Лабораторная работа № 4 по курсу криптографии

Выполнил студент группы М8О-308Б *Куликов Алексей.*

Условие

Сравнить

1. два осмысленных текста на естественном языке,
2. осмысленный текст и текст из случайных букв,
3. осмысленный текст и текст из случайных слов,
4. два текста из случайных букв,
5. два текста из случайных слов.

Как сравнивать: считать процент совпадения букв в сравниваемых текстах – получить дробное значение от 0 до 1 как результат деления количества совпадений на общее число букв. Расписать подробно в отчёте алгоритм сравнения и приложить сравниваемые тексты в отчёте хотя бы для одного запуска по всем пяти подпунктам. Осознать какие значения получаются в этих пяти подпунктах. Привести свои соображения о том почему так происходит.

Длина сравниваемых текстов должна совпадать. Привести соображения о том какой длины текста должно быть достаточно для корректного сравнения.

Описание алгоритма

Алгоритм сравнения текстов таков: Открываются два файла, подсчитывается их длина. Выбирается минимальная из длин n. Затем посимвольно считывается n символов параллельно из обоих файлов. При равенстве символов инкрементируется счетчик совпадений c. За результат подсчета принимается число $\frac{c}{n}$.

Текст из случайных букв генерируется следующим образом: Получаем случайное число от 0 до 51, смотрим меньше ли оно 26. Если да, то пишем символ с кодом 'a' + c, он будет из диапазона a-z. В противном случае пишем в файл символ с кодом 'A' + c `mod` 26. Повторяем нужное количество раз.

Текст из случайных слов генерируется следующим образом: Считывается словарь заблаговременно полученных из Python-библиотеки `nltk` английских слов. Генерируется случайно число от 0 до длины словаря. В поток пишется слово с таким номером в словаре. Повторяем нужное количество раз.

Эксперименты

Вычисление статистики для осмысленных текстов:

```
$ ./prog large_meaningful_file1 large_meaningful_file2
count to read: 2918308
match rate: 0.0646183
$ ./prog meaningful_file1 meaningful_file2
count to read: 43249
match rate: 0.0639552
$ ./prog small_meaningful_file1 small_meaningful_file2
count to read: 819
match rate: 0.0622711
$ ./prog really_small_meaningful_file1 really_small_meaningful_file2
count to read: 203
match rate: 0.0541872
```

Вычисление статистики для двух файлов из случайных букв:

```
$ ./random_char_file 3000000 random_char_file1
$ ./random_char_file 3000000 random_char_file2
$ ./prog random_char_file1 random_char_file2
count to read: 3000000
match rate: 0.019193
```

Вычисление статистики для осмысленного файла и файла из случайных букв:

```
$ ./prog random_char_file1 large_meaningful_file1
count to read: 3000000
match rate: 0.015113
$ ./prog random_char_file1 large_meaningful_file2
count to read: 2918308
match rate: 0.0150416
```

Вычисление статистики для двух файлов из случайных слов:

```
$ ./random_word_file 350000 random_word_file1
$ ./random_word_file 350000 random_word_file2
$ ./prog random_word_file1 random_word_file2
count to read: 3691084
match rate: 0.0599921
```

Вычисление статистики для осмысленного файла и файла из случайных слов:

```
$ ./prog random_word_file1 large_meaningful_file1
count to read: 2918308
```

```
match rate: 0.0580439
$ ./prog random_word_file1 large_meaningful_file2
count to read: 3426145
match rate: 0.0582859
```

Анализ

Из экспериментов видно, что шанс совпадения символов в двух осмысленных текстов на английском языке наибольший из представленных, и составляет 0.064. Шанс совпадения двух символов в случае двух файлов из случайных латинских символов составляет порядка 0.019. Для осмысленного текста и случайных букв – 0.015. Для файлов из случайных слов – 0.060. Для осмысленного текста и случайных слов – 0.058.

При подсчете статистики для осмысленных текстов различной длинны видно, что шанс совпадения более-менее "устаканивается" уже при длине текста в 800 символов, но 200 символов недостаточно. Поэтому можно сделать вывод, что длины в 400-500 символов будет вполне достаточно для объективной оценки.

В случае сравнения двух осмысленных текстов шанс совпадения значительно больше, чем у случайных символов, в силу специфики языка.

