Métodos Probabilísticos para Engenharia Informática

2018/2019

universidade de aveiro deti departamento de electrónica, telecomunicações e informática

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

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.

```
src/project/Lizard-Squad.txt
src/project/myspace.txt
src/project/porn-unknown.txt
src/project/faithwriters.txt
src/project/rockyou-50.txt
```

Figura 1 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.

```
1: Contador Estocastico
        Estimates the number of passwords in the file "Twitter_Passwords.txt"

2: Counting BloomFilter
        Loads all the passwords to the CountingBloomFilter and compare wich ones
        are on the file "WorstPasswords_10k" but are not on the list of the 370
        passwords banned by Twitter

3: MinHash
        Compares between a given list of leaked password files and returns
        which ones are similar, according to a given threshold

0: Exit
Option?
```

Figura 2 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.

```
------------------------------------------------
CONTADOR ESTOCASTICO

10514 elements estimated by the Contador Estocastico
------------------------------------------------
```

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.

```
------------------------------------------------
COUNTING BLOOM FILTER
Calculating...........

Passwords that are banned by Twitter but do not belong to the 10k most common passwords file:
>twitter
TOTAL: 1

369 passwords of 370 banned by Twitter are on the 10k most common passwords file!!!
------------------------------------------------
```

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.

```
------------------------------------------------
MINHASH

Lizard-Squad.txt and myspace.txt are a similar pair - 0.5900000000000001
myspace.txt and faithwriters.txt are a similar pair - 0.635
myspace.txt and rockyou-50.txt are a similar pair - 0.565
------------------------------------------------
```

Figura 5 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

Bibliografia

- GitHub - SecLists (Ficheiros c/ passwords)