Московский авиационный институт
(Национальный исследовательский университет)
Факультет прикладной математики и физики
Кафедра вычислительной математики и программирования

Лабораторная работа № 4

по курсу «Криптография»

Студент: Алексюнина Ю.В.
Группа: 80-307Б
Преподаватель: Борисов А. В.
Оценка:

Москва, 2020

Постановка задачи:

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.

Решение:

- В качестве осмысленных текстов на естественном языке были взяты первый и второй том романа «Война и мир» Л.Н. Толстого на английском языке. Данные тексты были выбраны неслучайно. Существует мнение, что все части этого романа чрезвычайно похожи с точки зрения написания, и мне хотелось также проверить и совпадение слов.

- Текст из случайных слов генерируется из следующего словаря (чуть меньше 25 тысяч английских слов): http://svnweb.freebsd.org/csrg/share/dict/words?view=cocontent-type=text/plain

- Текст из случайных букв генерируется из букв английского алфавита в обоих регистрах и состоит из слов длиной от 3 до 10 знаков.

- Тексты из случайных букв/слов сравниваются с первым томом «Война и мир»

- Алгоритм сравнения: параллельно обходим оба текста, сравниваем знаки на одинаковых позициях, исключая символы пробела, перевода строки и null. Если знаки совпадают, то увеличиваем счётчик совпавших символов на 1. Сравнение регистрозависимое.

- При выполнении данной задачи одной из сложностей было восприятие Python-ом текста в другой кодировке, т.к. текстовые файлы читаются только в Unicode, а на моем ноутбуке не было возможности сохранить их в этой кодировке.

Исходный код:

```
import random
import string
import urllib.request


cnt_rnd_txt = 10
len_rnd_txt = 500000


def count_common_letters(text1, text2):
    cnt = 0
    for char1, char2 in zip(text1, text2):
        if (char1 == char2) and (char1 != '\n') and (char1 != ' ') and (char1 != '\0'):
            # print(char1)
            cnt += 1
    return cnt


def match_perc(text1, text2):
    return count_common_letters(text1, text2) / len(text1)


def gen_random_letters(n):
    text = ''
    while len(text) < n:
        len_word = random.randint(3, 10)
        word = ''.join(random.choice(string.ascii_letters) for _ in range(len_word))
        text += ' ' + word
    rem = len(text) - n
    if rem != 0:
        text = text[:-rem]
    return text


def gen_random_words(n):
    url = 'http://svnweb.freebsd.org/csrg/share/dict/words?view=co&content-type=text/plain'
    response = urllib.request.urlopen(url)
```

```
    words = response.read().decode()
    words = words.splitlines()
    text = ''
    while len(text) < n:
        text += ' ' + random.choice(words)
    rem = len(text) - n
    if rem != 0:
        text = text[:-rem]
    return text


def full_of_sense():
    print("1. Два осмысленных текста на естественном языке.")
    handle1 = open('1.txt', 'r')
    text1 = handle1.read()
    # text1 = [line.rstrip() for line in text1]
    handle2 = open('2.txt', 'r')
    text2 = handle2.read()
    # text2 = [line.rstrip() for line in text2]
    min_len = min(len(text1), len(text2))
    text1 = text1[:min_len]
    text2 = text2[:min_len]
    print("Длина текста: {0}".format(min_len))
    print("Процент совпадений: {0}".format(match_perc(text1, text2)))


def sense_and_randoml():
    print("2. Осмысленный текст и текст из случайных букв.")
    handle1 = open('1.txt', 'r')
    text1 = handle1.read()
    s = 0
    for _ in range(cnt_rnd_txt):
        text2 = gen_random_letters(len(text1))
        s += match_perc(text1, text2)
    s /= cnt_rnd_txt
    print("Длина текста: {0}".format(len(text1)))
    print("Процент совпадений: {0}".format(s))


def sense_and_randomw():
    print("3. Осмысленный текст и текст из случайных слов.")
    handle1 = open('1.txt', 'r')
    text1 = handle1.read()
    s = 0
    for _ in range(cnt_rnd_txt):
        text2 = gen_random_words(len(text1))
        s += match_perc(text1, text2)
    s /= cnt_rnd_txt
    print("Длина текста: {0}".format(len(text1)))
    print("Процент совпадений: {0}".format(s))


def randoml():
    print("4. Два текста из случайных букв.")
    s = 0
    for _ in range(cnt_rnd_txt):
        text1 = gen_random_letters(len_rnd_txt)
        text2 = gen_random_letters(len_rnd_txt)
        s += match_perc(text1, text2)
    s /= cnt_rnd_txt
    print("Длина текста: {0}".format(len_rnd_txt))
    print("Процент совпадений: {0}".format(s))
```

```
def randomw():
    print("5. Два текста из случайных слов.")
    s = 0
    for _ in range(cnt_rnd_txt):
        text1 = gen_random_words(len_rnd_txt)
        text2 = gen_random_words(len_rnd_txt)
        s += match_perc(text1, text2)
    s /= cnt_rnd_txt
    print("Длина текста: {0}".format(len_rnd_txt))
    print("Процент совпадений: {0}".format(s))


full_of_sense()
sense_and_randoml()
sense_and_randomw()
randoml()
randomw()
```

