Le Python, c'est bon

Cours 6 : Cryptographie et décryptage

1 Code de César

Cette façon de coder des messages très ancienne, César l'aurait utilisée, repose sur une *substitution monoalphabétique* particulière, c'est à dire remplacer une lettre de l'alphabet par une autre.

Le principe en est simple, chaque lettre de l'alphabet possède une valeur de 1 à 26 (A vaut 1, B vaut 2, ..., Z vaut 26), que l'on code en lui rajoutant une constante modulo 26.

On a alors coutume de considérer que la clé permettant de coder le texte est la lettre correspondant au codage de 'A'. Ainsi, l'ajout de la constante 4 correspond à la clé 'E', comme le montre la table 1.

Table 1: Correspondance des lettre pour la clé 'E'

Lettre initiale	A	B	C	D	E	F	G	H	I	J	K	L	M
Lettre codée	E	F	G	H	I	J	K	L	M	N	O	P	Q
Lettre initiale	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Lettre codée	R	S	T	U	V	W	X	Y	Z	A	B	C	D

Le texte "si c'est jeudi c'est qu'il y a des frites" devient alors "WM G'IWX NIYHM G'IWX UY'MP C E HIW JVMXIW" avec 'E' comme clé.

1.1 Coder et décoder un texte

Dans tout ce TP, on considérera que nos textes à coder ne sont composés que de lettres de l'alphabet et d'espaces.

1.1.1 Travail préparatoire

Implémentez une fonction rendant un texte en majuscule. On pourra en première approche ne pas traiter les caractères accentués.

1.1.2 Codage

Implémentez une fonction qui, à partir d'un texte et d'un caractère rend le texte codé selon la méthode de César (tout ce qui n'est pas une lettre de l'alphabet en majuscule ne sera pas codé).

1.1.3 Décodage

Implémentez une fonction qui, à partir d'un texte dont les lettres sont codé selon la méthode de César et la clé ayant permis de le coder, rend le texte originel.

1.2 Décrypter un texte

Décrypter un texte est l'art de décoder un texte sans connaître la clé qui a permis de le coder. Décrypter un texte chiffré par le code de césar peut se faire aisément, il suffit d'essayer toutes les clés (il y en a 25 possibles).

Cette solution, certes viable, est cependant totalement dépourvue d'élégance et sera inapplicable pour des codes ayant un grand nombre de clés potentielles.

Une idée pour ne pas avoir à essayer toutes les possibilités est d'utiliser le fait que chaque lettre est codée par une autre de façon unique : les fréquences d'apparitions des lettres sont inchangées. Comme l'utilisation d'une lettre plutôt qu'une autre n'est pas équiprobable (en Français, le 'E' est la lettre la plus utilisée), la lettre revenant le plus grand nombre de fois dans le message codé a de forte chance d'être la lettre la plus fréquente dans la langue utilisée pour écrire le message.

La table 2 donne les pourcentages d'utilisation de chaque lettre pour la langue française.

Table 2: Fréquences d'apparitions des lettres en Français

E	17.76	O	5.34	B	0.80
S	8.23	D	3.60	H	0.64
A	7.68	C	3.32	X	0.54
N	7.61	P	3.24	Y	0.21
T	7.30	M	2.72	J	0.19
I	7.23	Q	1.34	Z	0.07
R	6.81	V	1.27	K	0.00
U	6.05	G	1.10	W	0.00
L	5.89	F	1.06		

Si le message est court, les fréquences théoriques et ceux du message risquent de différer, mais cela permet tout de même de se donner une idée de la clé. Pour le texte "si c'est jeudi c'est qu'il y a des frites", les différentes fréquences sont : 17% pour 'E' et 'S', 13% pour 'I', 10% pour T, etc.

1.2.1 Analyse des fréquences

Implémentez une fonction rendant les différentes fréquences de caractères d'un texte donné.

