Théorie des codes - TP 1
ZZ3 F5 - Réseaux et Sécurité Informatique
Cryptographie Classique : Chiffrement de Vigenère

Blaise de Vigenère est né le 5 avril 1523 à Saint-Pourçain sur Sioule dans l'Allier. Politique, il fut secrétaire du duc de Nevers, secrétaire de la Chambre du roi, chargé de mission diplomatique. Il fut aussi un savant, auteur de nombreuses traductions, critique d'art et cryptographe. Même si l'on connaissait depuis fort longtemps les faiblesses de la cryptographie par substitution, il n'y eut pas entre César et le XVI$^{\text{ième}}$ siècle de véritable nouveau procédé cryptographique (sûr et facile). Vigenère fut l'initiateur d'une nouvelle façon de chiffrer les messages qui mit en échec les cryptanalystes trois siècles durant (1863). C'est en 1586 qu'il publie son *Traité des chiffres*, qui explique son nouveau chiffre [a].

Blaise de Vigenère

a. Disponible à la BNF : `http://gallica.bnf.fr/ark:/12148/bpt6k1040608n/f7.image`

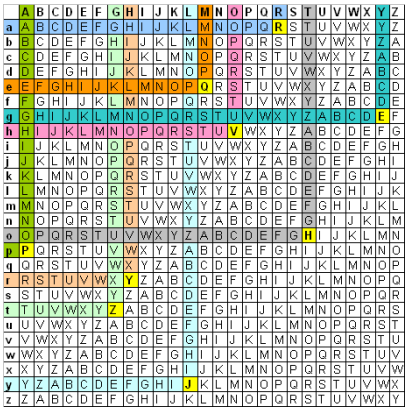
Chiffrement de Vigenère

Principe : Ce chiffrement introduit la notion de clef. Une clef se présente généralement sous la forme d'un mot ou d'une phrase. Pour pouvoir chiffrer notre texte, à chaque caractère nous utilisons une lettre de la clef pour effectuer la substitution. Évidemment, plus la clef sera longue et variée et mieux le texte sera chiffré.
A chaque lettre du texte clair on fait correspondre une lettre de la clef (la clef étant répétée autant de fois que nécessaire). La lettre du texte chiffré sera prise dans la colonne correspondante à la lettre du texte clair, et dans la ligne correspondante à la lettre de la clef.
En posant C le texte codé, T le texte et K la clef, on peut traduire ceci par la formule :

$$C = T + K \pmod{26}$$

Pour déchiffrer le message, il suffit de faire l'opération inverse : on prend la ligne correspondante à la lettre de la clef, et on la suit jusqu'à rencontrer le caractère codé ; la lettre décodée est alors la première de cette colonne. Ce qui se traduit par la formule :

$$T = C - K \pmod{26}$$



Exemple :
Dans cet exemple, nous chiffrons le texte T =`pythagore` à l'aide de la clef K =`algorythme`. Grâce au carré de Vigenère, on obtient C =`pjzvrehyq`. La première lettre $T[0]$ =`"p"` étant codée par un $K[0]$ =`"a"` devient un $C[0]$ =`"p"`, la deuxième lettre $T[1]$ =`"y"` codée par un $K[1]$ =`"l"` devient un $C[1]$ =`"j"`, même procédé pour toutes les autres lettres.

Dans ce TP, nous allons nous intéresser à la réalisation du chiffrement de Vigenère mais aussi à son décryptage ($\neq$ déchiffrement). Vous trouverez plus d'informations en annexe sur la cryptanalyse du chiffrement de Vigenère par le test de Friedman (appelé aussi test kappa).

1. Réalisez des fonctions permettant le chiffrement et le déchiffrement par substitution polyalphabétique de Vigenère.

2. Réalisez ensuite une fonction de décryptage du chiffrement de Vigenère par l'utilisation de l'indice de coïncidence et d'un test de χ^2 (cf. Annexe).