Результат работы программы:

1. Два осмысленных текста на естественном языке.
Длина текста: 476646
Процент совпадений: 0.01830289145403507
2. Осмысленный текст и текст из случайных букв.
Длина текста: 543202
Процент совпадений: 0.006512126244012356
3. Осмысленный текст и текст из случайных слов.
Длина текста: 543202
Процент совпадений: 0.0200013991111962
4. Два текста из случайных букв.
Длина текста: 500000
Процент совпадений: 0.004395799999999997
5. Два текста из случайных слов.
Длина текста: 500000
Процент совпадений: 0.014553359999999994

Выводы:

Как видно из результатов, наилучшие совпадения получаются путём сравнения двух осмысленных текстов и осмысленного текста с текстом из случайных слов. Худшие совпадения у осмысленного текста с текстом из случайных букв и у двух текстов из случайных букв.
К сожалению, первый и второй том произведения «Война и мир» на английском языке не совпадают настолько, насколько хотелось бы.
Думаю, полученные результаты можно было бы объяснить какими-то лингвистическими законами построения языка, но у меня нет достаточных знаний в этой области. Например, для букв английского языка характерна некоторая частотность, которая будет соблюдаться в осмысленных текстах, и которая не соблюдается в генерируемых из букв текстах. Эмпирически кажется, что размер слогов в осмысленных словах совпадает чаще, чем в случайных, и 5 букв, передающих гласные звуки в английском языке, будут совпадать чаще.

В текстах из случайных букв нет никаких ограничений на использование букв в верхнем регистре не на первой позиции в слове, что сильно снижает количество совпадений с осмысленным тестом.

Что касается достаточной длины текста для корректного сравнения, я нашла следующую информацию в книге «Определение жанра и автора литературного произведения статистическими методами» (Авторы: Ю. Орлов, К. Осминин):

Поскольку последовательность букв в тексте образует нестационарный временной ряд, необходимо понять, какой смысл имеет ВПФР. Ведь эмпирическая вероятность есть предел отношения (1) при $N \to \infty$, если таковой существует, поэтому значений k_i для каждого i должно быть достаточно много. Тогда ВПФР представляет собой набор вероятностей использования букв в тексте, объем которого должен быть достаточно большим, чтобы эти вероятности определялись с заданным уровнем точности в предположении стационарности выборки. Ошибка δ в определении вероятностей отличается от уровня квазистационарности ε, более того, она должна быть существенно меньше, иначе само понятие длины стационарности не будет иметь практического смысла. Оценим соответствующий минимальный объем текста.

Как известно (см., например, [10]) отклонение выборочного среднего значения $\bar{x}(N)$ стационарной случайной величины, определяемое по выборке объема N, от генерального среднего μ распределено асимптотически нормально с нулевым средним и стремящейся к нулю дисперсией σ^2 / N, где σ^2 есть дисперсия этой величины по гипотетической генеральной совокупности $f(i)$. Рассмотрим в качестве такой случайной величины количество n_i буквы «i» в тексте объема N. Тогда среднее значение этого количества n_i / N даст выборочную эмпирическую вероятность использования данной буквы. Значение σ_i представляет собой среднеквадратичное отклонение этой вероятности, а $\sigma_i / \sqrt{N}$ — отклонение среднего значения этой вероятности от значения по генеральной совокупности. Однако в условиях, когда генеральная дисперсия не известна, а оценивается только по выборочной дисперсии $s^2(N)$, следует рассматривать статистику

