SR06

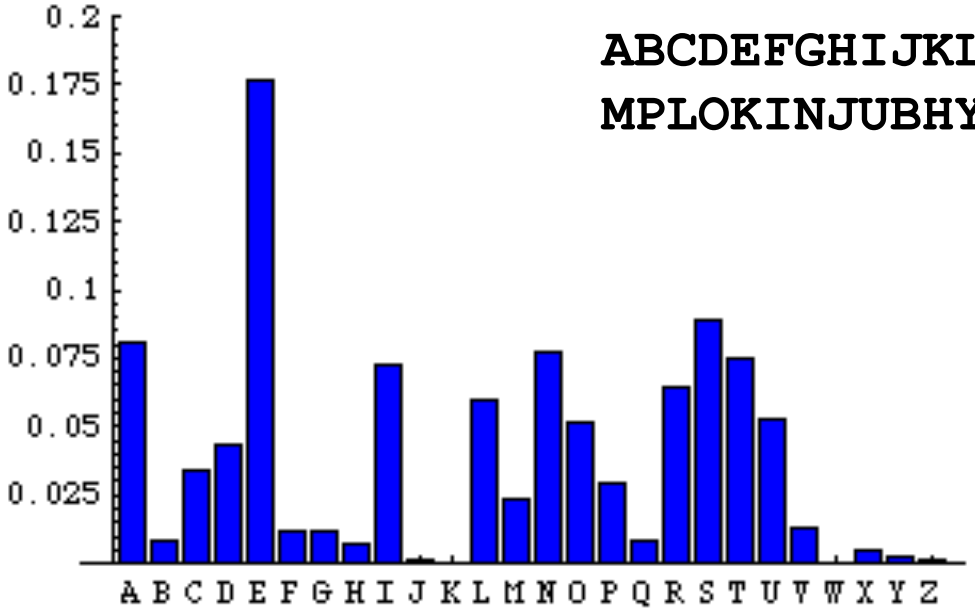
Cryptographie avancée

TD / TP de cryptanalyse appliquée

Walter SCHÖN

Rappels de cours

- La cryptanalyse du chiffre de Cesar est trivial car il n'y a que 26 décalages possibles
- La cryptanalyse d'un alphabet mélangé (ou toute substitution mono-alphabétique associant à toute lettre, toujours le même symbole) est un jeu à la portée de tout le monde, à réaliser grâce à l'analyse des fréquences.



ABCDEFGHIJKLMNOPQRSTUVWXYZ
MPLOKINJUBHYVGTCFRXDEWSZQA

Une excellente description de ce genre de cryptanalyse figure dans la nouvelle d'Edgar Poë « Le scarabée d'or »

Travaux pratiques

En écrivant de petits programmes dans le langage de votre choix (C, C++, Java, Python, Matlab…)

✓Prenez un grand texte Français et analysez le pour calculer les fréquences d'occurrence p_A,p_B,…p_Z prenez garde de ne compter que les lettres en majuscule ou en minuscules (dans le texte donné en exemple il n'y a que des majuscules), ne comptez pas les espaces (dans le texte donné en exemple il n'y a pas d'espace).

✓Utilisez ces valeurs de fréquences pour casser le code alphabet mélangé donné dans le fichier exemple (faites des hypothèses que vous validerez par recoupement, possibilité d'utiliser des digrammes fréquents…).

✓Gare (quand vous faites des substitutions) à différencier ce qui a été substitué de ce qui ne l'a pas été (idée possible : majuscules / minuscules).

Rappels de cours

- Le chiffre de Vigenère (1596), est obtenu en «additionnant» (A = 0 décalage, B = 1 décalage…) le texte à chiffrer avec un texte clé répété autant de fois que nécessaire :

```
  TO BE OR NOT TO BE
+ CL EC LE CLE CL EC
= VZ FG ZV PZX VZ FG
```

Revient à utiliser autant d'alphabets que de caractères de la clé, l'alphabet utilisé étant fonction de la position

La cryptanalyse du Vigenère est un est un exercice nettement plus compliqué mais réalisable si la longueur de la clé est faible devant la longueur du texte
Rappel : si la clé est réellement aléatoire, de longueur égale à celle du texte et utilisée une seule fois, le système est indéchiffrable.

Les pistes de cryptanalyse : séquences répétées

- Avant tout, il faut trouver la longueur de la clé, car une fois trouvée on sait que toutes les lettres en même position par rapport à la clé sont codés de la même manière, d'autant plus triviale à trouver que c'est un chiffre de Cesar (simple décalage)
- Une séquence de 2 lettres identiques a en effet de fortes chances d'être une séquence de 2 lettres identiques du texte clair, chiffrée avec la clé dans la même position. Le décalage correspondant est alors un multiple de la longueur de clé :

  VZ FG ZV PZX **VZ FG** Longueur de clé : diviseur de 9

  - S'il y a trop de digrammes répétés et que les tests sur les digrammes ne s'avèrent pas assez significatifs, refaire sur les trigrammes. Ne retenir que les indices les plus flagrants et prendre le pgcd des pas trouvés.

