Федеральное государственное автономное
образовательное учреждение
высшего профессионального образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий

Кафедра вычислительной техники

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №6

«Подсчёт количества информации в сообщении»

Преподаватель Пушкарёв К.В.
подпись, дата инициалы, фамилия

Студент КИ15-08Б Войченко В.В.
подпись, дата инициалы, фамилия

Красноярск 2016

Цели работы:

1. Формирование представления об избыточности естественного языка.
2. Формирование понимания теоретического основания сжатия данных.
3. Получение навыков работы с файлами в MATLAB.
4. Получение навыков работы со строками в MATLAB.

Порядок выполнения работы:

1. Выполнить все задания.
2. Продемонстрировать выполнение заданий преподавателю.
3. Подготовить отчёт.
4. Защитить лабораторную работу перед преподавателем.

Указания:

1. Работу выполнять индивидуально.
2. Данные для анализа взять из электронного курса.
3. Символы новой строки в обрабатываемых файлах игнорировать.
4. В MATLAB закрыть все открытые в данный момент с помощью fopen() файлы можно командой fclose all.
5. Внимание! Перед началом работы генератор случайных чисел MATLAB необходимо инициализировать (см. «Генератор случайных чисел MATLAB и его инициализация»).

Задания

1. Написать функцию, подсчитывающую частоты встречаемости символов в тексте (частота встречаемости i-го символа: $f_i = \frac{N_i}{N}$, где N_i — сколько раз встретился данный символ, N — общее количество символов):
F = calc_freq(fname), где *fname* -- имя входного текстового файла; *F* -- результат, массив структур *(F(i).c* -- символ, *F(i).freq* -- его частота в тексте из файла *fname*).
Указание: для чтения строк из файла можно использовать функцию *fgetl()*.

2. Написать функцию, определяющую количество информации в тексте, рассматриваемом как цепочка независимых символов:
b = calc_info(fname, F), где *fname* -- имя входного текстового файла; *F* -- встречаемость символов в файле *fname* (результат *calc_freq()*); *b* -- количество информации в файле *fname* в байтах, подсчитанное на основе встречаемости *F*.

Указание: для определения количества информации в одном символе использовать формулу $log_2 p$, где p -- вероятность появления этого символа (приближённо равна частоте встречаемости).

3. С помощью функций **calc_freq(), calc_info()** определить частоты встречаемости символов и количество информации в приложенном к заданию тексте. Определить среднее количество информации, приходящееся на один символ. Определить энтропию и избыточность алфавита.

4. Сравнить среднее количество информации на один символ с энтропией алфавита.

5. Сравнить количество информации с объёмом (1 символ -- 1 байт).

6. Сгенерировать в MATLAB текст командой **gen_txt('text.txt', 64, 64)** с помощью приложенной к заданию программы и выполнить для него задания 3-5.

Результаты работы:

1. Отчёт, включающий программный код и результаты.
2. **Текстовый файл, сгенерированный в п. 6.**

I. Функция **calc_freq :**

```
function[F] = calc_freq(fname)
fid = fopen(fname, 'r');

if fid == -1
    error('Файл не был открыт!');
end

kolvo = 0;
string = fgetl(fid);
F(1).c = string(1);
F(1).freq = 0;
while ischar(string)
    kolvo = kolvo + numel(string);
    for i = 1:numel(string)
        charPos = strfind([F.c], string(i));
        if isempty(charPos)
            add_new.c = string(i);
            add_new.freq = 1;
            F(end + 1) = add_new;
        elseif isempty(charPos) == false
            F(charPos).freq = F(charPos).freq + 1;
        end
    end
    string = fgetl(fid);
end

for i = 1:numel(F)
    F(i).freq = F(i).freq / kolvo;
end

fclose(fid);
end
```

II. Функция **calc_info:**

```
function [information, kolvo] = calc_info(fname, F)
fid = fopen(fname, 'r');

if fid == -1
    error('Файл не был открыт!');
end

FirstStringRead = 1;
kolvo = 0;
information = 0;

while (ischar(string)) || (FirstStringRead)
string = fgetl(fid);
    kolvo = kolvo + numel(string);
    for i = 1:numel(string)
        information = information - log2(F(strfind([F.c],
string(i))).freq);
    end
    string = fgetl(fid);
end
fclose(fid);
information = information/8;
end
```