1.2.2 Hack da world

Quels textes se cachent derrière :

- NQX TSY IJX HMFUJFZC WTSIX ANAJ QF GWJYFLSJ NQX TSY IJX HMFUJFZC WTSIX ANAJ QJX GWJYTSX
- YB UJQYJ KD FUJYJ DQLYHU GKY D QLQYJ ZQCQYI DQLYWKU XEXU XEXU

2 Code de Vigenere

Le code de Vigenère (1523-1596) permet de casser le point faible du code de César, à savoir l'analyse des fréquences.

Il est cependant basé sur le même principe que celui de César, mais il utilise des clés plus longues. Chaque lettre peut alors être codée de plusieurs façons, rendant caduque l'analyse des fréquences.

Par exemple, si la clé est 'bcd', toutes les lettres en positions 3*k seront codés selon le code de César avec 'b' comme clé, toutes les lettres en positions 3*k+1 seront codées avec la clé 'c' et toutes les lettres en positions 3*k+2 avec la clé 'd'. Ainsi, en codant le texte "si c'est jeudi c'est qu'il y a des frites" avec la clé "RAVIOLI" on obtient "JI X'MGE RVUYQ Q'PAK QP'QZ J I UEN NFTBVS" comme le montre la table 3.

Table 3: Code de Vigenère, exemple

Texte	S	I	C	E	S	T	J	E	U	D	I	C	E	S	T
Clé	R	A	V	I	O	L	I	R	A	V	I	O	L	I	R
Code	J	I	X	M	G	E	R	V	U	Y	Q	Q	P	A	K
Texte	Q	U	I	L	Y	A	D	E	S	F	R	I	T	E	S
Clé	A	V	I	O	L	I	R	A	V	I	O	L	I	R	A
Code	Q	P	Q	Z	J	I	U	E	N	N	F	T	B	V	S

Plus la clé est longue, moins il est facile de le décrypter. Un cas particulier de cette méthode où la longueur de la clé est exactement la longueur du texte à coder est appelée *code de Vernam*. C'est le seul code dont sont inviolabilité est prouvée. Il semblerait que c'est ce type de code qui était utilisé pour feu le téléphone rouge.

2.1 Coder et décoder un texte

Modifiez les fonctions permettant de coder et décoder un texte selon la méthode de César pour qu'elles permettent de coder et décoder un texte sans espaces selon la méthode de Vigenère.

2.2 Décrypter un texte

Si l est la longueur de la clé utilisée pour coder un texte selon la méthode de Vigenère, on peut remarquer que toutes les lettres aux positions $k * l + p$ où est fixé sont codées par la même lettre. Une analyse des fréquences pour les différents p (de 0 à $l-1$) nous permet alors, de même que pour le code de César de deviner la clé utilisée.

Le principal problème pour décrypter un texte encodé selon la méthode de Vigenère est donc de trouver la longuer de la clé.

En 1920, Wolfe Friedman introduit pour cela la notion *d'indice de coïncidence mutuel* (IC). Cet indice est la probabilité que deux lettres choisies aléatoirement dans un texte soient identiques. Si on se donne un texte de n lettres possédant n_1 lettres 'A', n_2 lettres 'B', ..., n_{26} lettres 'Z', la probabilité de tirer 2 'A' parmi les n lettre du texte est alors :

$$\frac{C_2^{n_1}}{C_2^n} = \frac{n_1(n_1-1)}{n(n-1)}$$

.

De là, la probabilité de tirer deux lettres identiques est égal à la somme des probabilités de tirer deux lettres particulières, et donc cette probabilité (IC) est égale à :

$$\sum_{1 \leq i \leq 26} \frac{n_i(n_i-1)}{n(n-1)}$$

Pour la langue française, IC est environ égal à 0.074. Attention cependant, si le texte est petit, les variations de l'IC vont être importante.

2.2.1

Modifiez la fonction permettant de calculer les fréquences pour que pour un entier p donné, elle rende les fréquences associées aux lettres en positions k*p, k*p+1, k*p+2, ..., k*p + (p-1).

2.2.2

Implementez une fonction qui, à partir d'un texte sans espace et un entier p, rend l'IC associé à une longueur de clé égale à p.