Les pistes de cryptanalyse : indice de coïncidence

- On appelle indice de coïncidence d'un texte, la probabilité de tomber sur deux lettres identiques lorsqu'on tire deux lettres au hasard.
- Dans tout ce qui suit on considère un texte ne comportant que les lettres A à Z en majuscules sans chiffres ni espaces ni caractères accentués ou spéciaux (séparer les mots est l'étape finale très facile d'une cryptanalyse).
- On note $n_A \ldots n_Z$ le nombre d'occurrences des lettres A…Z dans le texte composé de n lettres au total. Calculer l'indice de coïncidence.

Les pistes de cryptanalyse : indice de coïncidence

- On note $n_A \ldots n_Z$ le nombre d'occurrences des lettres A…Z dans le texte chiffré à analyser composé de n lettres. Calculer l'indice de coïncidence.
- Réponse : $$I = \frac{n_A(n_A-1)}{n(n-1)} + \ldots + \frac{n_Z(n_Z-1)}{n(n-1)}$$

$$C_{n_A}^2 = \frac{n_A(n_A-1)}{2}$$ manières de tirer 2 lettres A, parmi les n_A figurant dans le texte

$$C_n^2 = \frac{n(n-1)}{2}$$ manières de tirer 2 lettres, d'où la probabilité de tirer deux lettres A, idem jusqu'à Z et on additionne

Les pistes de cryptanalyse : indice de coïncidence

- Si maintenant le texte est long et vraiment aléatoire que vaut l'indice de coïncidence I_{Alea} ?
- Rappel : pour tout texte de longueur n, l'indice vaut :

$$I = \frac{n_A(n_A - 1)}{n(n-1)} + \ldots + \frac{n_Z(n_Z - 1)}{n(n-1)}$$

Les pistes de cryptanalyse : indice de coïncidence

- Si maintenant le texte est long et vraiment aléatoire que vaut l'indice de coïncidence I_{Alea} ? Rappel : pour tout texte de longueur n, l'indice vaut : $$I = \frac{n_A(n_A-1)}{n(n-1)} + \ldots + \frac{n_Z(n_Z-1)}{n(n-1)}$$

- Réponse : pour un texte long si l'on note $p_A = \frac{n_A}{n}$

la probabilité de tirer un A en tirant une lettre au hasard, la probabilité de tirer deux A en tirant deux lettres au hasard est proche de p_A^2

Mais pour un texte aléatoire $p_A=p_B=\ldots p_Z=1/26$

$$I_{Alea} = 26\left(\frac{1}{26}\right)^2 = \frac{1}{26}$$

Les pistes de cryptanalyse : indice de coïncidence

- Si maintenant le texte est long et écrit en Français que vaut I_{French} sachant que vous connaissez les fréquences en Français (les probabilités $p_A=p_B=\ldots p_Z$) ?

Les pistes de cryptanalyse : indice de coïncidence

- Si maintenant le texte est long et écrit en Français que vaut I_{French} sachant que vous connaissez les fréquences en Français (les probabilités $p_A, p_B, \ldots p_Z$) ?

- Réponse : $I_{French} = p_A^2 + \ldots + p_Z^2$

Travaux Pratiques : Utilisant les résultats de $p_A, p_B, \ldots p_Z$ donnés plus haut, calculez I_{French} et I_{Alea} que remarquez vous ?

Les pistes de cryptanalyse : indice de coïncidence

Lettre	Proba (%)	Proba^2
A	8,11	0,00657721
B	0,81	0,00006561
C	3,38	0,00114244
D	4,28	0,00183184
E	17,69	0,03129361
F	1,13	0,00012769
G	1,19	0,00014161
H	0,74	0,00005476
I	7,24	0,00524176
J	0,18	0,00000324
K	0,02	0,00000004
L	5,99	0,00358801
M	2,29	0,00052441
N	7,68	0,00589824
O	5,2	0,002704
P	2,92	0,00085264
Q	0,83	0,00006889
R	6,43	0,00413449
S	8,87	0,00786769
T	7,44	0,00553536
U	5,23	0,00273529
V	1,28	0,00016384
W	0,06	0,00000036
X	0,53	0,00002809
Y	0,26	0,00000676
Z	0,12	0,00000144
Somme	99,9	0,08058932