III. Функция **alph_ redudancy:**

```
function [redudancy] = alph_redundancy(P)
redudancy = 1 - alph_entropy(P) / log2(numel(P));
end
```

IV. Функция **alph_entropy:**

```
function [entropy] = alph_entropy(P)
entropy = -sum(P(P > 0).*log2(P(P > 0)));
end
```

Код для работы программы:

```
F = calc_freq('crime.txt'); % частота втречаемости символов
в 'crime.txt'
[information, length] = calc_info('crime.txt', F);
%количество информации в 'crime.txt'
freq = [F.freq];  %частота встречаемости символов в
'crime.txt'
entropy = alph_entropy(freq);
redudancy = alph_redundancy(freq);
sred_on_symb = (information/length)*8;
%среднее количество информации, приходящееся на один символ
if (sred_on_symb > entropy)
    disp('<crime.txt>Среднее количество информации больше
энтропии на '); abs (sred_on_symb - entropy)
end
if (sred_on_symb < entropy)
    disp('<crime.txt>Среднее количество информации меньше
энтропии на '); abs (sred_on_symb - entropy)
end
if (sred_on_symb == entropy)
    disp('<crime.txt>Среднее количество информации равно
энтропии.');
end
if (information > length)
    disp('<crime.txt>количество информации больше объёма на
'); abs (information - length)
end
if (information < length)
    disp('<crime.txt>количество информации меньше объёма на
'); abs (information - length)
end
if (information == length)
    disp('<crime.txt>количество информации равно объёму');
end
disp ('Количество информации в crime.txt: '); information
disp ('Длина текста в crime.txt: '); length
disp ('Средняя информация на символ в crime.txt');
sred_on_symb
disp ('Избыточность алфавита в crime.txt'); redudancy
%%
%gen_text('text.txt', 64, 64);
F1 = calc_freq('text.txt');
[information1, length1] = calc_info('text.txt', F1);
```

Продолжение

```
freq1 = [F1.freq];
entropy1= alph_entropy(freq1);
redundancy1 = alph_redundancy(freq1);
sred_on_symb1 = (information1/length1)*8;
if (sred_on_symb1 > entropy1)
    disp('<GENtext>Среднее количество информации больше
энтропии на '); abs (sred_on_symb1 - entropy1)
end
if (sred_on_symb1 < entropy1)
    disp('<GENtext>Среднее количество информации меньше
энтропии на '); abs (sred_on_symb1 - entropy1)
end
if (sred_on_symb1 == entropy1)
    disp('<GENtext>Среднее количество информации равно
энтропии.');
end
if (information1 > length1)
    disp('<GENtext>количество информации больше объёма на
'); abs (information1 - length1)
end
if (information1 < length1)
    disp('<GENtext>количество информации меньше объёма на
'); abs (information1 - length1)
end
if (information1 == length1)
    disp('<GENtext>количество информации равно объёму');
end
disp ('Количество информации в сгенерированном файле: ');
information1
disp ('Длина текста в сгенерированном файле: '); length1
disp ('Средняя информация на символ в сгенерированом файле
'); sred_on_symb1
disp ('Избыточность алфавита в сгенерированном файле ');
redundancy1
```

Результаты работы программы:

Таблица 1
"Встречаемость символов в тексте crime.txt"
(содержимое структуры F)

