EXTRACTION DES TITRES ET BLOCS D’UNE ENCYCLOPEDIE NUMERISEE

MANUEL UTILISATEUR

**I. Récupération des données**

* Fichier txt généré à partir de l'url sur gallica + « /texteBrut »

*Exemple* : <https://gallica.bnf.fr/ark:/12148/bpt6k62608307/texteBrut>

* Copier/coller tout le texte dans un fichier txt

**II. Lissage préliminaire du document (à la main)**

* Enlever les parties introductives et la conclusion
* Enlever la table des matières
* Enlever les figures, tableaux synoptiques...
* Mettre en minuscule les deux premières lettres de certains termes problématiques récurrents

*Exemple* : « prEMIERE SECTION », « clASSE PREMIERE », « chAPITRE DEUXIEME »

* En profiter pour corriger les erreurs d'océrisation (surtout sur les titres) si on en repère
* Eventuellement régulariser la présentation des titres si l'on souhaite faciliter la recherche de règles de sélection

*Exemple* : mettre un point après chaque titre

* Cette étape permet également de repérer les règles de sélection qui seront pertinentes, puisque l’on parcourt l’ensemble du document

**III. Prétraitement pour encodage des caractères spéciaux**

SCRIPT *pretraitement.py*

* Input : fichier txt (à passer en ligne de commande ou bien demandé si oubli)
* Output : fichier txt avec les caractères spéciaux (notamment caractères accentués) remplacés par le caractère simple le plus proche

**IV. Extraction des titres et blocs**

SCRIPT *script\_extraction\_encyclopedie.py*

* Input : fichier txt
* Output : dossier contenant la liste des titres, la liste des doublons, la liste des titres en contexte (optionnel), la liste des titres en désordre alphabétique, la liste des titres à la suite ayant un mot commun, la liste des règles de sélection, deux fichier txt contenant chaque titre et son bloc extrait (l’un avec le titre contenu dans le bloc, l’autre sans)
* Entrer un numéro d’essai : numéro du dossier qui contiendra tous les fichiers d’output
* Règles de sélection (itérations successives permettant de les modifier afin de trouver les règles les plus précises possibles)

*Exemple* : "\n[A-Z][A-Z](?:[^.,\n])\*[A-Z](?: )\*\,""\n[A-Z][A-Z](?:[^.,\n])\*[A-Z](?: )\*\.""\n[A-Z][A-Z](?:[^.,\n])\*\)(?: )\*\,""\n[A-Z][A-Z](?:[^.,\n])\*\)(?: )\*\."

Pour aider (les fichiers à consulter sont générés automatiquement dans le dossier) :

* Titres en désordre alphabétique (sélectionner ceux à conserver) : permet de repérer les termes qui ne sont de manière évidente pas des titres

*Exemple* : DECA,

@MERRAIN, @MESURE , @METRE, @MEUBLE,

(Regarder directement dans le document pour les termes bizarres)

* Titres ayant des mots en commun (sélectionner ceux à conserver) : permet d’éviter les familles de mots si on le souhaite

*Exemple* : NOYER COMMUN,

@NOYAU, @NOYER , @NOYER COMMUN,

Remarque : Décider si l'on veut des titres très précis (auquel cas on prend toutes les déclinaisons d'une famille) ou non (on garde seulement le terme très général)

* Ajouter des titres (à la main)

*Exemple* : titre non sélectionné par les règles de sélection à cause d’une erreur d’océrisation

* Supprimer des titres

*Exemple* : « III. », « ART. »

* Tri des doublons : Conserver un titre parmi les doublons, en regardant le contexte

Possibilité de conserver aucun des doublons

Impossibilité d’en conserver plusieurs (conservation du doublon apparaissant le plus tard le cas échéant)

* Fournir un nouveau fichier txt contenant les titres "lissés" (voir section **V.**) : il s’agira des titres définitifs, qui seront notamment les noms des fichiers titre/bloc
* Permet de mettre les titres en meilleure forme (enlever la ponctuation, les ajouts éventuels)

**V. Polissage des termes**

SCRIPT *polissage.py*

* Input : fichier txt contenant la liste des titres
* Output : fichier txt contenant la liste des titres « lissés »
* Le programme : enlève les points/virgules en fin de titre, ainsi que tout ce qui se trouve entre parenthèses
* A la main : enlever d’éventuels autres caractères qui n’appartiennent pas au titre à proprement parler
* Essayer de ne pas modifier l'intérieur d'un titre (problème d'océrisation par exemple) : ce sera possible dans une étape ultérieure

*Exemple* : « PLACENTA (placenta seu receptacvlum seminis). » devient « PLACENTA »

**VI. Génération des fichiers titres-blocs**

SCRIPT *fichiers\_blocs.py*

* Input : fichier txt contenant chaque titre et son bloc extrait
* Output : dossier contenant tous les fichiers txt ayant pour nom un titre et pour contenu le bloc correspondant, ainsi qu’un fichier txt contenant la liste des titres
* Parcourir rapidement les fichiers et corriger à la main les éventuelles erreurs d'océrisation (modifier le nom du fichier + le mot dans le bloc + le mot dans la liste des titres), ainsi que les éventuelles erreurs de découpage