Министерство образования Республики Беларусь

Учреждение образования «Белорусский государственный университет   
информатики и радиоэлектроники»

Факультет компьютерных систем и сетей

Кафедра информатики

Дисциплина «Операционные среды и системное программирование»

|  |  |
| --- | --- |
|  |  |

**ОТЧЕТ**

к лабораторной работе № 2

на тему «Обработка текстовой информации. Регулярные выражения»

Выполнил             А. К. Хрищанович

Проверил                          Н. Ю. Гриценко

Минск 2024

**СОДЕРЖАНИЕ**

[1 Постановка задачи 3](#_Toc157722973)

[2 Краткие теоретические сведения 4](#_Toc157722974)

[3 Результаты выполнения лабораторной работы 5](#_Toc157722975)

[Выводы 6](#_Toc157722976)

[Список использованных источников 7](#_Toc157722977)

[Приложение А (обязательное) Листинг исходного кода 8](#_Toc157722978)

# **1 ПОСТАНОВКА ЗАДАЧИ**

Целью выполнения данной лабораторной работы является изучение методов и средств обработки текстовой информации, включая регулярные выражения, и использующих их утилит. Кроме того, необходимо написать скрипт, обращающийся к необходимым программам, для обработки входных данных, а также предусмотреть поведение скрипта при ошибочных или «неочищенных» входных данных.

# **2 КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ**

Обработка текстовой информации в среде оболочки часто выполняется с использованием различных утили и команд, таких как grep, sed, awk.[1]

Утилита grep, используется для поиска строк в текстовых файлах, которые соответствуют определенному шаблону или регулярному выражению. Данная утилита эффективна для быстрого поиска строк по паттернам, но не имеет возможности выполнения сложных манипуляция с данными.

Утилита sed используются для выполнения текстовых замен и преобразований в потоковых данных или в файлах. Данная утилита удобна для выполнения простых текстовых замен, но может быть неэффективна для сложных обработок данных.

Утилита awk является языком программирования для обработки и анализа текстовых данных, разбивки строк на поля и выполнения операций над ними. Данная утилита является мощным инструментом для обработки и анализа текстовых данных.

Регулярные выражения – мощный инструмент для работы с текстовой информацией. Они используются для поиска, сопоставления и манипулирования строками текста с использованием определенных шаблонов. Регулярные выражения широко используются в программировании, обработке текстов, поиске информации в текстовых файлах и многочисленных других областях.

Для выполнения данной лабораторной работы были использованы следующие сведения и концепции:

1 Файлы с расширением csv: файлы с расширением csv представляют собой файлы, в которых данные разделены запятыми. Этот формат данных часто используется для хранения табличных данных или списков информации, где каждая строка представляет собой отдельную запись, а значение внутри строки разделены запятыми.

2 Скрипты bash: скрипты bash являются подмножеством скриптов shell, предоставляющим большие возможности для автоматизации задач.

3 Обработка полей: утилита awk предоставляет удобные средства для обработки полей в строках. С помощью параметров задается разделить полей, что делает обработку файла проще.

4 Арифметические операции: при помощи утилиты awk выполняются арифметические операции для вычисления суммы и среднего значения данных.

5 Циклы и агрегация данных: утилита awk позволяет легко использовать циклы для обработки данных и агрегировать данные с помощью встроенных структур, таких как массивы.

# **3 РЕЗУЛЬТАТЫ ВЫПОЛНЕНИЯ ЛАБОРАТОРНОЙ             РАБОТЫ**

В ходе выполнения лабораторной работы был разработан небольшой bash-скрипт, который позволяет обрабатывать файлы в формате csv с данными о товарах. Программа подсчитывает суммарное количество товаров с одинаковым названием, средней ценой и общей стоимостью. Результат работы bash-скрипта представлен на рисунке 3.1.

