TP - Codage du texte

M. Tellene

1 Création de l'environnement de travail

Avant de commencer le TP, vous allez créer votre environnement de travail. Pour ce faire vous allez dans votre dossier personnel et vous allez créer un dossier « codage du texte ».

Une fois le dossier créé, vous irez sur **Pronote** et récupérerez le dossier « fichier.zip » qui se trouve dans le cahier de texte du jour. Une fois le dossier télécharger, vous le copierez dans votre dossier « codage du texte » et vous l'extrairez. Si ça c'est bien passé, vous devriez avoir un dossier « fichiers » contenant 5 fichiers « bonjour » et un dossier « fichier_mystere » qui contient 5 fichiers.

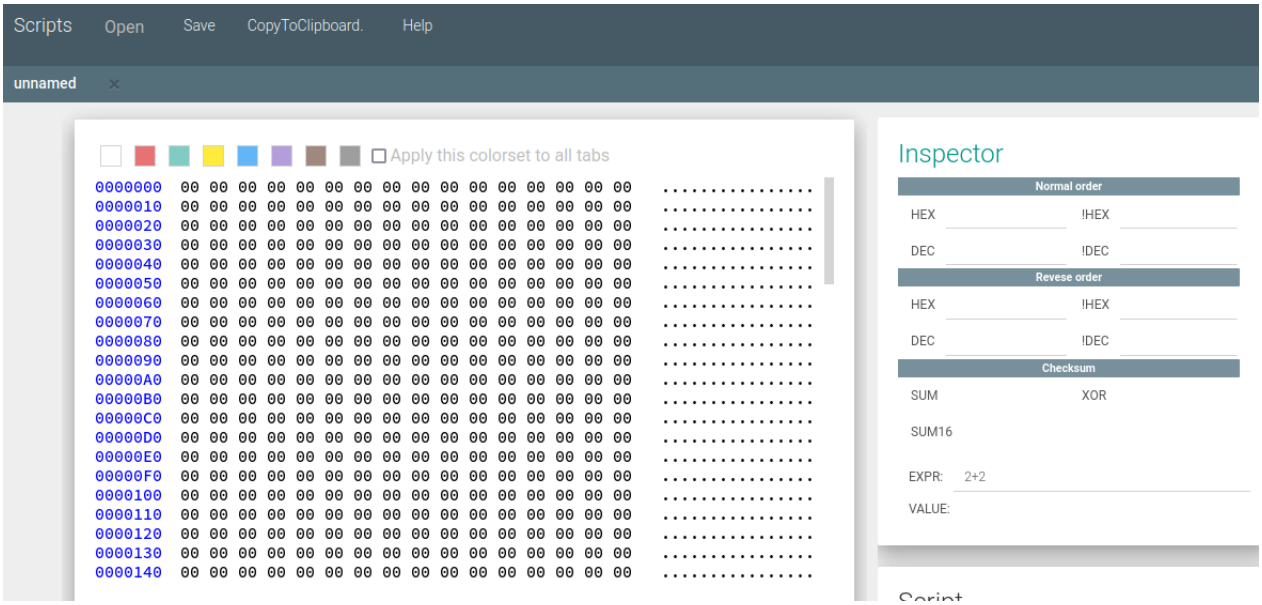
2 Type et taille de fichier

Dans le dossier « fichiers », vous trouverez plusieurs fichiers « bonjour. ... », avec des extensions différentes (.txt, .gif, .odt, .pdf).

1. **Double-cliquer** sur chacun d'entre eux pour voir comment ils s'ouvrent et compléter la colonne « Nom du logiciel qui permet de l'ouvrir » du tableau ci-dessous :

Nom du fichier	Taille (en octets)	Nom du logiciel qu permet de l'ouvrir
bonjour.gif		
bonjour.html		
bonjour.odt		
bonjour.pdf		
bonjour.txt		

2. Pour regarder leur taille, faire un **clic droit** et aller dans **« propriété »**. Vous trouverez les tailles des différents fichiers que vous pouvez reporter dans le tableau précédent. La taille d'un fichier se mesure en octet (kilo octet, mégaoctet, giga octet...).
3. Aller sur le site https://hex-works.com/eng. Vous obtenez une fenêtre comme celle-ci :



Ce que vous avez devant les yeux est un **éditeur hexadécimal**. Il permet de voir en détail ce que contient un fichier, sans se soucier de ce que l'on veut faire avec.

