Cybersécurité
Introduction Crypto
TP1

Vasco VALADARES SEMANA et Guillaume BLAS

Donner votre avis en répondant aux questions suivantes :

Qu'est-ce qu'une donnée personnelle ? (donner des exemples)

Une donnée personnelle est toute information se rapportant à une personne physique identifiée ou identifiable. Par exemple, le nom, l'adresse e-mail, le numéro de téléphone, l'adresse IP, les données de localisation, les informations de santé, etc.

Que sont les données sensibles ? (donner des exemples)

Les données sensibles sont des données personnelles permettant d'identifier une personne physique de manière unique ou de révéler des informations intimes sur elle. Par exemple, les données de santé, les opinions politiques, les convictions religieuses, l'origine raciale ou ethnique, les données biométriques, etc.

Qu'est-ce que les métadonnées ? (donner des exemples)

Les métadonnées sont des données qui décrivent d'autres données. Elles fournissent des informations sur la structure, le contenu, la provenance ou l'utilisation d'une donnée. Par exemple, les métadonnées d'une photo peuvent inclure la date et l'heure où celle-ci a été prise, les coordonnées GPS, le modèle de l'appareil photo, etc.

Qu'est-ce que le cyberespace ?

Le cyberespace est un espace virtuel créé par les réseaux informatiques et les technologies de l'information. Il englobe l'ensemble des interactions, des communications et des activités qui se déroulent en ligne, que ce soit sur Internet, les réseaux sociaux, les applications mobiles, etc.

Pourquoi les données personnelles passionnent tant ?

Les données personnelles passionnent tant car elles sont au coeur de nombreuses activités en ligne. Ce sont des ressources préciseuses pour comprendre les comportements, les préférences et les besoins des utilisateurs, ce qui en fait un enjeu majeur pour la vie privée et la sécurité. Elles permettent aux entreprises d'appliquer des stratégies de marketing ciblé, aux gouvernements de surveiller les citoyens, et aux individus de partager des informations sur eux-mêmes.

Internet, est-il un cimetière de données personnelles ?

Il n'y a pas de bonne ou mauvaise réponse, cependant, nous pouvons considérer qu'Internet est un cimetière de données personnelles dans la mesure où de nombreuses informations personnelles y sont stockées, souvent sans le consentement explicite des utilisateurs, et ce, même après leur mort.

Si je n'utilise pas ou très peu Internet, suis-je concerné par la sécurité des données personnelles ? Pourquoi ?

Oui, même si on n'utilise pas ou très peu Internet, on est concerné par la sécurité des données personnelles, car de nombreuses données personnelles peuvent être collectées à notre sujet par d'autres moyens, tels que les transactions bancaires, les achats en magasin, les interactions avec les services publics, etc.

Le « sentiment de sécurité dans le cloud ».

Je pense que le sentiment de sécurité dans le cloud peut être trompeur, car nous ne savons pas toujours où nos données sont stockées, qui y a accès et comment elles sont protégées, on peut croire ou être persuadés que nos données sont en sécurité, alors qu'en réalité, elles peuvent être vulnérables à des attaques ou à des fuites.

« Pas grave si on prend mes données, je n'ai rien à cacher ».

Ce n'est pas une question d'avoir des choses à cacher, mais plutôt une question de respect de la vie privée et de contrôle sur ses propres informations. Même si on n'a rien à cacher, on peut ne pas vouloir que nos données soient utilisées à des fins commerciales, surveillées par des gouvernements ou exposées à des risques de sécurité. Ce n'est pas parce qu'on n'a rien à cacher que l'on a pas quelque chose à perdre.

Quels sont vos interrogations sur le monde digital ?

Comment puis-je protéger ma vie privée en ligne si tout ce que je fais est suivi sans que je n'en sois au courant ?

1 - Code César

1.1 - Chiffrement et déchiffrement César

Nous avons implémenté deux fonctions : `caesar_encode` pour chiffrer un message en décalant chaque lettre de l'alphabet d'un nombre donné, et `caesar_decode` pour l'inverser. Les fonctions gèrent les majuscules, les minuscules, et préservent les ponctuations, espaces, etc..

```python
def caesar_encode(text, shift):
    result = ""
    for char in text:
        if char.isalpha():
            base = ord('A') if char.isupper() else ord('a')
            result += chr((ord(char) - base + shift) % 26 + base)
        else:
            result += char
    return result

def caesar_decode(text, shift):
    return caesar_encode(text, -shift)

encoded = caesar_encode("cybersecurite", 11)
print("Chiffrement de 'cybersecurite' avec clé 11 :", encoded)
print(f"Déchiffrement de '{encoded}' avec la clé 11 :", caesar_decode(encoded, 11))
```

