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.

**Задание и последовательность выполнения работы:**

1. Загрузите копию рейтингового списка популярных доменных имён 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 утилиты curl и распакуйте. Файл top-1m.csv содержит 1 миллион самых популярных доменных имён и состоит из двух полей, разделённых запятой: рейтинговый номер и доменное имя. В файл results.txt необходимо сохранить 15 самых часто встречающихся в файле dns-tunneling.log доменных имён из рейтинга CISCO Umbrella Popularity List. Если доменные имена встречаются с одинаковой частотой, отсортируйте их в лексикографическом порядке. В переменную VAR_2 записать рейтинговый номер (из списка 1 миллиона самых популярных доменных имён CISCO Umbrella) четвёртого по популярности доменного имени.

**Описание выходных 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.
- **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:**

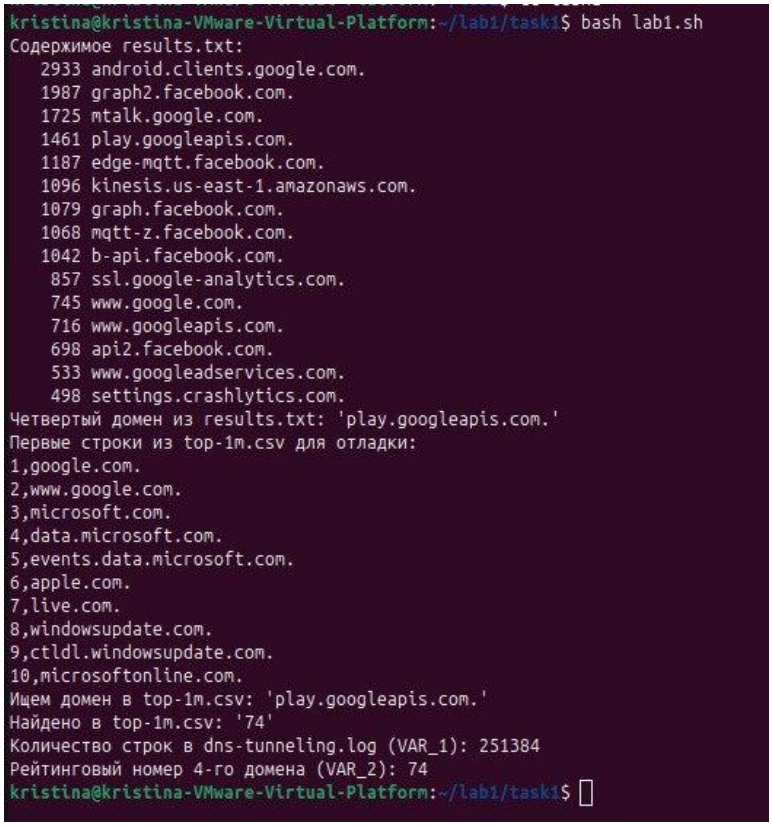
```
#!/bin/bash

# Задание: объявляем переменную TASKID
TASKID=1

# Подсчитываем количество строк в dns-tunneling.log
VAR_1=$(wc -l < dns-tunneling.log)
export VAR_1

# Извлекаем и сортируем домены из dns-tunneling.log, берем только 15-й столбец (предположим, что домены там)
cut -f15 dns-tunneling.log | sort | uniq -c | sort -nr > domain_counts.txt

# Извлекаем доменные имена из top-1m.csv (вторую колонку с доменами), приводим их к нижнему регистру для корректного сравнения
cut -d',' -f2 top-1m.csv | tr '[:upper:]' '[:lower:]' | sed 's/^[[:space:]]*//;s/[[:space:]]*$//' > top_domains.txt

# Находим домены из dns-tunneling.log, которые есть в top-1m.csv, сортируем по частоте и лексикографически
grep -Ff top_domains.txt domain_counts.txt | sort -nr -k1,1 -k2,2 | head -15 > results.txt

# Отладка: выводим содержимое results.txt для проверки
echo "Содержимое results.txt:"
cat results.txt

# Извлекаем 4-й домен из results.txt, учитывая возможные пробелы и табуляции, и не удаляем точку в конце домена
FOURTH_DOMAIN=$(sed -n '4p' results.txt | awk '{print $2}' | tr '[:upper:]' '[:lower:]' | sed 's/^[[:space:]]*//;s/[[:space:]]*$//')

# Отладка: выводим 4-й домен
echo "Четвертый домен из results.txt: '$FOURTH_DOMAIN'"
```

```
# Отладка: выводим содержимое top-1m.csv и проверяем, как выглядит первый и 4-й домен
echo "Первые строки из top-1m.csv для отладки:"
head -n 10 top-1m.csv

echo "Ищем домен в top-1m.csv: '$FOURTH_DOMAIN'"

# Ищем рейтинг для 4-го домена в top-1m.csv
VAR_2=$(grep -m 1 ",$FOURTH_DOMAIN" top-1m.csv | cut -d ',' -f 1)

# Отладка: выводим, что мы нашли в top-1m.csv
echo "Найдено в top-1m.csv: '$VAR_2'"

# Проверка, если VAR_2 пустое, значит домен не найден в top-1m.csv
if [ -z "$VAR_2" ]; then
    echo "Ошибка: не удалось найти домен '$FOURTH_DOMAIN' в top-1m.csv"
    exit 1
fi

# Экспортируем переменную VAR_2
export VAR_2

# Выводим количество строк в dns-tunneling.log и рейтинг 4-го домена
echo "Количество строк в dns-tunneling.log (VAR_1): $VAR_1"
echo "Рейтинговый номер 4-го домена (VAR_2): $VAR_2"
```

