Métodos Probabilísticos para Engenharia Informática

**2018/2019**



**Docentes:** Prof. Carlos Bastos | Prof. António Teixeira

Implementação de um Contador Estocástico, Counting Bloom Filter e Encontrar Similares (Usando MinHash)

Tomás Batista **89296**

Rodrigo Oliveira **90514**

ÍNDICE

[Introdução 3](#_Toc532059716)

[Desenvolvimento 3](#_Toc532059717)

[Conclusão 5](#_Toc532059718)

[Ferramentas de Suporte 5](#_Toc532059719)

[Bibliografia 5](#_Toc532059720)

# Introdução

No âmbito da unidade curricular de Métodos Probabilísticos para a Engenharia Informática, foi-nos proposto a realização de um trabalho cujo objetivo é a implementação de um Contador Estocástico, um Counting Bloom Filter e Deteção de Similares usando MinHash. Cada um destes três módulos foi testado com os conteúdos dos guiões das aulas práticas. Foi também criada uma aplicação conjunta que implementa os três módulos.

# Desenvolvimento

A nossa ideia de implementação foi trabalharmos com ficheiros que contêm Passwords. Usámos um ficheiro com as 10 mil passwords mais comuns, um com as 370 passwords banidas pelo Twitter, entre outros com diversas passwords.

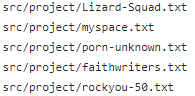


Figura Lista de ficheiros usados na opção 3 (MinHash)

De modo a facilitar o utilizador criámos um **menu** em que o utilizador decide que módulo que experimentar usando apenas inputs de teclado.

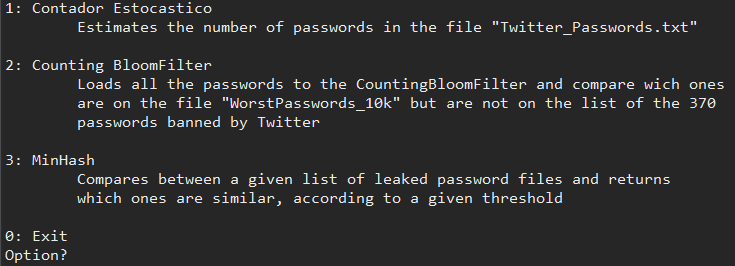


Figura Menu da Aplicação Conjunta

No módulo do **Contador Estocástico** estimamos a quantidade de Passwords nos ficheiros “MostCommom10kPasswords” e “Twitter\_Passwords” (soma de ambas). O número real de passwords é 10370.

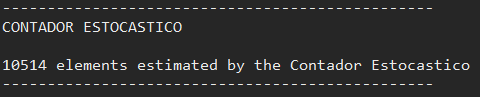
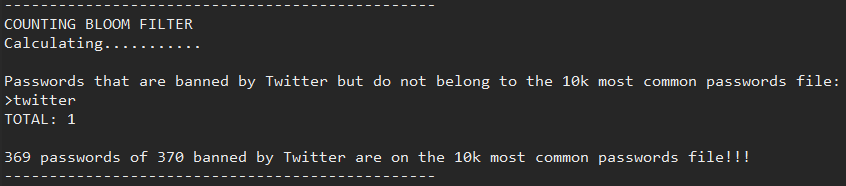


Figura 3 Teste Contador Estocástico

No módulo do **Counting Bloom Filter** testamos quais as passwords banidas pelo Twitter que não constam nas 10 mil passwords mais comuns.

Figura 4 Teste Counting Bloom Filter



No módulo do **Encontrar Similares (c/ MinHash)** usamos vários ficheiros com passwords que comparamos (usando um threshold de 0.65) vendo assim quais são similares e o respetivo índice de similaridade.

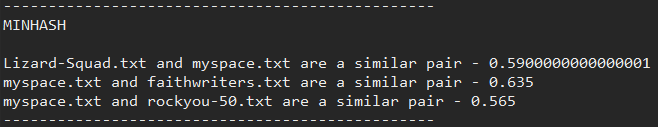


Figura Teste Encontrar Similares (c/ MinHash)

# Conclusão

Considerando todos os requisitos impostos no enunciado, podemos dizer que o resultado foi alcançado. Com os ficheiros à disposição e com todo o trabalho realizado podemos concluir que os utilizadores usam passwords comuns semelhantes em vários sites, independentemente de quão diferentes são os sites em termos de conteúdo.

# Ferramentas de Suporte

* [GitHub/Projeto\_MPEI](https://github.com/tomas99batista/Projeto_MPEI)

# Bibliografia

* [GitHub - SecLists](https://github.com/danielmiessler/SecLists) (Ficheiros c/ passwords)