4. Cliquer sur **Open** et **ouvrir tous les fichiers « bonjour »**. Les « 00 » se sont normalement transformés pour la plupart en nombres hexadécimaux.
5. Prendre le fichier « bonjour.txt » et le convertir en binaire, combien contient-il de bits ?.....

   .....

   Pour rappel, un octet = 8 bits.
6. Toujours avec « bonjour.txt », quel est le rapport entre la taille du fichier et ce que vous avez à l'écran ? .....

   .....
7. Si vous êtes attentifs, on peut retrouver l'extension de chaque fichier dans leur code hexadécimal. Dans le dossier « fichiers_mystere », vous trouverez des fichiers dont l'extension a été enlevée. A l'aide de l'éditeur hexadécimal, retrouver chacune des extensions de chacun des fichiers. Compléter alors le tableau ci-dessous :

Nom du fichier	Extension
fichier_mystere_1	
fichier_mystere_2	
fichier_mystere_3	
fichier_mystere_4	
fichier_mystere_5	

8. Une fois les extensions retrouvées, renommer les fichiers en ajoutant l'extension **(clic droit → renommer)** puis double-cliquer dessus pour les ouvrir.

3 Encodage de fichier

Si vous avez réussi la partie précédente, vous avez remarqué que « fichier_mystere_4 » possède des caractères illisibles. Il y a quelque chose de bizarre avec ce fichier. C'est ce que l'on va expliquer dans cette deuxième partie.

3.1 Quel est le problème avec le fichier mystère 4 ?

1. Dans l'éditeur hexadécimal, quel est le premier octet de « bonjour.txt » ? .....
2. Aller sur le site https://shop.alterlinks.com/ascii-table/ascii-table-fr.php. Dans la colonne « Valeur », chercher B. Que peut-on remarquer ? .....

   .....
3. Chercher les autres caractères du mot « bonjour » et vérifier que cela correspond bien.
4. Ce que vous venez d'utiliser s'appelle une **table ASCII**. La norme ASCII est ce que l'on appelle une **norme d'encodage des caractères**. Elle donne la correspondance entre le codage binaire des caractères et les caractères d'un fichier texte.
5.
6. Écrire votre prénom en ASCII : .....

   .....
7. A l'aide de la table ASCII décoder le message binaire suivant :

Caractère en binaire	Caractère
01010000	
01010010	
01010010	
01010010	
01010010	
01100001	
01101101	
01001101	
01001110	
01000001	
01110100	
01101001	
01101111	

   .....
8. Revenons au fichier_mystere_4. Rechercher le code ASCII du caractère « é ». Quel est le problème ? .....

   .....

   .....

   .....

   .....

9. Si vous avez terminé les questions précédentes, alors vous avez pu vous apercevoir que la table ASCII ne traite pas tous les caractères existants

3.2 Vers une explication et une solution

1. Faire une recherche Internet sur l'ASCII et expliquer en quelques mots pourquoi cet encodage ne traite pas tous les caractères : .....

   .....
2. Aller sur le site https://www.ascii-code.com/, vous indiquerez quel est le code hexadécimal du « é » et comment se nomme la table auquel il appartient : .....
3. Faire une recherche Internet sur l'encodage ISO/CEI 8859-1 et sur le mot clef `charset`.

   Faire un petit résumé ici : .....

   .....

   .....

   .....
4. Ouvrez « fichier_mystère_4 »avec le bloc note **(clic droit → ouvrir avec bloc note)**. Il y a un endroit à modifier pour que le « é » s'affiche correctement. A vous de le trouver. Une fois la modification faite, sauvegarder et recharger le fichier dans le navigateur.
5. Faire une recherche sur l'Unicode et l'UTF8 : .....

   .....

RÉSUMÉ :

- L'extension d'un nom de fichier permet au système d'exploitation de l'ordinateur de savoir quel logiciel utiliser pour lire le fichier. (par exemple .py, .mp3, .html ...)
- Un fichier est une suite de nombres binaires qui nécessitent d'être interprétés pour être lus correctement
- Un éditeur hexadécimal permet de lire directement le code binaire contenu dans un fichier
- Pour écrire du texte, on a besoin d'un encodage. L'encodage détermine les caractères que l'on pourra afficher
- Il y a beaucoup de normes différentes, mais la plus répandue actuellement est l'Unicode qui contient quasiment tous les caractères de toutes les langues
- Dans un fichier HTML, on utilise l'attribut `charset` afin d'indiquer quel jeu de caractères sera utiliser dans la page web afin de bien les afficher