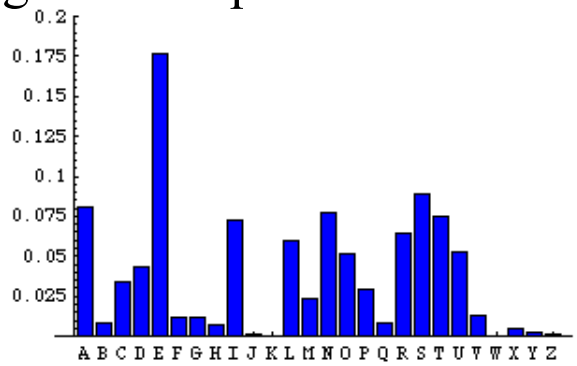
Voici mes résultats, donc pour moi

I_{French} = 0,0806 (sur le web ou trouve 0,0746 en Français et 0,065 pour l'Anglais)

Quels sont vos résultats ?

$I_{Aléa}$ = 0,03846

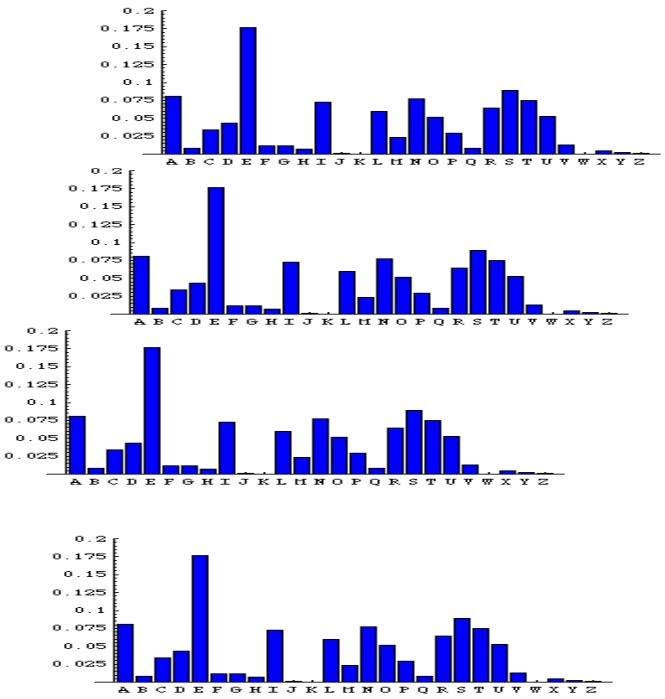
Comme on pouvait s'y attendre, l'indice de coïncidence est donc beaucoup plus grand lorsqu'on tire les lettres suivant l'histogramme très irrégulier de la langue, que lorsqu'on les tire sur un histogramme « plat »



Les pistes de cryptanalyse : indice de coïncidence

Et après ? Voici l'idée :

Si on effectue les statistiques sur les k sous-ensembles de lettres qui sont séparés d'une distance k (lettres en position 1, k+1, 2k+1… et on calcule un indice, lettres en position 2, k+2, 2k+2 et on calcule un autre indice etc. et on fait la moyenne des k indices calculés) , on trouve par définition l'indice de la langue si k est la longueur de clé



Si au contraire k n'est pas la longueur de clé, comme tente de l'illustrer cette figure on va mélanger des histogrammes décalés et trouver un indice se rapprochant de l'indice aléatoire qui est beaucoup plus faible.

D'où méthode 1 : calculer pour k=1, 2… les valeurs de cet indice et s'arrêter quand on trouve un indice anormalement grand (et voisin de celui de la langue).

Attention l'indice vaut aussi celui de la langue pour tout multiple de la longueur de clé.

D'un autre côté si la longueur de clé admet des diviseurs il faut aussi s'attendre à des valeurs d'indice important pour ces diviseurs…

Les pistes de cryptanalyse : indice de coïncidence

Méthode 2 : on calcule une valeur théorique de l'indice de coïncidence du texte complet, en fonction de la longueur de clé noté k et de la longueur n du texte.

Lorsque l'on prend deux lettres du texte complet (on suppose pour simplifier que n est multiple de k, on imagine que le texte est écrit en n/k lignes de k colonnes) :

✓Soit elles sont dans la même position par rapport à la clé : combien y a-t-il de paires qui sont dans ce cas ?

