	Travaux pratiques de téléinformatique **TP05 – Codage de source**	**Classe** **T-1a/f**
EIA-FR	J. Robadey et R. Scheurer	1.12.2017

1. OBJECTIFS

- exercer les algorithmes de codage de source selon “Shannon-Fano” et “Huffman”
- calculer l’*entropie*, la *quantité de décision* et la *redondance* de sources (par l’utilisation d’une feuille de calcul *MSExcel*)
- tout cela sur des exemples concrets (3 textes dans les 3 langues, dont un en français)

2. TÂCHES

Lire **attentivement** les instructions ci-dessous. Si des points semblent flous, discuter avec vos collègues. Si le problème persiste, appeler le professeur pour de l’assistance.

2.1. Préparation (max 15 min) – Rassembler des exemples de textes

Avant d’appliquer un algorithme de codage, vous devez chercher sur le web 3 textes **en prose** (pas de code de programmation, pas de formules, pas de choses techniques, etc.) écrites en **français** et deux autres langues. Chaque texte doit contenir **au minimum 100'000 caractères** de prose.

1. Obtenir les 3 textes, un pour chaque langue mentionnée (exemple : http://www.gutenberg.org ou liens à la fin de cette donnée).
2. **Relever la référence de la source** (URL ou autre type de référence), vous en aurez besoin pour l’indiquer dans le fichier Excel.
3. Copier ces textes dans un fichier en format texte et nommer le avec l’extension “`<langue>.txt`” (par ex. “`anglais.txt`”).
4. Contrôler à l’aide de la classe Java `CharCount` mise à disposition si ces textes contiennent réellement au minimum 100'000 caractères (p.ex. avec la commande “`java CharCount anglais.txt`”). *Pour faire ce test, il faut utiliser une fenêtre DOS, aller dans le répertoire des fichiers texte et écrire les instructions java***.**

2.2. Calcul des Probabilités

Les deux algorithmes Shannon-Fano et Huffman, sont basés sur la probabilité des caractères à être connu à l’avance, avant le codage.
Soit vous déterminez cette probabilité des caractères à la main (☺), ou vous pouvez utiliser la classe java `CharProbs` déjà préparée pour vous. Ce programme ne va pas seulement vous indiquer la probabilité des caractères, il va aussi vous créer un fichier appelé “`<langue>.xls`” contenant ces informations.

1. Démarrer le programme java (utiliser une fenêtre DOS) avec le nom de votre fichier texte (en incluant l’extension “`.txt`”) comme seul paramètre (par exemple “`java CharProbs anglais.txt`”). Le programme va générer deux fichiers : `<langue>.filtered.txt` (texte filtré) et `<langue>.xls` (statistiques/probabilités).
2. Ouvrir le fichier généré “`<langue>.xls`” dans MSExcel (double-click sur le fichier) et contrôler que l’information est présentée en colonnes.
3. Effectuer les étapes 1-3 **pour les trois fichiers texte** (les trois langues).

2.3. Calcul de l’Entropie / Quantité de Décision / Redondance

Nous allons maintenant calculer les paramètres importants des sources. Pour cela vous pouvez utiliser le fichier Excel déjà préparé pour vous pour y entrer vos données.

1. Ouvrir le fichier “`TP05.xls`” et regarder comment il est organisé. Entrer vos noms dans les champs prévus à cet effet. Sauvegarder le fichier sous **le nouveau nom** de "`TP05_nom1_nom2.xls`".

2. **Pour chaque langue**: copier et coller vos données dans les feuilles de calcul correspondantes (c.-à-d. créer une seconde et troisième feuille de calcul dans le **même fichier**). Trier les données suivant l'ordre décroissant des probabilités.
3. Entrer les formules correctes dans les cellules indiquées. Utiliser la fonction d'aide pour déterminer quelle fonction utiliser et comment. Les colonnes A à G doivent être remplies pour les trois langues. Les colonnes doivent être réordonnées en fonction du nombre de caractère (maximum vers minimum)

2.4. Pratiquer le Codage de Source

Sélectionner la langue pour laquelle vous allez construire le codage de source en utilisant les algorithmes de Shannon-Fano et de Huffman. Effectuer le travail sur papier, et, une fois terminé, présenter vos résultats au professeur.