$$t = \sqrt{N-1}\frac{\bar{x}(N)-\mu}{s(N)}. \qquad (6)$$

Предположим, что выборочные отклонения частот использования букв с увеличением объемов выборки асимптотически нормальны. Тогда для каждой из n букв статистика (6) имеет распределение Стьюдента с $N - 1$ степенями свободы. Пренебрегая отличием N от $N - 1$, с доверительной вероятностью α получаем, что $|f_N(i) - f(i)|$ не превосходит $t_\alpha s / \sqrt{N}$, где t_α оценим сверху как α-квантиль предельного распределения Стьюдента с бесконечным числом степеней свободы. В частности, для $\alpha = 0{,}95$; 0,97; 0,99 соответствующие значения t_α приближенно равны 1,96; 2,20; 2,58 [10]. В качестве оценки выборочной дисперсии также возьмем максимальную по 32 буквам: $s = \max s_i$. Тогда из (6) получаем следующую оценку для минимального объема текста:

$$\sum_{i=1}^{n}\left|f_N(i) - f_{max}(i)\right| \le \frac{t_\alpha}{\sqrt{N}}\sum_{i=1}^{n} s_i \le \frac{t_\alpha n s}{\sqrt{N}}. \qquad (7)$$

Зададим число λ как величину интегральной близости $f_N(i)$ к некоторой гипотетической $f(i)$: $\sum_{i=1}^{n}\left|f_N(i) - f(i)\right| \le \lambda$. Тогда из (7) получаем, что если объем текста превосходит величину N_{min}, приближенно являющуюся решением уравнения

$$N = \left(\frac{t_\alpha n s(N)}{\lambda}\right)^2, \qquad (8)$$

то с вероятностью α его распределение на этом объеме близко к стационарному с точностью λ. Эмпирическая зависимость $s(N)$ была проанализирована для 100 произведений различных авторов и жанров (см. далее п. 3). Полагая $n = 32$ и $\lambda = 0{,}01$, получаем в результате численного решения уравнения (8), что для вышеуказанных значений α величины N_{min} соответственно равны примерно 8 тыс., 10 тыс. и 15 тыс. знаков. Для корректного сравнения текстов между собой их уровень стационарности на этих длинах должен во всяком случае превосходить уровень ошибки, с которой были определены эмпирические частоты, т. е. $\varepsilon > \lambda$.

Анализ текстов показал, что чем меньше ε, т. е. выше задаваемый уровень стационарности, тем больший разброс наблюдается в длинах $L(\varepsilon)$. Для $\varepsilon = 0{,}05$ всевозможные $L(0{,}05)$ заключены между 10 тыс.

и 40 тыс. знаков. Из оценки (8) следует, что соответствующие вероятности определены с точностью λ от 0,005 до 0,01. Ошибка в определении ε для каждой длины из диапазона 10÷40 тыс. знаков, обусловленная неточностью определения вероятностей, имеет величину порядка $\lambda^2 / (2\varepsilon)$, что не превосходит 0,001 (относительная ошибка менее 2% по сравнению с $\varepsilon = 0{,}05$). Это означает, что 0,05-стационарность определена достаточно корректно. Такой же вывод можно сделать и для 0,03-стационарности. В то же время разброс для $L(0{,}01)$ оказался очень велик, от 40 тыс. до почти 400 тыс. знаков. Поэтому, чтобы иметь относительную ошибку на уровне 2%, необходимо рассматривать тексты с длинами, большими, чем 250 тыс. знаков. В противном случае ошибка, вносимая неточностью в эмпирических вероятностях, может повлиять на статистические выводы о длине стационарности текста, и, в конечном счете, на критерий группировки текста.

Кроме того, анализ показал, что функции $L(\varepsilon)$ для разных произведений одного и того же автора могут существенно различаться, а для разных авторов, напротив, быть весьма близки. Поэтому $L(\varepsilon)$ не может служить опознавательным знаком отдельного писателя. В то же время стабилизация ПФР самих произведений позволяет сделать предположение, что ПФР различных авторов могут быть статистически различимы. Основанием для корректного сравнения авторских ПФР является 0,03-стабилизация всех произведений с объемом более 100 тыс. знаков на этом минимальном объеме независимо от объема самого произведения. Важно также и то, что установление достаточно высокого уровня стационарности происходит на объемах, существенно меньших тех, которые следуют из формулы (3).