Cryptography and Data Security (CSCE-544)

Major Addison Betances

Air Force Institute of Technology

27-March-2019

Overview

The Substitution Cipher

The goal of the substitution cipher is the encryption of text (as opposed to bits in modern digital systems). The idea is very simple: We substitute each letter of the alphabet with another one. Example:

$$\begin{aligned} \texttt{A} &\rightarrow \texttt{k} \\ \texttt{B} &\rightarrow \texttt{d} \\ \texttt{C} &\rightarrow \texttt{w} \\ &\cdots \end{aligned}$$

For instance, the pop group `ABBA` would be encrypted as `kddk`.

The Substitution Cipher

Let's look at another ciphertext:

`iq ifcc vqqr fb rdq vfllcq na rdq cfjwhwz hr bnnb hcc hwwhbsqvqbre hwq vhlq`

This does not seem to make too much sense and looks like decent cryptography. The Substitution Cipher has $26! \approx 4.0329E26$ possible keys, and it looks like it provides adequate protection against cryptanalysis.

However, the substitution cipher is not secure at all!

Brute-Force or Exhaustive Key Search

Definition

Let (x, y) denote the pair of plaintext and ciphertext, and let $K = \{k_1, ..., k_k\}$be the key space of all possible keys k_i . A brute-force attack now checks for every $k_i \in K$ if

$$d_{k_i} \stackrel{?}{=} x.$$

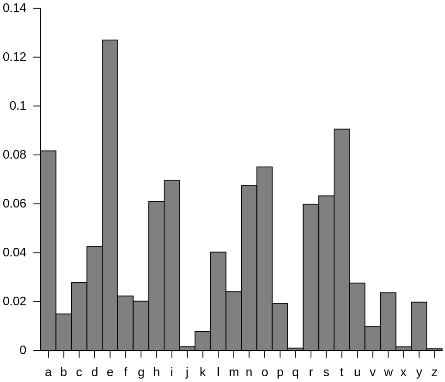
If the equality holds, a possible correct key is found; if not, proceed with the next key.

Letter Frequency Analysis

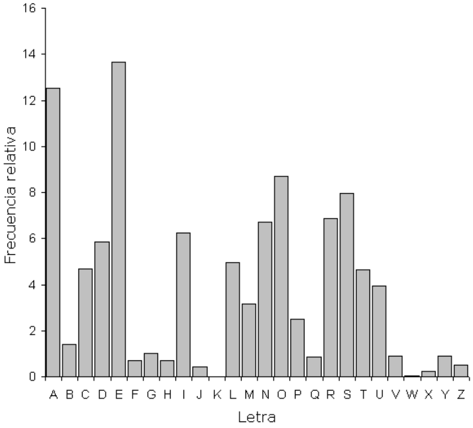
For practical attacks, the following properties of language can be exploited:

1. Determine the frequency of every ciphertext letter. The frequency distribution, often even of relatively short pieces of encrypted text, will be close to that of the given language in general.
2. The method above can be generalized by looking at pairs or triples, or quadruples, and so on of ciphertext symbols.
3. If we assume that word separators (blanks) have been found (which is only sometimes the case), one can often detect frequent short words such as `THE`, `AND`, etc.

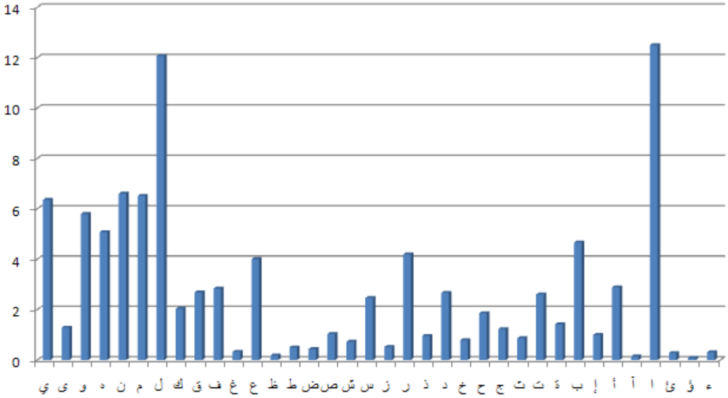
Relative letter frequencies of the English language



Letter frequencies of the Spanish language



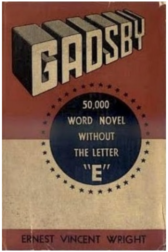
Arabic Letter Frequency distribution



Unusual Texts

Lipogram: is a kind of constrained writing or word game consisting of writing paragraphs or longer works in which a particular letter or group of letters is avoided—usually a common vowel, and frequently E, the most common letter in the English language. Larousse defines a lipogram as a "*literary work in which one compels oneself strictly to exclude one or several letters of the alphabet*".

