Consignes du projet - 3

Dans ce document, je vous présente les dernières consignes du projet.

- Après avoir aspiré les substances actives à partir du VIDAL et généré le dictionnaire « subst.dic », vous devrez écrire un deuxième script Python permettant d'**enrichir** le dictionnaire « **subst.dic** » avec de nouveaux médicaments par nom commercial ou par substance active, à partir du fichier « **corpus-medical.txt** » donné en argument. **L'encodage de ce fichier ne doit pas être modifié**.

  Le dictionnaire « subst.dic » après enrichissement, contiendra donc toutes les substances issues du VIDAL selon l'intervalle choisi **+** les nouveaux médicaments issus du corpus Le script d'enrichissement **doit garder une trace** des médicaments trouvés dans le fichier « corpus-medical.txt », en les stockant dans un autre fichier qui **doit** s'appeler « **subst_enri.dic** », qui **ne doit subir aucun post-traitement** (tri et suppression de doublons). Ce fichier doit être encodé en **« UTF-16 LE avec BOM »** (UCS-2 LE BOM).

- Les médicaments par nom commercial ou par substance active **issus de l'enrichissement** doivent également être affichés sur la console, avec un compteur commençant à 1.

- Générer un fichier nommé « infos2.txt » contenant :
  - le nombre de médicaments **issus de l'enrichissement** pour chaque lettre de l'alphabet ;
  - et le nombre total de médicaments **issus de l'enrichissement**.
- Vous devrez éliminer les doublons du dictionnaire « subst.dic » enrichi.
- Vous devrez aussi trier par ordre **croissant** (a-z) les éléments du dictionnaire « subst.dic » enrichi.
- Le dictionnaire enrichi **doit conserver son encodage de départ**, à savoir « **UTF-16 LE avec BOM** » (UCS-2 LE BOM).

- Construire un graphe d'extraction sous UNITEX, qui se base sur l'étiquette **<N+subst>** du dictionnaire « subst.dic », afin d'**extraire les occurrences de** « **posologies de traitement** » à partir du fichier « corpus-medical.txt ». Le graphe d'extraction **doit s'appeler** « **posologie.grf** ». Le résultat de cette extraction sera placé par UNITEX dans le fichier « **concord.html** » qui se trouve dans le dossier « corpus-medical_snt ».

Remarque 1 : une « posologie de traitement » contient le nom du médicament, le dosage du médicament, la dose (usuelle ou maximale), le rythme d'administration (ou fréquence d'administration), l'heure-moment de prise du médicament et la durée de traitement.
Par exemple :

SIMVASTATINE 20 mg : 1 cp/j à 8 heures pendant un mois
CYTARABINE 100 mg/m² de J1 à J7
ZOLPIDEM 10 mg 1 cp au coucher
METFORMINE 850 mg 3 fois par jour
SPECIAFOLDINE 5 mg : 1 cp matin
ALADACTONE 25 mg : 1 cp/jour le midi
INEXIUM 40 1 cp par jour le soir

Remarque 2 : une « posologie de traitement » n'est pas forcément précédée du token « posologie ». Il faut donc bien analyser le fichier « corpus-medical.txt » que je vous ai envoyé en pièce-jointe, afin de découvrir les différentes façons d'exprimer une « posologie de traitement ».

Remarque 3 : lors de la phase d'extraction, il est **nécessaire** d'utiliser comme ressource supplémentaire le dictionnaire système « **Delaf.bin** » fourni par UNITEX, afin de pouvoir exploiter les masques lexicaux comme **<PREP>, <DET> ou <PREPDET>**, etc. **Vérifiez aussi que vous avez bien « Delaf.inf » à côté du « Delaf.bin », afin que ce dernier puisse être exploité.**

Pour l'**évaluation de votre travail**, vous **devrez m'envoyer par mail** :

1- **Le script d'aspiration** : « aspirer.py » doit générer « subst.dic » et « infos.txt ». Ce script prend deux arguments :
   I. l'intervalle des pages à traiter, en respectant le format : **B-H, E-S** ou **A-W**, etc ;
   II. le port utilisé dans le fichier de configuration du serveur « Apache ».

2- **Le script d'enrichissement** : « enrichir.py » doit enrichir le DELAF « subst.dic » à partir du fichier « corpus-medical.txt » donné en argument. Ce script doit générer 3 fichiers :
   I. « subst.dic » (dictionnaire enrichi) ;
   II. « subt_enri.dic » (trace d'enrichissement) ;
   III. « infos2.txt ».

3- **Le script Python qui appelle UNITEX** : « unitex.py » doit appeler UNITEX à partir de l'emplacement **C:\.....\Unitex-GramLab\App>**
   a. **Pour appeler UNITEX, vous devrez utiliser le script du cours 7, slide 2 (Lancer UNITEX à partir d'un script Python). Pour plus de détails sur les programmes externes d'UNITEX, vous pouvez vous référer au chapitre 13 du manuel d'UNITEX.**
   b. Placer vos 3 scripts (aspirer.py, enrichir.py et unitex.py) dans l'emplacement C:\.....\Unitex-GramLab\App>
   c. Ce troisième script Python **doit exploiter** les **ressources** suivantes :
      I. le dossier « **corpus-medical_snt** » créé automatiquement à chaque lancement du script « unitex.py » ;
      II. le fichier : « **corpus-medical.txt** » ;
      III. le fichier : « **corpus-medical.snt** » ;
      IV. le fichier : « **Norm.txt** » (facultatif) ;
      V. le fichier : « **Alphabet.txt** » ;
      VI. le fichier : « **subst.dic** » ;

VII. le fichier : « **subst.bin** » ;
VIII. le fichier : « **delaf.bin** » ;
IX. le fichier : « **posologie.grf** » ;
X. le fichier : « **posologie.fst2** » ;
XI. le fichier : « **concord.ind** ».

4- **Le graphe d'extraction** : par exemple « posologie.grf » doit extraire à partir du fichier « corpus-medical.txt » les posologies de traitement.

Pour résumer, vous devrez m'envoyer **4 fichiers** :

- les 3 scripts **Python** ;
- et le **graphe d'extraction** au format **.grf**.

Le non respect des consignes du projet entraînera des pénalités lors de l'évaluation de votre travail.

La date limite d'envoi de votre projet par mail est fixée au jeudi 2 janvier à 23h59.

Cdt,
N.Z