Les pistes de cryptanalyse : indice de coïncidence

Méthode 2 : on calcule une valeur théorique de l'indice de coïncidence du texte complet, en fonction de la longueur de clé noté k et de la longueur n du texte.
Lorsque l'on prend deux lettres du texte complet (on suppose pour simplifier que n est multiple de k, on imagine que le texte est écrit en n/k lignes de k colonnes)

✓Soit elles sont dans la même position par rapport à la clé : combien y a-t-il de paires qui sont dans ce cas ?

$$kC_{n/k}^2 = \frac{k}{2}\left(\frac{n}{k}\right)\left(\frac{n}{k}-1\right) = \frac{n(n-k)}{2k}$$

k : choix de la colonne

$C_{n/k}^2$: choix des deux lettres dans la colonne

Les pistes de cryptanalyse : indice de coïncidence

Méthode 2 : on calcule une valeur théorique de l'indice de coïncidence du texte complet, en fonction de la longueur de clé noté k et de la longueur n du texte.
Lorsque l'on prend deux lettres du texte complet (on suppose pour simplifier que n est multiple de k, on imagine que le texte est écrit en n/k lignes de k colonnes)

✓Soit elles ne sont dans la même position par rapport à la clé : combien y a-t-il de paires qui sont dans ce cas ?

Les pistes de cryptanalyse : indice de coïncidence

Méthode 2 : on calcule une valeur théorique de l'indice de coïncidence du texte complet, en fonction de la longueur de clé noté k et de la longueur n du texte.

Lorsque l'on prend deux lettres du texte complet (on suppose pour simplifier que n est multiple de k, on imagine que le texte est écrit en n/k lignes de k colonnes)

✓Soit elles ne sont dans la même position par rapport à la clé : combien y a-t-il de paires qui sont dans ce cas ?

$$C_k^2\left(\frac{n}{k}\right)^2 = \frac{k(k-1)}{2}\frac{n^2}{k^2} = \frac{n^2(k-1)}{2k}$$

C_k^2 : choix des deux colonnes différentes

$\left(\frac{n}{k}\right)^2$: choix d'une lettre quelconque de chacune de ces deux colonnes

On vérifie bien que : $\frac{n(n-k)}{2k} + \frac{n^2(k-1)}{2k} = \frac{n(n-1)}{2}$

Les pistes de cryptanalyse : indice de coïncidence

Méthode 2 : on calcule une valeur théorique de l'indice de coïncidence du texte complet, en fonction de la longueur de clé noté k et de la longueur n du texte.

Compte tenu de ce qui précède estimer l'indice I_{Cipher} du texte chiffré en se souvenant que l'indice est I_{French} dans l'ensemble des paires issues d'une même colonne, et en considérant qu'il sera I_{Alea} pour les paires issues de colonnes différentes (ce qui est évidemment une approximation, on suppose que l'effet de tirages dans plusieurs histogrammes décalés correspondant au Français est équivalent à un tirage uniforme).

Les pistes de cryptanalyse : indice de coïncidence

Méthode 2 : on calcule une valeur théorique de l'indice de coïncidence du texte complet, en fonction de la longueur de clé noté k et de la longueur n du texte.

Compte tenu de ce qui précède estimer l'indice I_{Cipher} du texte chiffré en se souvenant que l'indice est I_{French} dans l'ensemble des paires issues d'une même colonne, et en considérant qu'il sera I_{Alea} pour les paires issues de colonnes différentes (ce qui est évidemment une approximation, on suppose que l'effet de tirages dans plusieurs histogrammes décalés correspondant au Français est équivalent à un tirage uniforme).

$$\frac{n(n-1)}{2} I_{Cipher} = \frac{n(n-k)}{2k} I_{French} + \frac{n^2(k-1)}{2k} I_{Alea}$$

Les deux expressions de part et d'autres du signe égale représentent le nombre estimé de paires de lettres identiques

Dans l'expression précédente, seul k n'est pas connu. Calculer k en fonction des indices et de n…

Les pistes de cryptanalyse : indice de coïncidence

Méthode 2 :

$$\frac{n(n-1)}{2} I_{Cipher} = \frac{n(n-k)}{2k} I_{French} + \frac{n^2(k-1)}{2k} I_{Alea}$$

Donne après un calcul facile :

$$k = \frac{n(I_{French} - I_{Alea})}{I_{French} - I_{Cipher} + n(I_{Cipher} - I_{Alea})}$$

Travaux pratiques :

✓A partir du long texte Vigenere fourni (deux variantes du même texte chiffré avec deux clés de longueur différente)

✓Trouver la longueur de clé par l'une des méthodes précédentes

✓Pour chaque sous-ensemble trouver la lettre la plus fréquente et deviner le décalage de César correspondant

✓Déchiffrer le texte

Quelques sites

- http://www.bibmath.net/crypto
- http://www.dil.univ-mrs.fr/~vancan/inf7/old/documents/EARLYCIPHERS/Vigenere.html
- http://www.apprendre-en-ligne.net/crypto/crypto