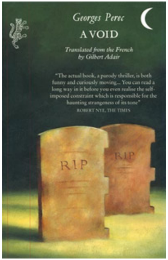
Unusual Texts



Gadsby is a 1939 novel by Ernest Vincent Wright written as a lipogram, which does not include words that contain the letter E. The plot revolves around the dying fictional city of Branton Hills, which is revitalized as a result of the efforts of protagonist John Gadsby and a youth group he organizes.

Though vanity published and little noticed in its time, the book is a favourite of fans of constrained writing and is a sought-after rarity among some book collectors. Later editions of the book have sometimes carried the alternative subtitle 50,000 Word Novel Without the Letter"E".

Unusual Texts



A Void, translated from the original French La Disparition (literally, "The Disappearance"), is a 300-page French lipogrammatic novel, written in 1969 by Georges Perec, entirely without using the letter e (except for the author's name), following Oulipo constraints.

How Many Key Bits Are Enough?

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0	
0	0	0	0	0	0	0	0	**0**
128	64	32	16	8	4	2	1	**1**

The discussion of key lengths for symmetric crypto algorithms is only relevant if a brute-force attack is the best known attack. The key lengths for symmetric and asymmetric algorithms are dramatically different. For instance, an 80-bit symmetric key provides roughly the same security as a 1024-bit RSA (RSA is a popular asymmetric algorithm) key.

Modular Arithmetic

Arithmetic within a discrete, finite set of elements (i.e. integers in a range). Example operations in this set :

- $1 + 2 = 3$ addition
- $4 \times 3 =$ multiplication
- $7 + 7 = 2 \mod 12$

Modular Arithmetic

Uses the following rule: Perform regular arithmetic and then divide the result by the number of elements to find the remainder.

Modification: element 12 becomes 0! Remainder is all we want

Modular Arithmetic

Consider the general form of a number a, $a \in \mathbb{Z}$. Where

$$a = q * m + r \qquad ,$$

denotes the division of a by m then,

$$a - r = q * m.$$

In this case, "m" is the modulus, "r" is the remainder, and "q" is the quotient.

Modular Arithmetic

- There are infinite r solutions. For example:
  - $15 \equiv 3 \mod 12$ since $12 - 0/12 = 1\ r\ 0$
  - $15 \equiv 15 \mod 12$ since $15 - 15/12 = 0\ 1r\ 0$
  - $15 \equiv -9 \mod 12$ since $15 - (-9)/12 = 2\ r\ 0$
- Remainder r is not constrained to the range of the finite set:
  - Means $r < 0$ and $r > m$ are valid
  - Also means that there are infinitely valid remainders
- All valid r form an equivalency class (infinite set) as follows: $\{\ldots, -21, -9, 3, 15, 27, \ldots\}$

Modular Arithmetic

All r in an equivalence class are equivalent.

1. Replacement of any number within a equivalency class can be done at any time in a calculation
2. More valuable - Modular reduction can be performed at any time in a calculation

$$
\begin{aligned}
3^8 &= 2 \mod 7 \\
3^8 &= 3^4 \times 3^4 \\
&= 81 \times 81 \\
&\equiv 4 \times 4 \mod 7 \qquad \text{Replace 81 with 4 in equiv class} \\
&\equiv 16 \mod 7 \\
&\equiv 2 \mod 7
\end{aligned}
$$

Modular Arithmetic

Which r to choose?

Choose smallest positive integer (naturally below m)
$a = q * m + r$, **where** $0 \leq r \leq m - 1$.

- $\mathbf{15 \equiv 3 \mod 12}$ $\quad$ **since** $\mathbf{12 - 0/12 = 1 \text{ r } 0}$
- $15 \equiv 15 \mod 12$ $\quad$ since $15 - 15/12 = 0\ 1r\ 0$
- $15 \equiv -9 \mod 12$ $\quad$ since $15 - (-9)/12 = 2\ r\ 0$