Consignes du projet - étape 1

- Écrire un premier script Python permettant d'aspirer (collecter) les entités médicales de type **noms de médicaments par substance active** de A à Z, à partir des 26 pages HTML du dossier « VIDAL » que je vous ai mis en pièce-jointe.

- Générer en sortie un dictionnaire au format **.dic** (format DELAF vu en cours 4) encodé en **UTF-16 LE avec BOM** (UCS-2 LE BOM).

- Ce dictionnaire **doit s'appeler** « **subst.dic** » et doit contenir les noms de médicaments par substance active des 26 pages HTML du dossier « VIDAL ».

- Chaque entrée lexicale de ce dictionnaire doit être suivie par les informations (codes) **,.N+subst**

- L'information **N** est de type grammatical et l'information **subst** est de type sémantique.

- Vous devez donc obtenir une sortie ayant le format DELAF-UNITEX suivant :

    - **abacavir,.N+subst**
    - **abatacept,.N+subst**
    - **abciximab,.N+subst**
    - **abiratérone,.N+subst**
    - **…….**

- L'aspiration doit être faite en local sur votre machine. Pour ce faire, vous devrez installer une plate-forme de développement Web, comme par exemple : **WampServer, XAMPP** ou **EasyPHP-DevServer**, qui contiennent, entre autres, un serveur Web Apache.

Remarque : L'encodage UTF-8 sans BOM des pages HTML du dossier « VIDAL » ne doit pas être modifié.

- Ensuite, donner la possibilité à l'utilisateur de déterminer l'intervalle des pages à traiter, en respectant le format : B-H, E-S ou A-W, etc. Cet intervalle **est le premier argument du premier script Python** « aspirer.py ».

- Générer un fichier nommé « infos.txt » contenant :

    - le nombre d'entités médicales de type noms de médicaments par substance active du dictionnaire « subst.dic » généré préalablement, pour chaque lettre de l'alphabet ;

    - et le nombre total d'entités médicales de type noms de médicaments par substance active de ce dictionnaire.

- Donner également la possibilité à l'utilisateur de saisir le « port http », qui est précisé dans le « fichier de configuration du serveur Web Apache ». Ce port est le deuxième argument du **premier script Python**.

- Autrement dit, une fois que vous avez choisi votre port **manuellement** dans ce fichier de configuration, vous le mettez ensuite comme deuxième argument à votre script « aspirer.py ». Ce script doit exploiter ce port pour accéder à l'URL des 26 pages HTML du dossier VIDAL, qui seront accessibles en local.

Remarque : Le port http par défaut est le port « **80** ».

Remarque : Ce premier script python « aspirer.py » **doit impérativement avoir 2 arguments : l'intervalle d'aspiration et le port http.**

Remarque : votre script python **ne doit pas modifier (écrire dans) le fichier de configuration Apache lors de l'aspiration.**