Pour cela, on calculera l'IC pour les lettres des positions $k*p+q$ (avec $0 \leq q \leq p-1$) que l'on nommera q-IC, et on effectuera la moyenne $1/p * \sum_{0 \leq q \leq p-1}$ q-IC.

2.2.3 Hack da Universe

Quels textes se cachent derrière :

- N SJB TLG DI JLIKM IE ELIRO XV LSYBV UE WOEOYP OL KLLH AM ZZWJ TID OLBVPG KWYE DIMXD O JM NPHVZ WFF DWM N SJB TLG DI JLIKM E XCZ AM U SEBIYRJ BSFH RCXZII LI XCZ PIWZF PIWZZ B ID O
- WM LR MCAG DRNU DT MRIP U XSK PQMK TFULGNRJ TQMCOLRU AE EJT DWTU CE XSNRZEP DX GVNVAE TIOWTTDFUT VXS CE NWGDVMCAG MRTKF BL IEUKXMSLG S ZAZNUTHUIG E WLT UITW LI TA EJTIET L SBMV PCK E ADOWJ

3 RSA

Les codes suent dans les sections précédentes sont des codes dit *symétriques*. En effet, les clés permettant de coder et de décoder le texte sont les mêmes.

Le code RSA (les initiales de ses "inventeurs", R. Rivest, A. Shamir et L. Adleman) quant à lui est un code *asymétrique*, puisque les clés servant à coder (clé publique) et à décoder (clé privée) sont différentes.

Les propriétés mathématiques soutenant ce système de code seraient trop longues à expliquer ici, nous nous contenterons donc d'en exposer les principes.

- on se donne trois nombres premiers x,y et s avec $s = \max\{x, y, s\}$,
- on note $N = x * y$,
- on calcule ensuite un entier p tel que $p * s$ modulo $(x-1)(y-1) = 1$ (on peut prouver que cet entier existe).

De là, on peut prouver que pour tout entier M, si on note $C = M^p$ modulo N, on a $M = C^s$ modulo N (ce qui correspond au fait que M^{ps} modulo $N = M$). La clé publique est alors le couple (N, p) (elle permet de coder le message M) et la clé privée le couple (N, s) (qui permet de décoder le message).

La confidentialité du message réside dans le fait que la factorisation de N n'est pas un problème aisé lorsque N est très grand.

3.1 Coder et décoder un texte

3.1.1 Travail préliminaire

Écrire une fonction permettant de rendre tous les nombres premiers inférieurs à un nombre n

3.1.2 Les nombres de RSA

Écrire une fonction permettant de rendre les entiers x,y,N,s et p. On se restreindra ici aux nombre à 2 chiffres.

3.1.3 Exponentiation

Implémentez une fonction permettant d'élever un nombre n à la puissance p modulo q.

3.1.4 Codage

Encoder un texte selon la méthode RSA peut se faire comme suit. On commence par transformer le texte à coder en entier. Pour cela, on peut utiliser le codage où chaque lettre est codée par deux chiffres, selon sa position dans l'alphabet. Ainsi, le message "CESSEZ LE FEU" sera codé : 03051919052600120500060521 (C a pour code 03, E à pour code 05, l'espace à un code égal à 00).

On découpe ensuite ce nombre en morceaux, chaque morceaux ayant autant de chiffres que N. Prenons par exemple x=47, y=79 et s=97. On trouve alors que N=3713 et p=37. Il faut donc découper notre mot en paquets de 4 chiffres. Notre mot n'ayant que 26 chiffres, on lui rajoute un espace à la fin. Comme $0305^{37} = 2496$ modulo 3713, $1919^{37} = 0655$ modulo 3713, ... Notre mot chiffré est : 24960655207609651589294518 82.

Écrire une fonction permettant de coder un texte selon la méthode RSA en connaissant N et p.

3.1.5 Décodage

Décoder un texte se fait de la façon duale au codage. On découpe le texte codé en autant de paquet que nécessaire (chaque paquet ayant un nombre de chiffre égal au nombre de chiffres de N), et on élève chaque paquet à la puissance s modulo N.

Écrire une fonction permettant de décoder un texte codé selon la méthode RSA.