✓ 0.0s

```
Chiffrement de 'cybersecurite' avec clé 11 : njmpcdpnfctep
Déchiffrement de 'njmpcdpnfctep' avec la clé 11 : cybersecurite
```

1.2 - Attaque par force brute

Nous avons implémenté une fonction de force brute (`brute_force_caesar`) qui teste les 26 clés possibles (0 à 25) pour déchiffrer un message César.

```python
def brute_force_caesar(cipher_text):
    for key in range(26):
        print(f"Clé {key}: {caesar_decode(cipher_text, key)}")

message1 = "WP NZOLRP PDE FY LCE"
message2 = "HA YKZWCA AOP QJ WNP"

print("============ Message 1 ============")
brute_force_caesar(message1)

print("\n============ Message 2 ============")
brute_force_caesar(message2)
```

✓ 0.0s

Extraits du brute force pour les messages 1 et 2 :

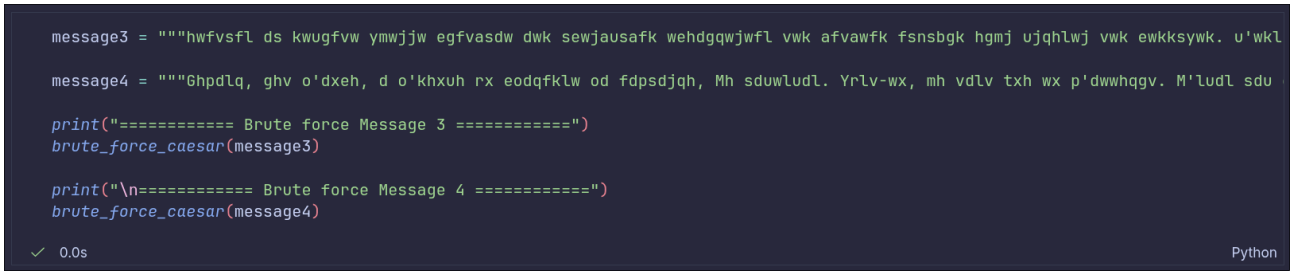
```
============ Message 1 ============
Clé  0: WP NZOLRP PDE FY LCE
Clé  1: VO MYNKQO OCD EX KBD
Clé  2: UN LXMJPN NBC DW JAC
  ...
Clé 10: MF DPEBHF FTU VO BSU
Clé 11: LE CODAGE EST UN ART
Clé 12: KD BNCZFD DRS TM ZQS
Clé 13: JC AMBYEC CQR SL YPR
Clé 14: IB ZLAXDB BPQ RK XOQ
Clé 15: HA YKZWCA AOP QJ WNP

============ Message 2 ============
Clé  0: HA YKZWCA AOP QJ WNP
Clé  1: GZ XJYVBZ ZNO PI VMO
Clé  2: FY WIXUAY YMN OH ULN
  ...
Clé 21: MF DPEBHF FTU VO BSU
Clé 22: LE CODAGE EST UN ART
Clé 23: KD BNCZFD DRS TM ZQS
Clé 24: JC AMBYEC CQR SL YPR
Clé 25: IB ZLAXDB BPQ RK XOQ
```

- Message 1 : « WP NZOLRP PDE FY LCE » → déchiffré avec la clé **11** : « LE CODAGE EST UN ART »
- Message 2 : « HA YKZWCA AOP QJ WNP » → déchiffré avec la clé **22** : « LE CODAGE EST UN ART »

On remarque que les deux messages donnent exactement le même texte en clair mais avec des clés différentes (11 et 22), ce qui montre qu'avec le chiffrement César, il existe plusieurs clés possibles pour obtenir le même résultat.