В английском языке буквы распределены неравномерно по частоте встречаемости, в отличие от случайно сгенерированного набора букв. По тем же данным из Википедии, например, буква 'e' встречается с частотой 11.162%, а суммарная частота всех гласных 37.47%, которая приходится всего на 5 гласных букв. Остальные 62.53% распределяются между оставшимися 21 согласными буквами. К тому же существуют буквы с крайне низкой частотой: на 'j', 'q', 'x', 'z' вместе приходится всего 0.475%. Видно, что шанс совпадения на гласных очень высок. Это и приводит показанному шансу совпадения.

Шанс совпадения для символов текста из случайных букв составляет 0.019 из-за того, что символы распределены равномерно, и вероятность каждого символа равняется $\frac{1}{52} \approx 0.01923$ (одинаковые буквы верхнего и нижнего регистров считаются различными).

Шанс совпадения символов для осмысленного текста и случайных букв и того ниже, по моему мнению, потому, что в осмысленном тексте достаточно часто встречаются символы, которые никогда не встречаются среди случайных букв. Среди них знаки препинания, цифры и т.п.

Для двух текстов из случайных слов вероятность совпадения меньше, чем для двух осмысленных, вероятно, из-за того, что из-за отсутствия в словах какого-либо порядка, по сравнению с осмысленными текстами, нарушается плотность частей речи, что ведед к изменению частот символов.

Для осмысленного текста и случайных слов вероятность совпадения символов несколько меньше, чем для двух осмысленных текстов, видимо, в силу того, что плотность распределения изменится из-за изменения состава частей речи в тексте, но среди случайных слов не присутствует знаков препинания, цифр и т.п.

Подсчитываемую таким образом статистику можно считать неким коэффициентом “похожести” текстов.

Стоит заметить, что из-за разных распределений вероятности букв в разных языках, подсчитываемая для них статистика будет различна. Также она может различаться для текстов на одном языке, но разного характера (техническая литература, художественная и т.д. и т.п.).

Выводы

Существует множество частотных характеристик текста, которые используются создании модели открытого текста, отражающей наиболее важные его свойства. Подобные модели могут исполховаться при автоматизации методов криптоанализа, связанных с перебором ключей, и распознаванием открытого текста.

Благодаря выявленным свойствам данную статистику можно использовать для распознавания открытого текста при попытках взлома различных шифрах.

Листинг кода

`main.cpp`

```
#include <iostream>
#include <fstream>

using namespace std;

int file_size(ifstream &f){
    f.seekg (0, std::ios::end);
    unsigned int size = f.tellg();
    f.seekg (0, std::ios::beg);
    return size;
}

int main(int argc, char **argv){
    if(argc < 3){
        cout << "usage: prog <file1> <file2>" << endl;
        return 0;
    }

    ifstream f1(argv[1]), f2(argv[2]);
    if(!f1.is_open()){
        cout << "Can't open " << argv[1] << " file" << endl;
        return 0;
    }

    if(!f2.is_open()){
        cout << "Can't open " << argv[2] << " file" << endl;
        return 0;
    }

    unsigned int count_to_read = min(file_size(f1), file_size(f2));

    cout << "count to read: " << count_to_read << endl;

    unsigned int match_count = 0;
    for(unsigned int i = 0; i < count_to_read; ++i){
        if(f1.get() == f2.get())
```

```
            ++match_count;
    }

    cout << "match rate: " << (double)match_count / count_to_read << endl;
}
```

random_word_file.cpp

```
#include <iostream>
#include <fstream>
#include <cstdlib>

#include <vector>
#include <string>

using namespace std;

int main(int argc, char **argv){
    if(argc < 2){
        cout << "usage: prog <file1>" << endl;
        return 0;
    }

    ofstream f(argv[2]);

    if(!f.is_open()){
        cout << "Can't open " << argv[2] << " file" << endl;
        return 0;
    }

    ifstream wf("word_list");
    unsigned int word_count;
    wf >> word_count;

    vector<string> words(word_count);
    for(unsigned int i = 0; i < word_count; ++i)
        wf >> words[i];

    srand(time(0));

    unsigned int n = atoi(argv[1]);

    for(unsigned int i = 0; i < n; ++i){
        f << words[rand() % word_count] << ' ';
    }
    f << endl;
}
```

random_char_file.cpp

```
#include <iostream>
#include <fstream>
#include <cstdlib>

using namespace std;

int main(int argc, char **argv){
    if(argc < 2){
        cout << "usage: prog <file1>" << endl;
        return 0;
    }

    ofstream f(argv[2]);

    if(!f.is_open()){
```

```
        cout << "Can't open " << argv[2] << " file" << endl;
        return 0;
    }

    srand(time(0));

    unsigned int n = atoi(argv[1]);

    for(unsigned int i = 0; i < n; ++i){
        unsigned int c = rand() % 52;
        f.put(c < 26 ? 'a' + c : 'A' + c % 26);
    }
}
```