Appendix Cryptanalysis of the Vigenere Cipher

The Vigenere cipher was though to be completely unbreakable for hundreds of years, and indeed, if very long keys are used the vigenere cipher can be unbreakable. But if short keys are used, or if we have a lot of ciphertext compared to the key length, the vigenere cipher is quite solvable. Cryptanalysis [1] of the Vigenere cipher has 2 main steps :

— identify the period of the cipher (the length of the key) ;
— then find the specific key.

To identify the period we use a test based on the Index of Coincidence, to find the specific key we use the Chi-squared statistic. For the purposes of this explanation, we will try to break the following message :

```
vptnvffuntshtarptymjwzirappljmhhqvsubwlzzygvtyitarptyiougxiuydtgzhhvvmum
shwkzgstfmekvmpkswdgbilvjljmglmjfqwioiivknulvvfemioiemojtywdsajtwmtcgluy
sdsumfbieugmvalvxkjduetukatymvkqzhvqvgvptytjwwldyeevquhlulwpkt
```

The first thing to note is that there is no guarantee that the period of key that we find is the actual key used. If the message is very long we can be almost certain of being correct, but the methods provided here are approximate.

Finding the Period : The Vigenere cipher applies different Caesar ciphers to consecutive letters. If the key is `"PUB"`, the first letter is enciphered with a Caesar cipher with key 16 (`P` is the 16th letter of the alphabet), the second letter with another, and the third letter with another. When we get to the 4th letter, it is enciphered using the same cipher as letter 1. As a result, if we gather letters 1,4,7,10,... we should get a sequence of characters, all of which were enciphered using the same Caesar cipher. The sequence of characters 2,5,8,11,... and 3,6,9,12,... will also be enciphered with their own Caesar cipher. The exact sequence will of course depend on the period of the cipher i.e. the key length.

The Index of Coincidence (IC) is a statistical technique (William F. Friedman in 1920) that gives an indication of how English-like (or French-like) a piece of text is.

$$IC = \sum_{q=A}^{q=Z} \frac{n_q(n_q - 1)}{n(n-1)}$$

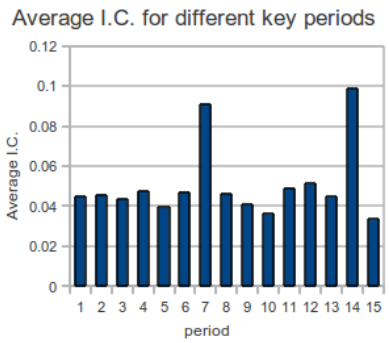
where n_q is the count of letter q and n is the total number of letters in the ciphertext. One of the useful properties of the technique is that the result of the IC does not change if you apply a substitution cipher to the text. This is because the IC is based on letter frequencies, and simple substitution ciphers do not modify the individual letter frequencies. If text is similar to english (or french) it will have an IC of around 0.0667 (or 0.0778), if the characters are uniformly distributed the IC is 0.0385. To determine the period of a Vigenere cipher we first assume the key length is 2. We extract the two sequences 1,3,5,7,... and 2,4,6,8,... from the ciphertext. For the example we are working with we get the following result (note that the IC is calculated using the whole sequences, not just the part shown)

```
                                                          IC
original:    vptnvffuntshtarptymjwzirappljmhhqvsubw...     0.049
if key were length 2:
sequence 1:  v t v f n s t r t m w i a p j h q s b ...     0.049
sequence 2:   p n f u t h a p y j z r p l m h v u w...     0.046
                                                 average:  0.048
if key were length 3:
sequence 1:  v  n  f  t  t  p  m  z  a  l  h  v  b ...     0.049
sequence 2:   p  v  u  s  a  t  j  i  p  j  h  s  w...     0.046
sequence 3:    t  f  n  h  r  y  w  r  p  m  q  u  ...     0.046
                                                 average:  0.047
```