1.3 - Déchiffrement avec majuscules et minuscules

Comme on a déjà implémenté ces contraintes dans nos fonctions on passe à la brute-force, nous avons identifié les messages :



```
message3 = """hwfvsfl ds kwugfvw ymwjjw egfvasdw dwk sewjausafk wehdgqwjwfl vwk afvawfk fsnsbgk hgmj ujqhlwj vwk ewkksywk. u'wkl

message4 = """Ghpdlq, ghv o'dxeh, d o'khxuh rx eodqfklw od fdpsdjqh, Mh sduwludl. Yrlv-wx, mh vdlv txh wx p'dwwhqgv. M'ludl sdu

print("============ Brute force Message 3 ============")
brute_force_caesar(message3)

print("\n============ Brute force Message 4 ============")
brute_force_caesar(message4)
```

✓ 0.0s Python

```
============ Brute force Message 3 ============
  ...
Clé 18: pendant la seconde guerre mondiale les americains employerent des indiens navajos
pour crypter des messages. c'est l'un des rares codes de l'histoire a n'avoir jamais ete
brise. l'impenetrabilite du code navajo vient en particulier du fait que cette langue n'a
aucun lien avec une quelconque langue europeenne ou asiatique.
  ...

============ Brute force Message 4 ============
  ...
Clé  3: Demain, des l'aube, a l'heure ou blanchit la campagne, Je partirai. Vois-tu, je
sais que tu m'attends. J'irai par la foret, j'irai par la montagne. Je ne puis demeurer
loin de toi plus longtemps. Je marcherai les yeux fixes sur mes pensees, Sans rien voir au
dehors, sans entendre aucun bruit, Seul, inconnu, le dos courbe, les mains croisees,
Triste, et le jour pour moi sera comme la nuit. Je ne regarderai ni l'or du soir qui
tombe, Ni les voiles au loin descendant vers Harfleur, Et quand j'arriverai, je mettrai
sur ta tombe Un bouquet de houx vert et de bruyere en fleur. Victor Hugo - Les
Contemplations
  ...
```

La vérification par re-chiffrement confirme que nos fonctions sont correctes :

```
key3 = 18
key4 = 3

decoded3 = caesar_decode(message3, key3)
decoded4 = caesar_decode(message4, key4)

print(f"\n============ Message 3 déchiffré ============")
print(decoded3)

print(f"\n============ Message 4 déchiffré ============")
print(decoded4)

print("\n============ Vérification par re-chiffrement ============")
encoded3 = caesar_encode(decoded3, key3)
encoded4 = caesar_encode(decoded4, key4)

print("Message 3 re-chiffré == original =", encoded3 == message3, ", message :\n", encoded3)
print("Message 4 re-chiffré == original =", encoded4 == message4, ", message :\n", encoded4)
```

2] ✓ 0.0s Python

```
============ Message 3 déchiffré ============
pendant la seconde guerre mondiale les americains employerent des indiens navajos pour crypter des messages. c'est l'un des rares c

============ Message 4 déchiffré ============
Demain, des l'aube, a l'heure ou blanchit la campagne, Je partirai. Vois-tu, je sais que tu m'attends. J'irai par la foret, j'irai

============ Vérification par re-chiffrement ============
Message 3 re-chiffré == original = True , message :
 hwfvsfl ds kwugfvw ymwjjw egfvasdw dwk sewjausafk wehdgqwjwfl vwk afvawfk fsnsbgk hgmj ujqhlwj vwk ewkksywk. u'wkl d'mf vwk jsjwk
Message 4 re-chiffré == original = True , message :
 Ghpdlq, ghv o'dxeh, d o'khxuh rx eodqfklw od fdpsdjqh, Mh sduwludl. Yrlv-wx, mh vdlv txh wx p'dwwhqgv. M'ludl sdu od iruhw, m'ludl
```

2 - Indice de Coïncidence

2.1 - Calcul de l'indice de coïncidence et détection de langue

Notre fonction `coincidence_index` calcule l'indice pour chaque message, puis `detect_language` compare la valeur obtenue avec les indices présents dans le tableau :

```python
def coincidence_index(text):
    freq = [0] * 26
    total_letters = 0
    for char in text:
        if char.isalpha():
            freq[ord(char.lower()) - ord('a')] += 1
            total_letters += 1
    if total_letters <= 1:
        return 0.0
    return sum(f * (f - 1) for f in freq) / (total_letters * (total_letters - 1))

INDICES = {
    "Suédois": 0.0644, "Serbe": 0.0643, "Russe": 0.0529, "Portugais": 0.0745,
    "Néerlandais": 0.0798, "Norvégien": 0.0694, "Malaysien": 0.0852,
    "Japonais": 0.0772, "Italien": 0.0738, "Hébreu": 0.0768, "Grec": 0.0691,
    "Français": 0.0778, "Finnois": 0.0737, "Esperanto": 0.069, "Espagnol": 0.077,
    "Danois": 0.0707, "Arabe": 0.0758, "Anglais": 0.0667, "Allemand": 0.0762,
}

def detect_language(text):
    ic = coincidence_index(text)
    closest = min(INDICES, key=lambda lang: abs(INDICES[lang] - ic))
    return closest, ic
```