1 Используемые тексты

Фрагмент осмысленного текста:

```
...
A member of the Hofkriegsrath from Vienna had come to Kutuzov the
day before with proposals and demands for him to join up with the army
of the Archduke Ferdinand and Mack, and Kutuzov, not considering
this junction advisable, meant, among other arguments in support of
his view, to show the Austrian general the wretched state in which the
troops arrived from Russia. With this object he intended to meet the
regiment; so the worse the condition it was in, the better pleased the
commander in chief would be. Though the aide-de-camp did not know
these circumstances, he nevertheless delivered the definite order that
the men should be in their greatcoats and in marching order, and
that the commander in chief would otherwise be dissatisfied. On
hearing this the regimental commander hung his head, silently shrugged
his shoulders, and spread out his arms with a choleric gesture.
...
```

Фрагмент текста из случайных букв:

```
...
LtFwjiQiQthkhTmlNnNBLsyTbeMzFPUqLzowkeECxnMggATTOHXbZXwbbLciAWbNXpLJWp
lVfzENCxjSegTedRHGCJOFhPseHGqDxByCbcqfAzXEhQkkHtQMcgrLylsfRIIpjItMLJRN
KOTtgeDqxVccEVNcgfhYNrPzZiNmuhZGVTaEzfUyDyAhtQLbxSZNLqOKbBXvIyDgTDmslI
tOgvxbliclCEYnUMXVPwSaVwgqBslMaEcIZzJNHnYjRyYoMWjBUEdrajIddTQFZSPYsanB
oNnfllVzJHDglGxNQHQvAGDzbsaVtPYHDloqZkQkRtqEBOrrvhpyqSzTkbOfRMMWADmzpE
JIXCmzqDsnmjnebNXOOLThaidALqBCUMkuowvgBpWNzjScyrsPdmWFWaHjSKLowVknrFVu
xRhWDBAbTTQyHNFfpmoJWZzUUKhNRdhowQMZUnbpItNRIUWzgkiELIyFUiULNDajVMLPBo
ejHtCSnyrVkcbVmAAgkWRXztgXIRMjHtVpoxheyyZIcbEodGunDOMEhuBrOQBvjYKxvtBT
udewGILloHaRvMYfgzwUpzSaxCavydQSIUQoCBzRjbkgNilwkkSbJKcJPEenHwHqrxgvBi
OmjBszJExTqpVZCzkrdQgmnpegoldPVSbGtufEACaQuxrWWePBUVNJMuPaFvQcPtIKNOoP
```

```
SQhOnbKLfzOZWelkyClFXDHMWpwmFnDadLOTMzEtAVUZZfjxIucFzlsYAqkIDNIHawAMxG
fYBaXcHIzpFdxeorcrHOZlejueFwRdcyDgaciIKkzRnWWenAXVRyIVHEZmaQRcRuKtYTbJ
dBcSZYyPbvmSvUNcyMRyfIcyFNRDGTOlWRFxremsbyMxUbzsOqsvBXTgKmlSHaFDtkDKQP
FSQTRKVqdljXiKUbsgPDAxffCaSfnIVSAnnTXikaVVAdhWGzdwEDTmlvmdBbLYvOlJjitt
LpQNsXJZYoXfuQrFNfkQhxOCNBNXmhSzwJmqGxRhMOmggfnUkxKTUbYIenhquZrsIDlRbc
YptMXbRkxDHKZdLXnrMukhWBbEfmXIRVXMjUNaGlEPVfSiEIbqElZANbHuPEcicBUlXImf
vsUSxoCbWDTBPTDewkylPcuRfoccyQhv
...
```

Фрагмент текста из случайных слов:

```
...
dirge patternless swampweed Anasazi exulceratory posttarsal
superintolerable figgle mimeographically nonsymphonic suppurant
Xylonite Montia nondeforestation hereticator balmony subversion
exasperate Enchelycephali neoclassicism fractocumulus axoplasm
drinkless nanosomus iliopsoas casino jagla repressionist wherewithal
periwinkle picksome yogasana whoremonging Dioscuri alnein apostatic
translater tersulphate capitulum chemotic commandingness upbrim
ungiftedness lobing mescal ethionic nondemobilization Flaminian
Greenland spontaneously periglandular precoiler precipitatedly
tersion untarred amygdal whelve holograph Neil Sclerodermi
seneschalsy cerebrogalactose
...
```