* está ou não relacionada no contexto de segurança da informação.

Obtenção dos Dados

- Utilizou-se o *github-csv-tools*[1] que possibilita a exportação dos dados de um repositório do *GitHub*, salvando as informações em um arquivo no formato CSV.
- Dados tratados para um CSV com 2 colunas:

```
security,PushObserver can be used to push ↩
    serverinitiated HTTP/2 requests into an ↩
    OkResponseCache...
not,Handle LOCKED in conversions.Motivation...
```

Código 1: CSV Exemplo com Base de Dados

[1] https://github.com/gavinr/github-csv-tools

Fonte de Dados

- Base de Testes: *issues* do projeto *Wildfly*[2];
- Base de Treinamento: *issues* dos projetos: *okhttp*[3], *jgit*[4] e *couchbase*[5]
- Os dados de treinamento possuem **199** entradas, enquanto que para a base de teste foram utilizadas **211** entradas.

[2] https://github.com/wildfly/wildfly
[3] https://github.com/square/okhttp
[4] https://github.com/eclipse/jgit
[5] https://github.com/couchbase

Pré-processamento

- Completar...

Extração de Características

- Completar...

Resultados

- Completar...

Conclusão

- Completar...

Trabalhos Futuros

- Completar...

References I

 Scott Chacon and Ben Straub, *Pro git*, Apress, 2020.