✓ 0.0s

```python
for i, msg in [(3, message3), (4, message4), (5, message5), (6, message6)]:
    lang, ic = detect_language(msg)
    print(f"\nMessage {i}:")
    print(f"  Indice = {ic:.4f}")
    print(f"  Langue détectée : {lang} (Indice Théorique : {INDICES[lang]:.4f})")
```

✓ 0.0s

```
Message 3:
  Indice = 0.0776
  Langue détectée : Français (Indice Théorique : 0.0778)

Message 4:
  Indice = 0.0710
  Langue détectée : Danois (Indice Théorique : 0.0707)

Message 5:
  Indice = 0.0832
  Langue détectée : Malaysien (Indice Théorique : 0.0852)

Message 6:
  Indice = 0.0737
  Langue détectée : Finnois (Indice Théorique : 0.0737)
```

Le seul correct est le message 3 qui est bien en français.

3 - L'Analyse Fréquentielle

3.1 - Déchiffrement par analyse fréquentielle

Notre fonction `decrypt_by_frequency_analysis` combine l'indice de coïncidence (pour déterminer la langue) et l'analyse des fréquences (pour trouver la clé).

Résultats :

- **Message 4** : la lettre la plus fréquente dans le texte chiffré est « H » (15.1%). L'écart entre H et E donne une clé de 3. Le texte déchiffré confirme le poème de Victor Hugo.
- **Message 5** : la lettre la plus fréquente est « T » (20.1%). L'écart donne une clé de 15. Le texte déchiffré parle de Julie et Jules qui inventent un moyen de communication secret basé sur un rectangle de transposition.

[AJOUTER SCREENSHOT ICI]

3.2 - Validation avec le message 6

Le message 6 est validé avec succès par notre fonction d'analyse fréquentielle :

- La lettre la plus fréquente dans le texte chiffré est « L » (13.1%).
- La clé trouvée est 7.
- Le texte déchiffré est le poème « The Tiger » (The Tyger) de William Blake, un poème anglais célèbre.

Cela confirme que notre méthode fonctionne aussi bien pour les textes en français qu'en anglais.

[AJOUTER SCREENSHOT ICI]

3.3 - Analyse par digrammes et trigrammes

L'analyse des digrammes (paires de lettres) et trigrammes (triplets de lettres) permet de déterminer si un texte est en clair ou chiffré, et d'identifier sa langue. On compare les n-grammes les plus fréquents du texte avec des listes de référence (par exemple : « ES », « DE », « LE » en français ; « TH », « HE », « IN » en anglais).

Nos fonctions `count_digrams`, `count_trigrams` et `is_clear_text` analysent un texte et calculent un score de correspondance avec les n-grammes de référence.

Résultats :

- **Message 4 chiffré** : les digrammes les plus fréquents (HV, DL, HQ…) ne correspondent pas aux digrammes français ou anglais courants → verdict : le message est **chiffré**.
- **Message 4 déchiffré** : les digrammes trouvés (ES, AI, EN, DE, ER, LE…) correspondent aux digrammes français les plus courants → verdict : le message est **en clair** (Français).
- **Message 6 chiffré** : les digrammes (AO, OL, OH…) ne correspondent pas → verdict : le message est **chiffré**.
- **Message 6 déchiffré** : les digrammes (TH, HE, AT, RE, AN…) correspondent aux digrammes anglais → verdict : le message est **en clair** (Anglais).

[AJOUTER SCREENSHOT ICI]

3.4 - (BONUS) Lettres répétées

L'analyse des lettres répétées (doublons consécutifs comme LL, SS, TT, etc.) fournit des indices supplémentaires sur la langue d'un texte.

Pour les messages déchiffrés :

- **Message 3 (FR)** : doublons trouvés : EE(2), RR(1), SS(1), TT(1), AA(1), NN(1)
- **Message 4 (FR)** : doublons trouvés : EE(4), SS(3), TT(2), NN(1), MM(1), RR(1)
- **Message 6 (EN)** : doublons trouvés : TT(6), MM(5), EE(5), LL(1), DD(1), RR(1)

On observe que le texte anglais présente plus de doublons TT et MM, tandis que les textes français ont davantage de EE et SS. Ces motifs sont caractéristiques des langues et peuvent compléter les autres techniques d'analyse pour améliorer la détection de la langue.

[AJOUTER SCREENSHOT ICI]