Introduction à la cryptographie
TD2 – Monoalphabétique automatique, Vigenère et Vernam

Cécile Pierrot

20 Octobre 2020

1 Substitution monoalphabétique

Le *chiffre de César* (du nom de l'empereur) consiste à remplacer les lettres du message clair par la lettre se trouvant trois positions plus loin. Par exemple, la lettre `A` est remplacée par `D`, la lettre `B` est remplacée par `E`, etc. Schématiquement, on peut représenter ce chiffre par

Lettre en clair	A	B	C	D	E	F	G	H	I	J	K	L	M
Lettre chiffrée	D	E	F	G	H	I	J	K	L	M	N	O	P

Lettre en clair	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Lettre chiffrée	Q	R	S	T	U	V	W	X	Y	Z	A	B	C

Dans le chiffre de César, la clé est le nombre de décalages à appliquer pour obtenir l'alphabet chiffré. Il y a donc 25 possibilités.
Le chiffre de César peut être généralisé en considérant toutes les permutations possibles de l'alphabet. Il y en a 26!, ce qui est plus que considérable. Cependant, quelle que soit la permutation choisie, aucune n'est particulièrement robuste. Et ce résultat est connu depuis le IX$^{\text{ème}}$ siècle : en effet, le traducteur arabe Al-Kindi (801-873) a établi que chaque langue avait une fréquence d'utilisation de ses symboles très caractéristique. Par exemple, en français, on obtient les fréquences suivantes [1] :

Lettre	A	B	C	D	E	F	G	H	I
Fréquence	8,40 %	1,06 %	3,03 %	4,18 %	17,26 %	1,12 %	1,27 %	0,92 %	7,34 %

Lettre	J	K	L	M	N	O	P	Q	R
Fréquence	0,31 %	0,05 %	6,01 %	2,96 %	7,13 %	5,26 %	3,01 %	0,99 %	6,55 %

Lettre	S	T	U	V	W	X	Y	Z
Fréquence	8,08 %	7,07 %	5,74 %	1,32 %	0,04 %	0,45 %	0,30 %	0,12 %

On peut également constater que certains groupe de deux lettres (bigrammes ou digrammes) sont plus fréquents que d'autres, ainsi que des groupes de trois lettres (trigrammes). Les tableaux suivants donnent le nombre d'occurrences des bigrammes et trigrammes les plus fréquents dans un corpus de 100 000 lettres :

Bigramme	ES	DE	LE	EN	RE	NT	ON	ER	TE	EL
Nb. d'occurrences	3318	2409	2366	2121	1885	1694	1646	1514	1484	1382

Bigramme	AN	SE	ET	LA	AI	IT	ME	OU	EM	IE
Nb. d'occurrences	1378	1377	1307	1270	1255	1243	1099	1086	1056	1030

1. `http://www.apprendre-en-ligne.net/crypto/stat/francais.html`

Trigramme	ENT	LES	EDE	DES	QUE	AIT	LLE	SDE	ION	EME
Nb. d'occurrences	900	801	630	609	607	542	509	508	477	472

Trigramme	ELA	RES	MEN	ESE	DEL	ANT	TIO	PAR	ESD	TDE
Nb. d'occurrences	437	432	425	416	404	397	383	360	351	350

En s'appuyant sur ces résultats, il est possible et même facile de décrypter à peu près n'importe quel message chiffré par cette technique, appelée *chiffrement par substitution monoalphabétique*, si le texte est suffisamment long.
La feuille de calcul `td2.ods` vous propose un défi. Il faut déchiffrer le texte donné simplement en s'appuyant sur les fréquences des lettres, bigrammes et trigrammes, ainsi que sur votre bon sens (vous serez amenés à faire des hypothèses qui, si elles sont fausses, produiront des anomalies dans le texte clair).
La feuille de calcul présente un premier tableau composé de 4 lignes :

— la première ligne correspond à l'alphabet chiffré (en rouge) ;
— la deuxième ligne, initialement vide, doit être remplie par vos soins : il vous faudra y indiquer les lettres de l'alphabet en clair que vous aurez identifiées (N.B. seule cette ligne est modifiable dans la feuille de calcul) ;
— les deux lignes suivantes vous indiquent le nombre d'occurrences de chaque lettre dans le message chiffré, ainsi que le pourcentage correspondant.

Les deux tableaux suivants vous donnent les mêmes informations pour les 20 bigrammes et trigrammes les plus fréquents. Les versions en clair des bigrammes et trigrammes sont automatiquement calculées en fonction des lettres renseignées dans l'alphabet en clair.
Le tableau « Lettres restantes », sur la droite, indique les lettres de l'alphabet en clair qui n'ont pas encore été associées à une lettre de l'alphabet chiffré.
Enfin, les deux derniers tableaux contiennent le texte chiffré et le texte en clair, respectivement. Initialement rempli avec des tirets, ce dernier se remplira automatiquement lorsque vous identifierez une lettre en clair dans le premier tableau.