1. Extract from :
http://practicalcryptography.com/cryptanalysis/stochastic-searching/cryptanalysis-vigenere-cipher/

This procedure of breaking up the ciphertext and calculating the IC for each subsequence is repeated for all the key lengths we wish to test. What we are most interested in is the average IC for a particular period, for the case of period = 2, the average IC is around 0.048. If you were to continue this procedure up to a period of 15 we get the following average I.C. values :

period	avg I.C.	period	avg I.C.
1 :	0.0449443523561	9 :	0.0407804755631
2 :	0.0457833618884	10:	0.0361152882206
3 :	0.0435885364312	11:	0.0491603339901
4 :	0.0474962292609	12:	0.0512663398693
5 :	0.0393612078978	13:	0.0446886446886
6 :	0.0471437059672	14:	0.0988487702773
7 :	0.0909922589726	15:	0.0334554334554
8 :	0.0461858974359		



We have 2 rows that have very high values of average IC. This indicates the key is probably of length 7, but could also be of length 14. Both of these probabilities should be tested.

Finding the Key : Since we now know the period is 7, we only have 7 Caesar ciphers to break, which is fairly easy. For this task we will use the Chi-squared statistic, which will compare the frequency distribution of our subsequences to the expected English (or French) frequency distribution.
The Chi-squared Statistic is a measure of how similar two categorical probability distributions are. If the two distributions are identical, the chi-squared statistic is 0, if the distributions are very different, some higher number will result. The formula for the chi-squared statistic is :

$$\chi^2(C,E) = \sum_{q=A}^{q=Z} \frac{(C_q - E_q)^2}{E_q}$$

where C_A is the count (not the probability) of letter `A`, and E_A is the expected count of letter `A`. Using every seventh letter starting with the first ; our first sequence is `VURZJUGRGGUGVGJQKEOAGUG KKQVWQP`. This is text all enciphered with the same Caesar cipher, we want to know what the key is. We try deciphering this sequence with each of the 25 possible Caesar ciphers, and compare the frequency distribution of the deciphered text with the frequency distribution of English (or of French) for each key. If we perform this, we get 26 values for the Chi-squared statistic. The correct key will correspond to the deciphered text with the lowest Chi-squared statistic (we hope, due to the statistical nature of the problem it may be the second or third lowest value).

This means our first Vigenere key letter is `"C"` (`"A"`=0, `"B"`=1, `"C"`=2,...). We have to repeat this procedure for each of the 7 key letters. If we continue this procedure of finding the keys corresponding to the Chi-squared minima, we get the sequence 2,8,0,7,4,17,18. This spells out `"CIAHERS`, which is wrong. This goes to show you can't rely on the technique fully unless very long ciphertexts are available. The correct key was `"CIPHERS"`, and indeed the Chi-square test had two very low values for that subsequence. Unfortunately the incorrect one was slightly lower. We show the results of this procedure here :

