OUTIL DE RECUPERATION DES RESULTATS D’UNE REQUETE SOUS GALLICA

MANUEL UTILISATEUR

**I. Requête avec mots simples > *rich\_pages.py***

* Télécharger tous les scripts dans un même répertoire
* Fichier à fournir :
  + fichier txt contenant la liste des mots de la requête (un terme par ligne, sans ligne vide)
  + ou fichier au format UNIMARC récupéré sur Gallica (avec les thèmes RAMEAU)
* Lancer le programme (voir README)
* Dans le terminal, il vous sera demandé de renseigner :
  + Quel type de fichier vous souhaitez fournir (txt ou UNIMARC)
  + Si vous souhaitez valider les termes un par un : cette étape permet de retirer des mots de la liste de vocabulaire
  + Si vous souhaitez effectuer une recherche souple : « o »génère une requête qui ressortira des documents contenant les termes à un ou deux caractères près, « n » ne ressortira que les documents qui contiennent exactement le terme. Il est conseillé d’effectuer une recherche souple pour couvrir les éventuels problèmes d’océrisation.
  + Le taux d’OCR minimal que vous souhaitez : si vous entrez « 90 », la requête ne ressortira que les documents ayant une qualité d’océrisation supérieure ou égale à 90%
  + Les paramètres de la requête (il est possible de renseigner plusieurs champs par paramètre)
    - Le type de document : *monographie*, *manuscrit*
    - L’auteur
    - Le titre
    - Le thème ou sous-thème (possibilité d’en renseigner uniquement un seul)
    - La librairie
    - La période : date de début et date de fin
  + Le nombre de pages que vous souhaitez conserver dans le résultat final
* Exécuter le script *jolie\_sortie.py* afin d’obtenir les résultats de manière claire (voir README)

**II. Requête par couple de mots > *rich\_croise.py***

* Télécharger tous les scripts dans un même répertoire
* Fichier à fournir :
  + fichier txt contenant une liste de mots (un terme par ligne, sans ligne vide)
  + et un autre fichier txt contenant une liste de mots (un terme par ligne, sans ligne vide)
* Lancer le programme (voir README)
* Dans le terminal, il vous sera demandé de renseigner :
  + Les deux fichiers txt qui serviront à former les couples de mots
  + La proximité maximale que vous souhaitez entre les deux mots du couple
  + Si vous souhaitez effectuer une recherche souple : « o »génère une requête qui ressortira des documents contenant les termes à un ou deux caractères près, « n » ne ressortira que les documents qui contiennent exactement le terme. Il est conseillé d’effectuer une recherche souple pour couvrir les éventuels problèmes d’océrisation.
  + Le taux d’OCR minimal que vous souhaitez : si vous entrez « 90 », la requête ne ressortira que les documents ayant une qualité d’océrisation supérieure ou égale à 90%
  + Les paramètres de la requête (il est possible de renseigner plusieurs champs par paramètre)
    - Le type de document : *monographie*, *manuscrit*
    - L’auteur
    - Le titre
    - Le thème ou sous-thème (possibilité d’en renseigner uniquement un seul)
    - La librairie
    - La période : date de début et date de fin
  + Le nombre de pages que vous souhaitez conserver dans le résultat final
* Exécuter le script *jolie\_sortie.py* afin d’obtenir les résultats de manière claire (voir README)

**II. Sortie du programme**

A la fin de l’exécution, vous aurez la création d’un nouveau dossier « Depot\_Gallica » dans votre répertoire. Ce dossier contient 5 fichiers :

* Tableau.xml : Tableau contenant pour chaque résultat de la requête la page, le lien ark, le nombre de termes (ou couples) trouvés, le nombre de fois où un terme (ou couple) est trouvé, les termes (ou couples) trouvés avec leur nombre d’occurrence, le taux d’OCR, le score moyen (multiplication des occurrences de chaque terme), le titre, l’auteur, la date d’édition, le lien du document sous Gallica, le lien de l’image sous Gallica, le lien de la page sous Gallica, le lien du document océrisé et l’extrait dans lequel le terme (ou couple) a été trouvé
* Extraits.txt : Extraits où l’on trouve les occurrences de mot (pour chaque document trouvé par la requête)
* Liens\_ARK.txt : Lien ark de chaque document trouvé
* Liste\_1.txt : Liste des termes (ou couples) testés
* Matrix.xml : Matrice ayant pour lignes les termes (ou couples) et pour colonnes les documents, une cellule correspondant au nombre d’occurrences d’un terme (ou couple) dans un document

**Remarque :**

Cet outil faisant appel à une API du site Gallica, il est possible de rencontrer des erreurs de réseau empêchant le chargement d’une page (erreur du style HTTPError). Il est dans ce cas conseillé de relancer la requête un peu plus tard. Durant l’exécution du programme, plusieurs fichiers intermédiaires sont créés puis supprimés à la fin : s’il y a une quelconque interruption du programme, il est possible que ces fichiers demeurent et il faut dans ce cas les supprimer à la main.