Question 1. Retrouvez le message en clair.

2 Chiffre de Vigenère

L'analyse fréquentielle est une attaque imparable contre les substitutions monoalphabétiques. Au XVI$^{\text{ème}}$ siècle, Blaise de Vigenère inventa un système polyalphabétique nécessitant un mot clé. Le système polyalphabétique est un carré (le carré de Vigenère) de 26 colonnes sur 26 lignes. Chaque ligne est en fait un décalage de César :

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

La première ligne et la première colonne servent également « d'indice » pour le chiffrement et le déchiffrement. Ainsi, pour chiffrer la lettre m avec la lettre k de la clé, on regarde la ligne d'indice k du carré de Vigenère : la lettre chiffrée est donnée dans la colonne d'indice m.
Exemple de chiffrement avec le mot clé `TARTE` :

Message en clair (colonne)	M	A	C	H	I	N	E	...
Clé (ligne)	T	A	R	T	E	T	A	...
Message chiffré (carré[ligne, colonne])	F	A	T	A	M	G	E	...

Notez que la clé est répétée autant de fois que le nécessite la longueur du message. De plus, le principe est complètement inversible : connaissant la clé (donc les lignes) et le message chiffré (la valeur dans le carré de Vigenère), retrouver la colonne est immédiat.

Question 1. Déchiffrez `KNFLGJIHMVCBNL` avec le mot clé `CACHE`.

En 1863, un officier prussien à la retraite, Friedrich Wilhelm Kasiski, découvre la faille. Il remarque que la répétition de la clé fait apparaître des « motifs » qui se répètent dans le texte

chiffré. Ces motifs correspondent à des groupes de lettres du message en clair qui ont été chiffrés avec les mêmes lettres de la clé. Cette coïncidence va permettre de déduire la longueur de la clé.
La démarche est la suivante :

1. identifier des groupes de lettres qui se répètent ;
2. calculer les écarts entre les occurrences des groupes de lettres ;
3. calculer le PGCD de ces écarts.

Ce PGCD est très probablement la longueur de la clé (il arrive que ce ne soit pas le cas, mais c'est très inhabituel). Mais que faire avec la longueur de la clé ?
Prenons un exemple : supposons que la longueur de la clé est $\ell = 5$. Cette clé peut s'écrire comme la chaîne de 5 lettres $k_1k_2k_3k_4k_5$. Comme on connaît le principe de chiffrement, il est facile de voir que la 1ère lettre du message en clair a été chiffrée avec la lettre k_1 de la clé, tout comme la 6ème lettre, la 11ème, la 16ème, ainsi que les autres lettres d'indice $5k + 1$.
Toutes ces lettres ont subi le même chiffrement avec la lettre k_1, ce qui revient à dire qu'elles ont subi un chiffrement monoalphabétique. Par conséquent, on peut appliquer l'analyse fréquentielle qui mettra sûrement en évidence une lettre chiffrée beaucoup plus utilisée que les autres, et qui correspondra à la lettre en clair `E`.
Si l'on connaît la lettre en clair (le `E`) et la lettre chiffrée correspondante, retrouver la valeur de k_1 est alors immédiat. De manière générale, on peut identifier chaque lettre k_i de la clé grâce à l'ensemble des lettres chiffrées d'indice $5k + i$.
Dans le classeur `td2.ods`, vous trouverez une feuille de calcul intitulée « 2. Vigenère », dont l'objectif est de retrouver la clé et de décrypter un message chiffré par ce système. Ici aussi, seules les cases en bleu clair sont modifiables.
Dans cette feuille de calcul, les deux gros tableaux en bas donnent le texte chiffré et le texte en clair, respectivement. Le contenu de ce dernier sera révélé au fur et à mesure de la découverte des caractères du mot clé. Ces caractères du mot clé sont à renseigner dans le tableau « Mot clé », situé juste au dessus du texte chiffré.
Le tableau précédent permet d'indiquer le nombre de caractères ℓ de la clé. La deuxième ligne de ce tableau, intitulée « Sélectionner lettres d'indice », permet de ne considérer que les lettres d'indice $k\ell + i$ du texte chiffré, lorsqu'une valeur i (pour $1 \leq i \leq \ell$) est renseignée dans la case correspondante.
Le tableau situé juste au dessus permet de calculer le PGCD des valeurs renseignées dans sa première ligne.
Les trois tableaux situés en haut de la feuille de calcul donnent des statistiques sur le texte chiffré (nombre d'occurrences des lettres, ainsi que des 10 bigrammes et trigrammes les plus fréquents). Il est à noter que, lorsqu'un indice $1 \leq i \leq \ell$ est sélectionné, le premier tableau donne les statistiques uniquement pour les lettres d'indice $k\ell + i$.
Enfin, le graphe des fréquences, sur la droite, donne une représentation graphique de la fréquence de chaque lettre dans le texte chiffré. De la même manière, si un indice i est sélectionné, les fréquences sont mesurées uniquement pour les lettres d'indice $k\ell + i$.