1. **Sélectionner une langue** avec laquelle vous voulez travailler (c.-à-d. faire le codage)
2. **Copier le fichier texte correspondant** dans le répertoire indiqué par le professeur.
3. Construire le codage de source selon ***Shannon-Fano*** pour cette langue sur une feuille blanche A3 (recommandation: il est beaucoup plus facile de baser vos calculs sur le nombre d'apparition que sur les probabilités). **Règle de codage**: assigner le code bit "0" pour les probabilités élevées, le code bit "1" pour les basses. **Montrer vos résultats au professeur** avant de continuer.
4. Entrer les mots de codes de votre codage dans la feuille de calcul (**formater les cellules en tant que texte** pour être en mesure d'entrer des valeurs avec des zéros au début tel que "0010"). Calculer la longueur moyenne des mots du code (en utilisant une fonction Excel). Calculer la redondance après codage.
5. **Montrer une impression de vos résultats au professeur** (adapter l'impression afin de placer tout sur une seule page A4).
6. Après avoir terminé cette analyse, effectuer exactement le **même travail** pour la **même langue** avec l'algorithme de ***Huffman*** (c.-à-d. étape 3 à 5).

2.5. Qu'en est-il du taux de compression ?

- Calculer la quantité de place économisée si l'on applique l'algorithme de Huffman sur le fichier texte que nous avons analysé.
- Comprimer votre fichier texte (utiliser la version filtrée `<langue>.filtered.txt`) en utilisant un outil de compression. Utiliser comme valeur de référence non compressée la quantité de décision **D x le nombre de caractères**. Indiquer la taille du fichier comprimé et calculer le taux de compression. Comparer avec le résultat du codage Huffman et avec la valeur de l'entropie. Expliquer le résultat obtenu.

3. RAPPORT

A la fin de ce travail pratique, vous allez **retourner tous les fichiers et papiers** avec lesquelles vous avez travaillés. Rassembler tous les fichiers dans un **fichier ZIP** (appelé "`TP05_nom1_nom2.zip`") et mettez le sur « don_prof ». Le fichier zip doit contenir au minimum:

- **le fichier** "`TP05_nom1_nom2.xls`" contenant toutes les feuilles de calcul, une par langue et la feuille « Question 2.5 » remplie
- **tous les textes sources (**les trois fichiers `<langue>.txt` et les trois fichiers `<langue>.filtered.txt`). Contrôler que vous avez bien mentionné la référence de ces fichiers dans les champs prévus à cet effet dans le fichier `TP05.xls`.
- **les trois fichiers** `<langue>.xls` (générés par le programme Java `CharProbs`)

Imprimer toutes les feuilles de calcul du fichier "`TP05_nom1_nom2.xls`" et les transmettre avec **tous les papiers / documents** sur lesquelles vous avez construit les codes sources au professeur (n'oublier pas d'y mettre vos noms !).

4. LIENS POUR TROUVER DES TEXTES (SANS GARANTIE !)

http://www.gutenberg.org
http://www.booklinks.de/menu.php

http://literature.org
recherche « free book in various languages » ou « free ASCII text in finnish » ou …

5. Applet (sans garantie !)

http://www.cs.pitt.edu/~kirk/cs1501/animations/Huffman.html
http://projects.hudecof.net/diplomovka/online/ucebnica/applets/AppletShannonFano.html
http://www.tillwiebke.de/index.html#HuffmanShannonFano
http://www.cs.sfu.ca/CourseCentral/365/li/squeeze/LZW.html (Lempel-Ziv)
http://bigwww.epfl.ch/demo/basisdct/index.html (jpeg)