TP 1

1. Les formats de texte

Téléchargez xml_1-tp.zip et décompressez l'archive (clic droit > Extraire tout)

Le dossier apollinaire contient le même texte sous 6 formats différents (ne pas tenir compte du fichier style.css). Quelles sont les différences entre ces formats ?

– Combien de fichiers faut-il pour représenter le texte ?

– Quels logiciel permet d'ouvrir le fichier ? de modifier son contenu si cela est possible ? de modifier sa mise en forme (gras, italique, etc.) si cela est possible ? Outre le logiciel par défaut, tester différents logiciels en faisant un clic droit > « Ouvrir avec » (tester les logiciels listés dans la diapos sur les formats de fichier)

– Puis-je chercher un mot dans le texte ?

– Les différents éléments du texte (titre, titres de partie, corps du texte) sont-ils distinguées par une mise en forme graphique ?

– Changer l'extension de apollinaire.odt en apollinaire.zip (comment faire ?) et décompresser le dossier. Qu'y trouve-t-on ?

2. Encodage

Ouvrir dans Notepad++ le fichier encodage.txt (dans le dossier xml_1-tp), changer l'encodage dans le menu Encoding > ANSI (autre nom du ISO-8859-1). Sur une nouvelle ligne, taper le caractère é. Revenir à l'encodage UTF-8. Cet manipulation a simplement pour but de vous sensibiliser à l'importance de l'encodage : selon l'encodage coché, Notepad++ interprète différemment la suite de 0 et 1 qui constitue le fichier (et qui reste, elle, inchangée)

3. Optical character recognition (OCR)

À partir du fichier ocr/zola_1.jpg, comparez la qualité de ces deux logiciels d'OCR en ligne :

– https://www.onlineocr.net/

– http://www.newocr.com/

À partir d'un même logiciel d'OCR, comparez les résultats obtenus avec les fichiers zola_1.jpg, zola_2.jpg, zola_3.jpg. Essayez de comprendre d'où viennent les erreurs. Que se passe-t-il si on change la langue ?