The Vigenère cipher.

This is a more elaborate simple substitution cipher, than the shift cipher (which we discussed before).

In a way, the Vigenère cipher is a generalization of the shift ciphers, with the shifts being more 'chaotic'.

The cipher is performed as follows:

(1) choose a keyword (responsible for the shifts);

(2) shift k^{th} letter of the plaintext by the number corresponding to $k \pmod{l}$'s letter of the keyword (l = length (keyword)).

Example. Consider the plaintext 'bitcoin' and keyword 'cat' = '3 1 20' (using the correspondence $A \leftrightarrow 1, B \leftrightarrow 2, \dots, Z \leftrightarrow 26$).

$$b \xmapsto{+3} e$$
$$i \xmapsto{+1} j$$
$$t \xmapsto{+20} n$$
$$c \xmapsto{+3} f$$
$$o \xmapsto{+1} p$$
$$i \xmapsto{+20} c$$
$$n \xmapsto{+3} k$$

The ciphertext is 'ejnfpck'.

Cryptanalysis of the Vigenère cipher.

At some point Vigenère-type ciphers were thought to be 'unbreakable', but this is far from being true. Our next goal is to get acquainted with the basic statistical tools in cryptanalysis.

Def-n. Let $s = a_1 a_2 \ldots a_n$ be a ~~text~~ (string of letters). The index of coincidence of s, denoted via $\mathrm{IndCo}(s)$, is the probability that two randomly chosen characters coincide.

Example. Let's take $s = aabbcc$. There are $\binom{6}{2} = \frac{6!}{2!4!} = 15$ pairs and 3 of them consist of coinciding letters ('aa', 'bb' and 'cc'). Hence, $\mathrm{IndCo}(s) = \frac{3}{15} = 0.2$.

Next we obtain a general formula. For this purpose, denote the number of occurences of the i^{th} letter of the alphabet in s by F_i, then

$$\mathrm{IndCo}(s) = \frac{\sum_{i=1}^{26} \binom{F_i}{2}}{\binom{n}{2}} = \frac{\sum_{i=1}^{26} F_i(F_i - 1)}{n(n-1)}.$$

Remark. If the string s consists of random characters (the occurence of each letter in s is equally likely), then

$$\mathrm{IndCo}(s) = \frac{26 \cdot \frac{n}{26}\left(\frac{n}{26}-1\right)}{n(n-1)} = \frac{\left(\frac{n}{26}-1\right)}{n-1} \xrightarrow[n\to\infty]{} \frac{1}{26} \approx 0.0385.$$

$$\left(F_i = \frac{n}{26}\right)$$

However, if s consists of an actual text (in English), then $\mathrm{IndCo}(s) \approx 0.0685$ (length(n) sufficiently large). Notice that this value is almost 2 times greater than for a random text!

Another important remark. $\mathrm{IndCo}(s)$ does not change (is invariant) under permutation letters, i.e. withstands any simple substitution.

Strategy to break the Vigenère cipher.

Step 1. Find the length of the keyword (ℓ).

Let's take a number k and test if $k=\ell$. In order to perform such a test, consider for every $1 \le i \le k$ the subtext s_i of s, where

$$s_i = a_i a_{i+k} a_{i+2k} \dots$$

$(s = a_1 a_2 \dots a_n)$ is the ciphertext

If our guess was correct ($k=\ell$), then $\mathrm{IndCo}(s_i)$ will be close to the one for a text in English (see the remark above). On the other hand, if $k\neq\ell$, then $\mathrm{IndCo}(s_i)$ will be close to the one for the random text.

To sum up, we take the average of the indices of coincidence and compare it to 0.0385 and 0.0685:

- if $\dfrac{\sum_{i=1}^{k} \mathrm{IndCo}(s_i)}{k} \sim 0.0685$, then $k=\ell$ is likely.
- if $\dfrac{\sum_{i=1}^{k} \mathrm{IndCo}(s_i)}{k} \sim 0.0385$, then $k=\ell$ is unlikely.

So we try $k=1,2,3$, etc. and, as the keyword has some 'adequate' length, at some point find ℓ.

Step 2. Next one needs to figure out the actual keyword.

Def-n. Let $s=a_1\ldots a_n$ and $t=b_1\ldots b_m$ be two strings of letters. The mutual index of coincidence of s and t is the probability that a randomly chosen symbol in s and a randomly chosen symbol in t will be the same.

$$\mathrm{MutIndCo}(s,t) := \frac{1}{nm}\sum_{i=1}^{26} F_i(s)F_i(t),$$

where $F_i(s)$ is the number of occurrences of the i^{th} letter in s and $F_i(t)$ in t.

Example. Let $s=$ 'Catalonia' and $t=$ 'Barcelona'. Then $n=\mathrm{len}(s)=9$, $m=\mathrm{len}(t)=9$ and we find

$$\mathrm{MutIndCo}(s,t) = \frac{1}{9\cdot 9}(\underset{'a'}{3\cdot 2} + \underset{'c'}{1\cdot 1} + \underset{'l'}{1\cdot 1} + \underset{'n'}{1\cdot 1} + \underset{'o'}{1\cdot 1}) = \frac{10}{81}.$$

So, they have something incommon.

(☆) Remark. If two strings s and t are encryptions of plaintext with the help of the same SSC, then the value of $\mathrm{MutIndCo}(s,t)$ will be larger. This index is an analogue of correlation in probability (that's my intuition at least).

Next we will use the MutIndCo to 'finish decrypting' Vigenere cipher.

Let $s^{(1)} = a_1 a_{1+\ell} a_{1+2\ell} \dots$
$s^{(2)} = a_2 a_{2+\ell} a_{2+3\ell} \dots$
$s^{(\ell)} = a_\ell a_{2\ell} a_{3\ell} \dots$

$P = p_1 p_2 p_3 \dots$ plaintext
$S = a_1 a_2 a_3 \dots$ ciphertext

$$\begin{array}{c} \beta_1 \quad \beta_2 \quad \beta_3 \quad\ \ \beta_\ell \\ \uparrow \quad\ \uparrow \quad\ \uparrow \quad\quad \uparrow \\ K = k_1 k_2 k_3 \dots k_\ell \end{array}$$ - key word

β_{12}, β_{13}, $\beta_{1\ell}$

$\beta_i :=$ index of letter k_i in the alphabet

$\beta_{1i} := \beta_1 - \beta_i$

Remark. Notice that $s^{(i)} = a_i a_{i+\ell} a_{i+2\ell} \ldots$ is encryption of $p^{(i)} = p_i p_{i+\ell} p_{i+2\ell} \ldots$ via shift cipher (the shift is by k_i with index β_i).

Next we compare the mutual indices of coincidence MutIndCo($s^{(1)}$, $s^{(i)}$ shifted by k_i) for different values of k_i between 0 and 25. As for $k_i = \beta_{1i}$ we get that both $s^{(1)}$ and $s^{(i)}$ are shifts of the corresponding parts of plaintext by β_1 (notice that $\beta_i + \beta_1 - \beta_i = \beta_1$), the correspon-ding index MutIndCo($s^{(1)}$, $s^{(i)}$ shifted by β_{1i}) will be the largest (see Remark (☆) on the previous page). This allows to find the values of $\beta_{12}, \beta_{13}, \ldots, \beta_{1\ell}$. Now it suffices to establish β_1 in order to decode the key word K. For this we simply try the 26 possible options.

Remark. The β_{1i}'s are found independently, so the complexity is $\mathcal{O}(\ell)$, not $\mathcal{O}(25^{\ell-1})$ as for a nested loop.