

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

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

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

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

Exemplo Issues

Filters ▾

🔍 is:issue is:open

🏷 Labels 33

📅 Milestones 0

New issue

🔔 332 Open ✓ 8,850 Closed

Author ▾	Label ▾	Projects ▾	Milestones ▾	Assignee ▾	Sort ▾
🔔 Add feature to catch errors in a production build for compatibility with external error tracking	feature request				💬 2
#11666 opened 5 days ago by ssmulders					
🔔 transition-group with flex parent causes removed items to fly	transition				💬 1
#11654 opened 12 days ago by turbosheep44					
🔔 slot is reused with v-if/v-else	bug	has workaround			💬 5
#11652 opened 12 days ago by jiankafei					
🔔 `vnode.data.on` is empty for custom components					
#11635 opened 19 days ago by posva					
🔔 TemplateRender.renderScripts breaks when there are no preloaded files (empty array)				🔗 1	💬 1
#11612 opened 26 days ago by john-ko					

Figura: Exemplos de Issues do Projeto Vue.js

- 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 códigos?

- 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>

- 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 (Regas Aplicadas)

- Transformar todo texto em 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*.

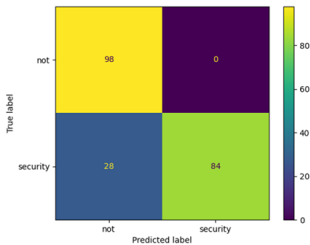
- Foram aplicadas as técnicas:
 - Bag-of-Words;
 - TF-IDF (term frequency-inverse document frequency);
- Palavras mais relevantes: ['security', 'secure', 'vulnerable', 'leak', 'exception', 'crash', 'malicious', 'sensitive', 'user', 'authentication', 'protect', 'vulnerability', 'authenticator', 'auth', 'npe']

Classifier	Accuracy	F1Score	Time (s)
LinearDiscriminantAnalysis	0.867	0.865	0.183
LogisticRegression	0.867	0.865	0.191
DecisionTreeClassifier	0.867	0.865	0.193
MLPClassifier	0.867	0.865	0.302
svm.LinearSVC	0.867	0.865	1.903
Perceptron	0.862	0.861	0.17
KNeighborsClassifier	0.533	0.696	0.229
GaussianNB	0.471	0.307	0.192

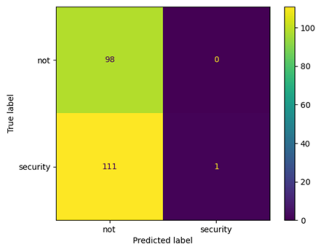
Tabela: Resultados dos Experimentos

Matrizes de Confusão

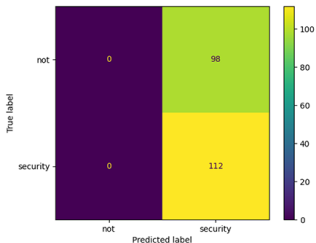
DecisionTreeClassifier



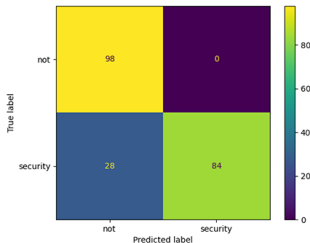
GaussianNB



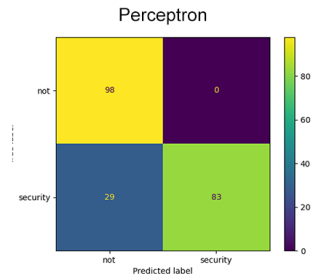
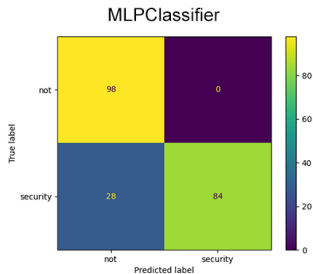
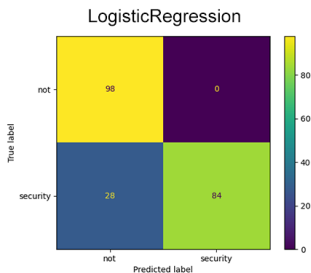
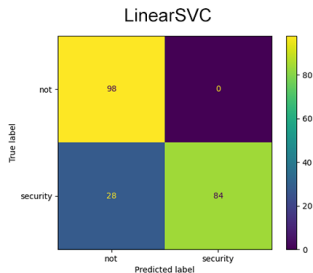
KNeighborsClassifier



LinearDiscriminantAnalysis






Matrizes de Confusão



- *LinearDiscriminantAnalysis*, *LogisticRegression*, *DecisionTreeClassifier*, *MLPClassifier*, *svm.LinearSVC* e *Perceptron* tiveram resultados bastante semelhantes;
- *KNeighborsClassifier* e *GaussianNB* mostraram resultados insatisfatórios;

- Criação de um mecanismo capaz de obter issues através da API do GitHub, filtrando labels relacionadas a segurança;

- <https://github.com/bmphx2/aprendizagem-de-maquina>

-  Scott Chacon and Ben Straub, *Pro git*, Apress, 2020.
-  Doaa Mohey El-Din, *Enhancement bag-of-words model for solving the challenges of sentiment analysis*, International Journal of Advanced Computer Science and Applications **7** (2016).
-  Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to information retrieval*, Cambridge University Press, Cambridge, UK, 2008.