

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:
студент гр. 153503
Татарин В.В.

Проверил:
Гриценко Н.Ю.

Минск 2024

СОДЕРЖАНИЕ

1 Цель работы.....	3
2 Теоретические сведения.....	4
3 Полученные результаты.....	5
Выводы	6
Список использованных источников.....	7
Приложение А (обязательное) листинг кода	8

1 ЦЕЛЬ РАБОТЫ

Изучить методы и средства обработки текстовой информации, включая регулярные выражения, и использующих их утилит. Написать скрипт для *sed*, *awk* и т.д., либо скрипт *shell*, обращающийся к необходимым программам, для обработки входных данных. Необходимо предусмотреть поведение скрипта при ошибочных входных данных.

Написать скрипт, реализующий замену строчных букв на заглавные в начале предложений, т.е. в начале документа и после знаков конца предложения.

2 ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

В командный файл могут быть переданы параметры. В *shell* используются позиционные параметры, т.е. существенна очередность их следования. В командном файле соответствующие параметрам переменные, аналогично переменным *shell*, начинаются с символа "\$", а далее следует одна из цифр от 0 до 9. При вызове командного файла и перечислении через пробел двух аргументов, они попадают в новую среду под стандартными именами "1" и "2". В стандартной переменной с именем "0" будет храниться имя командного файла.

Поскольку число переменных, в которые могут передаваться параметры, ограничено одной цифрой от 1 до 9, то для передачи большего числа параметров используется специальная команда *shift*. При каждом вызове этой команды происходит сдвиг влево передаваемых параметров относительно индексов, по которым происходит обращение к этим параметрам в среде командного файла. Таким образом, если вызвать команду *shift* 3 раза, то при обращении по индексу 1 будет получен параметр, который был передан четвертым по счету [1].

Регулярные выражения – это специальным образом записанные строки, используемые для поиска символьных шаблонов в тексте.

Как правило, регулярное выражение состоит из трёх основных частей:

1 Якорь – определяет позицию шаблона в строке текста. "^" – якорь, определяющий начало строки. "\$" – якорь, определяющий конец строки.

2 Набор символов – для поиска соответствий в заданных позициях строки текста. Символ "." соответствует любому произвольному символу. Алфавитные, цифровые символы и пробел представляют сами себя. Прочие символы – интерпретация зависит от диалекта.

3 Модификатор – задаёт количество повторов предыдущего символа или набора символов. "*" – любое количество повторов символа или набора, в том числе и нулевое. "?" соответствует нулю или одному экземпляру символа или набора. "+" соответствует одному или большему количеству экземпляров символа или набора.

Если возникает необходимость задать символ из определённой группы, например, только цифровой символ, или только гласную букву нижнего регистра, или только символы пунктуации, то используются квадратные скобки, внутри которых определяются требуемые символы [2].

3 ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

В результате выполнения лабораторной работы был написан скрипт, обращающийся к необходимым программам, для обработки входных данных.

Скрипт реализует замену строчных букв на заглавные в начале предложений, т.е. в начале документа и после знаков конца предложения. Скрипт считывает данные из исходного файла, производит обработку и записывает полученный результат в другой файл. Исходный текст для обработки содержится в файле *t.txt* (рисунок 1).

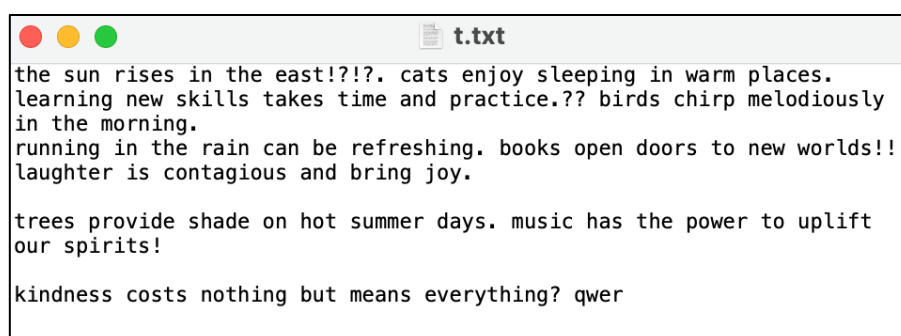


Рисунок 1 – Файл с исходным текстом для обработки

Обработанный текст с установленными заглавными буквами сохраняется в файл *out.txt* (рисунок 2).

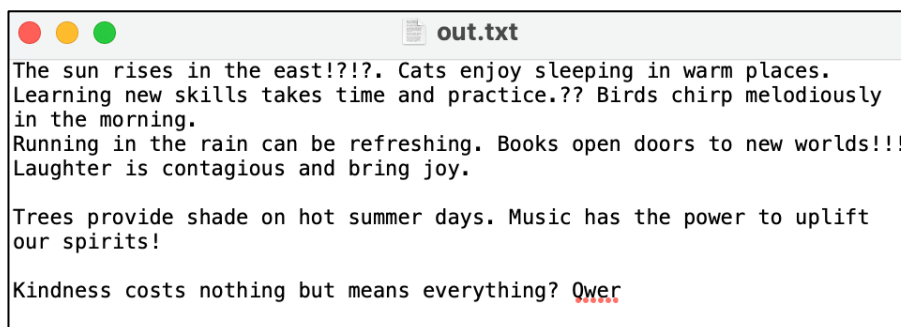


Рисунок 2 – Файл с обработанным текстом

ВЫВОДЫ

В результате выполнения лабораторной работы были изучены методы и средства обработки текстовой информации, включая регулярные выражения, и использующих их утилит. Написать скрипт, обращающийся к необходимым программам, для обработки входных данных.

Скрипт реализует замену строчных букв на заглавные в начале предложений, т.е. в начале документа и после знаков конца предложения.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Среда shell (переменные и параметры) [Электронный ресурс]. – Режим доступа: <http://const.uimech.org/man/shell4.html>.

[2] Регулярные выражения [Электронный ресурс]. – Режим доступа: http://asmodeus.com.ua/library/programing/shell/regular_expressions_bash.html.

ПРИЛОЖЕНИЕ А

(обязательное)

Листинг кода

Листинг 1 – Файл *auto.sh*:

```
#!/bin/zsh

if [ $# -lt 2 ];
then
    echo "Usage: $0 <input_file> <output_file>"
    exit 1
fi

filename=$1
output_file=$2

if [ ! -f "$filename" ]
then
    echo "File $filename does not exist."
    exit 7
fi

content=$(cat "$filename")

capitalized_content=$(echo "$content" | perl -pe 'BEGIN{undef $/;}
s/([.!?]\s+|^)(\w)/$1\u$2/gsm')

echo "$capitalized_content" > "$output_file"

echo "Capitalized content has been written to $output_file"
```