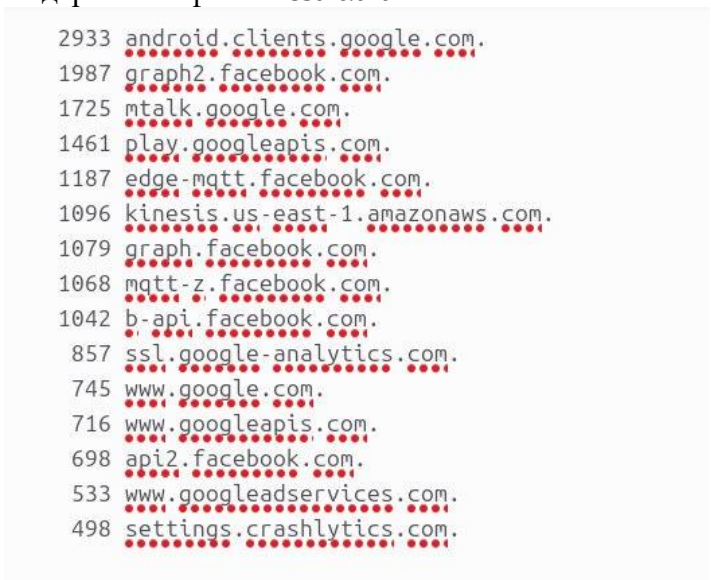
**test.sh**

```
#!/bin/bash
source ./lab1.sh

# Проверка наличия файла results.txt
if [ ! -f results.txt ]; then
  echo "Ошибка: файл results.txt не найден!"
  exit 1
fi

# Проверка, что файл results.txt содержит хотя бы 1 строку
if [ ! -s results.txt ]; then
  echo "Ошибка: файл results.txt пуст!"
  exit 1
fi

# Проверка, что в файле results.txt есть хотя бы 15 строк
lines=$(wc -l < results.txt)
if [ "$lines" -lt 15 ]; then
  echo "Ошибка: в файле results.txt меньше 15 строк!"
  exit 1
fi

# Проверка корректности данных в переменной VAR_1
if [ -z "$VAR_1" ]; then
  echo "Ошибка: переменная VAR_1 не установлена!"
  exit 1
fi

# Проверка корректности данных в переменной VAR_2
if [ -z "$VAR_2" ]; then
  echo "Ошибка: переменная VAR_2 не установлена!"
  exit 1
fi

# Проверка корректности значений VAR_2 (например, оно должно быть числом)
if ! [[ "$VAR_2" =~ ^[0-9]+$ ]]; then
  echo "Ошибка: VAR_2 не является числом!"
  exit 1
fi

# Проверка, что VAR_2 соответствует одному из рейтинговых номеров
if ! grep -q "^$VAR_2," top-1m.csv; then
  echo "Ошибка: VAR_2 не найден в рейтинговом списке!"
  exit 1
fi

# Если все проверки пройдены успешно
echo "Тесты пройдены успешно!"
```

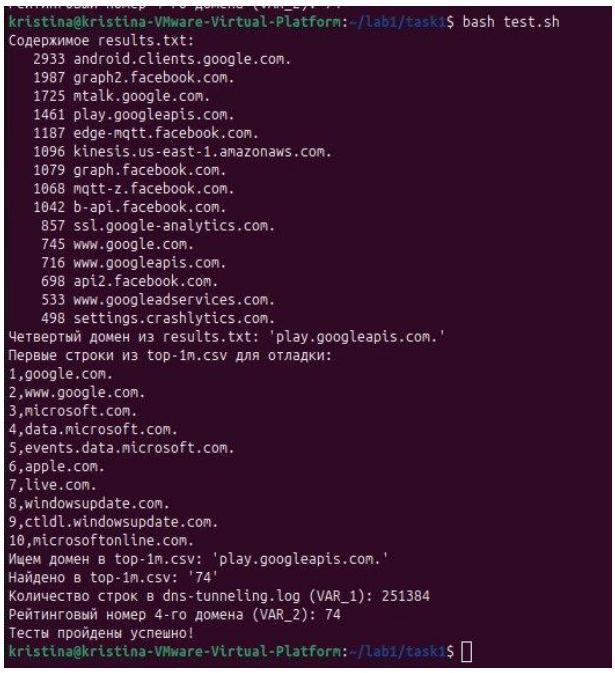
Результат запуска lab1.sh




Содержимое файла result.txt




Результат работы test.sh




**Вывод:**

В ходе выполнения данной лабораторной работы были изучены принципы работы с командным интерпретатором GNU/Linux и освоены базовые инструменты обработки текстовых файлов, включая команды `grep`, `awk` и `sed`.

Была загружена и распакована копия рейтингового списка популярных доменных имен CISCO Umbrella Popularity List Top 1 Million. В ходе анализа данных из `dns-tunneling.log` с использованием команд обработки текстовых файлов были выявлены 15 наиболее часто встречающихся доменных имен, содержащихся в списке CISCO Umbrella. Эти доменные имена были отсортированы по частоте встречаемости, а при одинаковых значениях частоты — в лексикографическом порядке, после чего результаты были сохранены в `results.txt`.

Дополнительно был определен рейтинговый номер четвертого по популярности доменного имени и записан в переменную `VAR_2`.

Таким образом, в рамках лабораторной работы были получены практические навыки работы с командным интерпретатором GNU/Linux и инструментами обработки текстовых данных, что позволило эффективно анализировать большие объемы информации с использованием стандартных утилит.