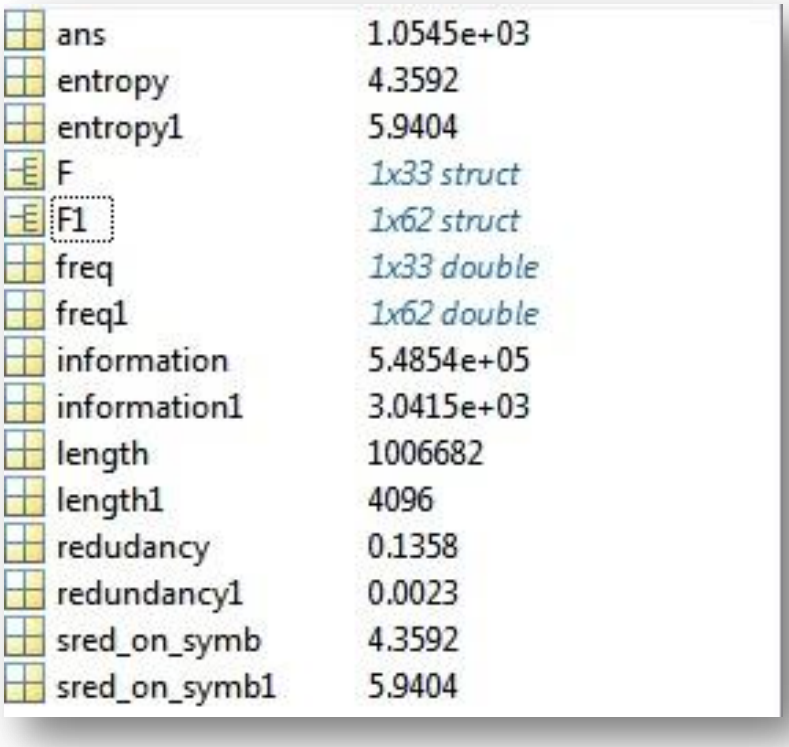
Символ	Частота встречаемости
	0,1669 *
а	0,0664
б	0,0145
в	0,0386
г	0,0141
д	0,0267
е	0,0717
ж	0,0095
з	0,0128
и	0,0541
й	0,0083
к	0,0275
л	0,0383
м	0,0262
н	0,0543
о	0,0958
п	0,0229
р	0,0348
с	0,0442
т	0,0540
у	0,0248
ф	0,0010
х	0,0071
ц	0,0023
ч	0,0151
ш	0,0069
щ	0,0025
ъ	0,0002
ы	0,0138
ь	0,0192
э	0,0030
ю	0,0047
я	0,0178

*- частота встречаемости символа ПРОБЕЛ

Таблица 2
"Встречаемость символов в тексте text.txt"
(содержимое структуры F1)

Символ	Частота встречаемости	Символ	Частота встречаемости
0	0,0142	L	0,0198
1	0,0137	l	0,0176
2	0,0149	M	0,0181
3	0,0129	m	0,0156
4	0,0154	N	0,0151
5	0,0203	n	0,0151
6	0,0134	O	0,0181
7	0,0161	o	0,0146
8	0,0149	P	0,0161
9	0,0159	p	0,0198
a	0,0161	q	0,0161
A	0,0168	Q	0,0183
b	0,0137	R	0,0144
c	0,0173	r	0,0205
C	0,0156	S	0,0166
d	0,0198	s	0,0137
D	0,0156	T	0,0142
e	0,0159	t	0,0166
E	0,0125	U	0,0188
F	0,0134	u	0,0134
f	0,0186	v	0,0129
G	0,0173	V	0,0154
g	0,0168	w	0,0171
H	0,0181	W	0,0190
h	0,0142	x	0,0166
I	0,0168	X	0,0181
i	0,0173	y	0,0166
J	0,0137	Y	0,0200
j	0,0217	Z	0,0146
k	0,0134	z	0,0154
K	0,0120		

Данные, полученные после выполнения программы:



Имя переменной/(ед. измерения)	**Значение**
sred_on_symb1/(бит/символ)	средняя информация на 1 символ сгенерированного документа
sred_on_symb/(бит/символ)	средняя информация на 1 символ документа 'crime.txt'
redudancy1/(бит)	избыточность алфавита сгенерированного документа
redudancy/(бит)	избыточность алфавита в документе 'crime.txt'
length1/(кол-во символов)	длина сгенерированного документа
length/(кол-во символов)	длина документа 'crime.txt'
information1/(бит)	объем информации сгенерированного текста
information/(бит)	объем информации текста 'crime.txt'
entropy1/(бит)	энтропия алфавита сгенерированного текста
entropy/(бит)	энтропия алфавита документа 'crime.txt'

Продолжение:

Command Window

<crime.txt>Среднее количество информации больше энтропии на $1.4684*10^{-11}$

<crime.txt>Количество информации меньше объёма на $4.5814*10^{5}$

Количество информации в crime.txt: $5.4854*10^{5}$

Средняя информация на символ в crime.txt: 4.3592

Избыточность алфавита в crime.txt: 0.1358

<GENtext>Среднее количество информации больше энтропии на $1.3056*10^{-13}$

<GENtext>количество информации меньше объёма на $1.0545*10^{3}$

Количество информации в сгенерированном файле: $3.0415*10^{3}$

Средняя информация на символ в сгенерированном файле: 5.9404

Избыточность алфавита в сгенерированном файле: 0.0023