Question 2. Retrouvez le message en clair grâce à la méthode de Kasiski.

3 Chiffre de Vernam : le masque jetable

La méthode du *masque jetable* (ou *chiffre de Vernam*, ou encore *one-time pad* en anglais) consiste à chiffrer un message en clair M avec une clé (ou *masque*) K choisie de la manière suivante :

— la clé doit être une suite de caractères de même longueur que le message M ;

— les caractères de la clé doivent être choisis de manière totalement aléatoire ;
— la clé ne doit jamais être réutilisée pour chiffrer un autre message.

Dans le cadre de cet exercice, les messages en clair et chiffrés ainsi que les clés seront des suites de lettres de l'alphabet. Nous considérons ainsi l'encodage des 26 lettres de `A` à `Z` comme les entiers de 0 à 25 de la manière suivante :

Lettre	A	B	C	D	E	F	G	H	I	J	K	L	M
Code	0	1	2	3	4	5	6	7	8	9	10	11	12

Lettre	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Code	13	14	15	16	17	18	19	20	21	22	23	24	25

On peut alors définir l'opération $\oplus$ sur les lettres comme l'addition dans $\mathbb{Z}/26\mathbb{Z}$ (c'est-à-dire, modulo 26 : tant que la somme est supérieure ou égale à 26, on soustrait 26) des codes correspondants. Ainsi, par exemple :

$$\begin{aligned} \texttt{C} \oplus \texttt{P} &\equiv (\ 2 + 15) \bmod 26 = 17 \equiv \texttt{R}, \\ \texttt{U} \oplus \texttt{K} &\equiv (20 + 10) \bmod 26 = \ 4 \equiv \texttt{E}. \end{aligned}$$

On peut aussi définir l'opération inverse, notée $\ominus$:

$$\begin{aligned} \texttt{W} \ominus \texttt{G} &\equiv (22 - 6) \bmod 26 = 16 \equiv \texttt{Q}, \\ \texttt{D} \ominus \texttt{J} &\equiv (\ 3 - 9) \bmod 26 = 20 \equiv \texttt{U}. \end{aligned}$$

Étant donné un message en clair M de ℓ caractères, noté $M = (m_1, m_2, \ldots, m_\ell)$, ainsi qu'une clé $K = (k_1, k_2, \ldots, k_\ell)$ de même longueur, le message chiffré correspondant est la chaîne de ℓ caractères $C = E_K(M) = (c_1, c_2, \ldots, c_\ell)$ avec $c_i = m_i \oplus k_i$ pour tout $1 \leq i \leq \ell$. Pour plus de simplicité, on pourra noter cette opération

$$C = E_K(M) = M \oplus K.$$

Question 1. Chiffrez le message $M = \texttt{CRYPTO}$ avec la clé $K = \texttt{VSDQLK}$.

Question 2. Décrivez l'opération de déchiffrement : étant donné un message chiffré C et une clé K, comment retrouver le message en clair correspondant, c'est-à-dire le message M tel que $C = E_K(M)$? Le chiffre de Vernam est-il un algorithme à clé secrète ou à clé publique ?

Question 3. Déchiffrez le chiffré $C = \texttt{DSVSWA}$ à l'aide de la clé $K = \texttt{LOTBSH}$.

Question 4. Déchiffrez ce même chiffré C à l'aide de la clé $K' = \texttt{OYUHOY}$. Commentez.

Question 5. Soient C et M deux suites de ℓ caractères chacune. Montrez qu'il existe toujours une clé K telle que $C = E_K(M)$. Que pouvez-vous en conclure sur la probabilité qu'un attaquant parvienne à déchiffrer un chiffré donné sans connaître la clé ?

Supposons que deux messages en clair M_1 et M_2 de même longueur aient été chiffrés avec la même clé K, et que Eve ait intercepté les deux chiffrés correspondants $C_1 = E_K(M_1)$ et $C_2 = E_K(M_2)$. Supposons de plus que Eve soit parvenue par un autre biais à deviner M_1.

Question 6. Montrez qu'elle peut alors retrouver M_2. À la lumière de cette propriété indésirable, commentez le nom de *masque jetable* donné à ce chiffrement.

Question 7. Discutez de la question de la distribution de la clé K : qui doit posséder la clé ? l'émetteur ? le destinataire ? Mettez cela en regard de la taille de la clé (pour l'envoi un long message par exemple) et discutez du problème que cela soulève.