ГУАП

КАФЕДРА № 43

ОТЧЕТ
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

ассистент		Д.А. Кочин
должность, уч. степень, звание	подпись, дата	инициалы, фамилия

ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №1

Работа с текстовыми потоками в командном интерпретаторе Bash

по курсу: ОПЕРАЦИОННЫЕ СИСТЕМЫ

РАБОТУ ВЫПОЛНИЛ

СТУДЕНТ ГР. № 4231		К.А. Чистякова
	подпись, дата	инициалы, фамилия

Санкт-Петербург 2024

Цель работы

Изучение принципов работы с командным интерпретатором GNU/Linux и основ обработки текстовых файлов с помощью команд grep, awk, sed.

Вариант задания

1. Загрузить копию рейтингового списка популярных доменных имен CISCO Umbrella Popularity List Top 1 Million с помощью команды wget https://github.com/markpolyak/datasets/raw/main/data/top-1m.csv.tar.bz2 или с использованием утилиты curl, и распаковать. Файл top-1m.csv содержит 1 миллион самых популярных доменных имен и состоит из двух полей, разделенных запятой: рейтинговый номер и доменное имя. В файл results.txt необходимо сохранить топ-15 самых часто встречающихся в файле dns-tunneling.log доменных имен из числа присутствующих в рейтинге CISCO Umbrella Popularity List. Если доменные имена встречаются с одинаковой частотой, отсортировать их в лексикографическом порядке. В переменную VAR_2 записать рейтинговый номер (из CISCO Umbrella Popularity List Top 1 Million) четвертого самого часто встречающегося доменного имени.

Описание входных данных

1. **Файл `dns-tunneling.log`** — текстовый файл, содержащий логи DNS-сервера. Каждая строка представляет собой запись запроса и состоит из нескольких параметров, разделенных символом табуляции (`\t`). Параметры включают:

- **Название провайдера телекоммуникационных услуг** — строка.
- **Название узла** — строка.
- **Порядковый номер запроса** — целое число.
- **Отметка времени** — два числа, разделенных точкой (первое число — количество секунд с 1 января 1970 года, второе — количество микросекунд).
- **IP-адрес пользователя** — строка.
- **Порт пользователя** — целое число.
- **Локальный IP-адрес** — строка.
- **Локальный порт** — целое число.
- **Название оборудования DNS-сервера** — строка.
- **Класс запроса** — целое число.
- **Тип запроса** — целое число.
- **Код возвращаемого значения** — целое число.
- **Флаги** — целое число.
- **Вспомогательный идентификатор** — целое число.

- **Запрашиваемый URL** — строка.
- **Зона** — строка.
- **Вспомогательные поля 1–4** — строки.
- **Ответ сервера** — строка.
- **Вспомогательные поля 5 и 6** — строки.
- **Длина ответа** — целое число.

2. **Файл `top-1m.csv`** — CSV-файл, содержащий топ 1 миллион популярных доменных имен из рейтинга CISCO Umbrella. Формат файла:

- **Рейтинговый номер** — целое число (первое поле).
- **Доменное имя** — строка (второе поле).

Результат выполнения работы

1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.
1 . 934692,zzz.visaissuercert.visa.com.

TASKID = 1
VAR_1 = 251384
VAR_2 = 934692,zzz.visaissuercert.visa.com.

Исходный код программы

```
#!/bin/bash

TASKID=1
VAR_1=$(wc -l < dns-tunneling.log)

if [[ ! -f "$HOME/lab1/task1/top-1m.csv" ]]; then
  echo "Файл top-1m.csv не найден. Загрузите и распакуйте его перед запуском
скрипта."
  exit 1
fi

# Извлекаем доменные имена из логов, подсчитываем их частоту
```

```
awk -F'\t' '{print $15}' dns-tunneling.log | sort | uniq -c | sort -nr > domain_counts.txt

# Сортируем файл top-1m.csv по доменам
sort -t, -k2,2 "$HOME/lab1/task1/top-1m.csv" > sorted_top_1m.csv

# Используем awk для объединения данных вместо join
awk -F'\t' '
   BEGIN {
      while (getline < "sorted_top_1m.csv") {
         domains[$2] = $1  # Сохраняем рейтинг домена
      }
   }
   FNR==NR { count[$2] = $1; next }  # Читаем из domain_counts.txt
   {
      if ($2 in domains) {
         print count[$2], $2, domains[$2]  # Выводим частоту, домен и рейтинг
      }
   }
' domain_counts.txt sorted_top_1m.csv | sort -k1,1nr -k2,2 | head -15 > results.txt

# Извлекаем рейтинговый номер четвертого домена
VAR_2=$(awk 'NR==4 {print $3}' results.txt)

echo "TASKID = $TASKID"
echo "VAR_1 = $VAR_1"
echo "VAR_2 = $VAR_2"
```

Выводы

В ходе выполнения лабораторной работы была разработана и протестирована программа на Bash для анализа логов DNS-сервера и выявления наиболее часто встречающихся доменных имен среди популярных, используя рейтинг CISCO Umbrella Top 1 Million.