key	deciphered sequence	chi-sq
0	VURZJUGRGGUGVGJQKEOAGUGKKQVWQP	595.42
1	UTQYITFQFFTFUFIPJDNZFTFJJPUVPO	466.86
2	TSPXHSEPEESETEHOICMYESEIIOTUON	41.22
3	SROWGRDODDRDSDGNHBLXDRDHHNSTNM	67.73
4	RQNVFQCNCCQCRCFMGAKWCQCGGMRSML	642.37
5	QPMUEPBMBBPBQBELFZJVBPBFFLQRLK	451.49
6	POLTDOALAAOAPADKEYIUAOAEEKPQKJ	121.97
7	ONKSCNZKZZNZOZCJDXHTZNZDDJOPJI	2441.20
8	NMJRBMYJYYMYNYBICWGSYMYCCINOIH	190.46
9	MLIQALXIXXLXMXAHBVFRXLXBBHMNHG	1142.90
10	LKHPZKWHWWKWLWZGAUEQWKWAAGLMGF	358.87
11	KJGOYJVGVVJVKVYFZTDPVJVZZFKLFE	962.13
12	JIFNXIUFUUIUJUXEYSCOUIUYYEJKED	354.18
13	IHEMWHTETTHTITWDXRBNTHTXXDIJDC	241.97
14	HGDLVGSDSSGSHSVCWQAMSGSWWCHICB	107.36
15	GFCKUFRCRRFRGRUBVPZLRFRVVBGHBA	136.40
16	FEBJTEQBQQEQFQTAUOYKQEQUUAFGAZ	1801.65
17	EDAISDPAPPDPEPSZTNXJPDPTTZEFZY	531.22
18	DCZHRCOZOOCODORYSMWIOCOSSYDEYX	247.66
19	CBYGQBNYNNBNCNQXRLVHNBNRRXCDXW	377.60
20	BAXFPAMXMMAMBMPWQKUGMAMQQWBCWV	489.12
21	AZWEOZLWLLZLALOVPJTFLZLPPVABVU	815.45
22	ZYVDNYKVKKYKZKNUOISEKYKOOUZAUT	648.33
23	YXUCMXJUJJXJYJMTNHRDJXJNNTYZTS	1476.11
24	XWTBLWITIIWIXILSMGQCIWIMMSXYSR	279.93
25	WVSAKVHSHHVHWHKRLFPBHVHLLRWXRQ	158.53

Wrapping up : As shown above, statistical techniques can give you wrong answers. To get around this you may have to try decrypting the ciphertext with each of several likely candidates to find the true key.

Appendix Letter Frequencies

English letter frequencies (%):

A :	8.55	K :	0.81	U :	2.68
B :	1.60	L :	4.21	V :	1.06
C :	3.16	M :	2.53	W :	1.83
D :	3.87	N :	7.17	X :	0.19
E :	12.10	O :	7.47	Y :	1.72
F :	2.18	P :	2.07	Z :	0.11
G :	2.09	Q :	0.10		
H :	4.96	R :	6.33		
I :	7.33	S :	6.73		
J :	0.22	T :	8.94		

German letter frequencies (%):

A :	6.34	K :	1.50	U :	3.76
B :	2.21	L :	3.72	V :	0.94
C :	2.71	M :	2.75	W :	1.40
D :	4.92	N :	9.59	X :	0.07
E :	15.99	O :	2.75	Y :	0.13
F :	1.80	P :	1.06	Z :	1.22
G :	3.02	Q :	0.04	Ä :	0.54
H :	4.11	R :	7.71	Ö :	0.24
I :	7.60	S :	6.41	Ü :	0.63
J :	0.27	T :	6.43	ß :	0.15

French letter frequencies (%):

A :	7.60	G :	1.18	Q :	0.85
À :	0.43	H :	0.93	R :	6.86
Â :	0.05	I :	7.21	S :	7.98
Æ :	0.00	Î :	0.04	T :	7.11
B :	0.96	Ï :	0.01	U :	5.55
C :	3.39	J :	0.30	Ù :	0.02
Ç :	0.05	K :	0.16	Û :	0.02
D :	4.08	L :	5.86	Ü :	0.00
E :	14.47	M :	2.78	V :	1.29
É :	2.43	N :	7.32	W :	0.08
È :	0.42	O :	5.39	X :	0.43
Ê :	0.13	Ô :	0.05	Y :	0.34
Ë :	0.00	OE:	0.02	Z :	0.10
F :	1.12	P :	2.98		

Spanish letter frequencies (%):

A :	12.50	K :	0.08	T :	4.42
B :	1.27	L :	5.84	U :	4.00
C :	4.43	M :	2.61	V :	0.98
D :	5.14	N :	7.09	W :	0.03
E :	13.24	Ñ :	0.22	X :	0.19
F :	0.79	O :	8.98	Y :	0.79
G :	1.17	P :	2.75	Z :	0.42
H :	0.81	Q :	0.83		
I :	6.91	R :	6.62		
J :	0.45	S :	7.44		