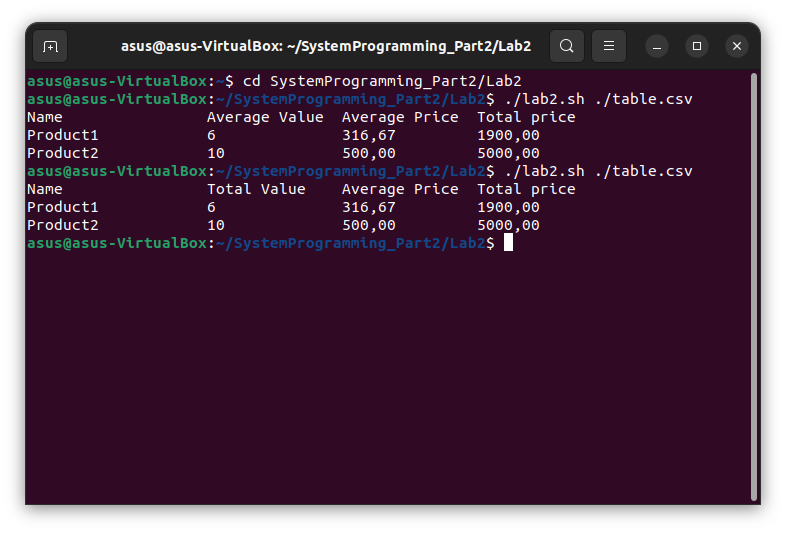


Рисунок 3.1 – Результат работы bash-скрипта

Если в нужном файле количество и цена товара будут прописаны неправильно, то скрипт будет завершен, вызывая ошибку и предупреждая о том, что данные в файле некорректны. Результат обработки некорректных данных представлен на рисунке 3.2.

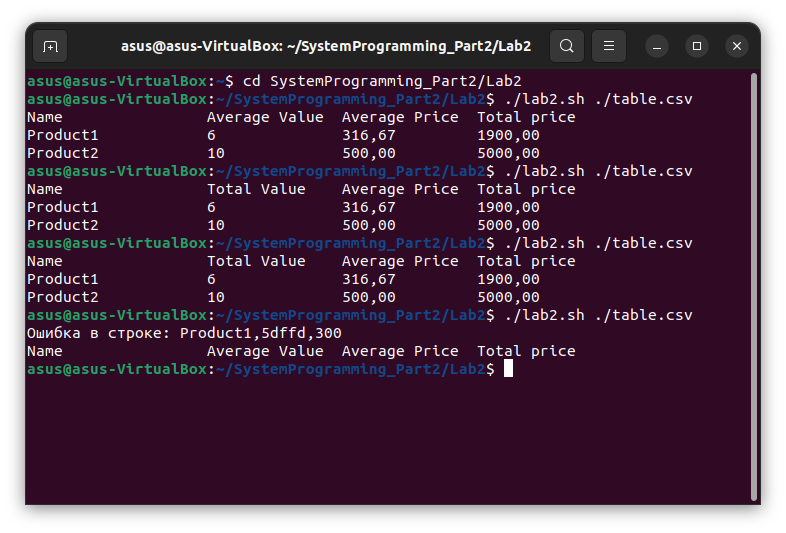


Рисунок 3.2 – Результат обработки некорректных данных

Наименования продуктов выводятся без повторений, подсчитывается их общее количество, общая цена и средняя цена, которая является результатом деления общей цены на общее количество товаров.

# **ВЫВОДЫ**

В ходе лабораторной работы были изучены методы и средства обработки текстовой информации, включая регулярные выражения, и использующих их утилит. Кроме того, в ходе лабораторной работы был разработан скрипт, обращающийся к необходимым программам, для обработки входных данных, а также было предусмотрено поведение скрипта при ошибочных или «неочищенных данных».

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

[1] Оболочка пользователя в Linux [Электронный ресурс]. – Режим доступа: https://habr.com/ru/companies/itsumma/articles/492932/. – Дата доступа: 07.02.2024.

[2] Скрипты shell [Электронный ресурс]. – Режим доступа: <https://support.apple.com/ru-ru/guide/terminal/>. – Дата доступа: 30.01.2024.

[3] BashTutorials [Электронный ресурс]. – Режим доступа:   
https://prog /bash-tutorials/. – Дата доступа: 30.01.2024.

## **ПРИЛОЖЕНИЕ А**

## **(обязательное)**

## **Листинг исходного кода**

Листинг 1 – Программный код разработанного приложения

#!/bin/bash

awk -F ',' '{

name=$1

value=$2

price=$3

if (value !~ /^[0-9]+$/ || price !~ /^[0-9]+(\.[0-9]+)?$/) {

print "Ошибка в строке:", $0

exit 1

}

total\_value[name] += value

total\_price[name] += value\*price

}

END {

printf "%-20s%-15s%-15s%-15s\n", "Name", "Average Value", "Average Price", "Total price"

for (name in total\_value) {

avg\_price = total\_price[name] / total\_value[name]

printf "%-20s%-15d%-15.2f%-15.2f\n", name, total\_value[name], avg\_price, total\_price[name]

}

}' table.csv