TD 7 – Scripts et formats d’échange

CSV, XML, JSON, YAML, SQLite

Projection lexicale

On va essayer de lemmatiser un texte par *projection lexicale*.

On part de (1) un texte plein de mots, et (2) un lexique *forme fléchie* / *POS* / *lemme*. Exemple de lexique :

mondassions	MONDER	Vvc
monde	MONDE	Nc
mondes	MONDE	Nc
mondeville	µMONDEVILLE	Np
mondial	MONDIAL	Ag
mondiale	MONDIAL	Ag
mondiales	MONDIAL	Ag

À chaque mot du texte, on ajoute une les informations trouvées dans le lexique.

Évidemment, il peut y avoir des ambiguïtés quand 1 mot = plusieurs lemmes, p. ex. salut = le nom commun SALUT *ou* le verbe SALUER.

Il peut aussi y avoir des mots inconnus (dans ce cas là on met le lemme UNK).

- Par exemple, pour l’entrée :
  - **Fichier XML** : `<text>Salut le monde</text>`
  - **Lexique CSV** :

```
le    LE      Dt
monde MONDE   Nc
salut SALUT   Nc
salut SALUER  Vvc
```

En sortie, on affichera un mot par ligne, on concaténera chaque lemme avec sa POS, et on utilisera le caractère | pour les cas ambigus (= plusieurs lemmes possibles) :

- Sortie :

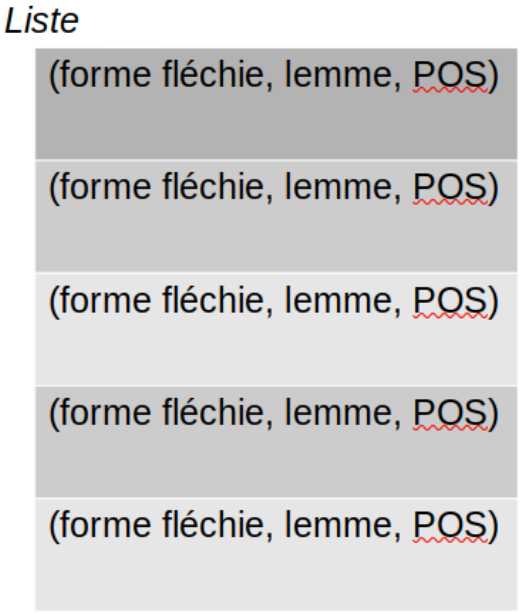
```
Salut SALUT_NC|SALUER_Vvc
le LE_Dt
monde MONDE_Nc
```

Dans notre cas, en entrée, on utilisera le fichier XML `2002-01-02.xml`, et le fichier CSV `lexicon.csv` comme lexique.

Exercice − structure du lexique

Quelle structure de données va-t-on utiliser pour stocker le lexique ? Dessinez-la.

Attention, un lexique peut contenir des millions de lignes. Pensez aux performances ! Par exemple, la structure de données ci-dessous va avoir des performances lamentables pour une tâche de lemmatisation.



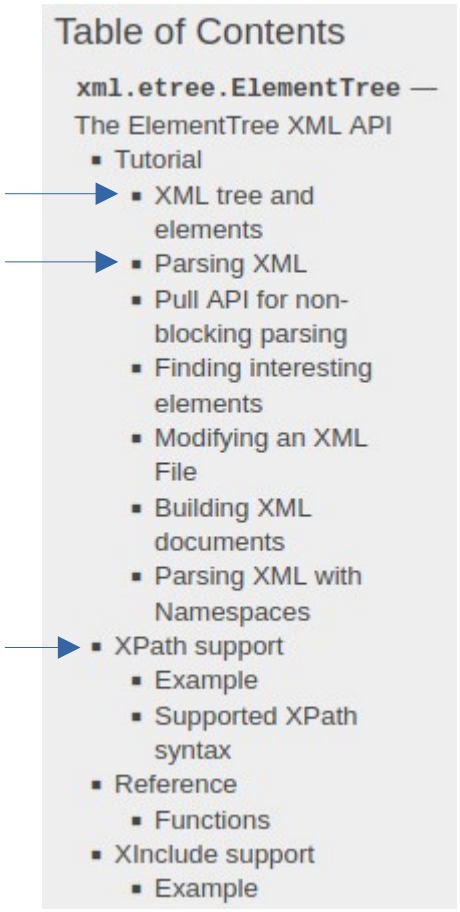
Exercice − chargement du lexique

Voir la doc de la bibliothèque qui lit/écrit du CSV :

https://docs.python.org/fr/3/library/csv.html

Le XML en Python

- Beautiful soup : https://beautiful-soup-4.readthedocs.io/en/latest/
- Xpath : https://docs.python.org/3/library/xml.etree.elementtree.html



Exercice − Lire du XML

Chargez le fichier avec *ElementTree*.

Pour chaque élément *p* du document, affichez son tag, ses attributs, et son texte. Faites de même avec *div*. Pourquoi ce dernier n'a-t-il pas de texte ?

Exercice − Accéder au texte d'un XML

https://docs.python.org/3/library/xml.etree.elementtree.html#xml.etree.ElementTree.Element.itertext

Affichez tout le texte de tous les éléments se trouvant dans ***p*** avec le méthode itertext() (voir la doc ci-dessus).

Exercice − Lemmatisation

Utilisez le lexique créé au début du TD pour afficher une version POS-lemmatisée du texte contenu dans les balises ***p***.

Fin de la sortie attendue :

```
Afin    AFIN_S|AFIN_SPS00
de  DE_Di|DE_S|DU_Dp|UN_Dn
tenter  TENTER_Vvn
d   CENT_Mc|CINQ_Mc|D_Nc
apaiser APAISER_Vvn
les IL_Pp|LE_Da
passions    PASSER_Vvc|PASSION_Nc
Magruder    UNK
patron  PATRON_Ag|PATRON_Nc
de  DE_Di|DE_S|DU_Dp|UN_Dn
la  IL_Pp|LA_Nc|LE_Da
North   µNORTH_Np
Coast   UNK
les IL_Pp|LE_Da
incite  INCITER_Vvc
à   À_S
participer  PARTICIPER_Vvn
à   À_S
une UN_Mo
fête    FÊTE_Nc|FÊTER_Vvc
organisée   ORGANISER_Ge|ORGANISÉ_Ag
pour    UNK
l   CINQUANTE_Mc|L_Nc|L_Xi
anniversaire ANNIVERSAIRE_Ag|ANNIVERSAIRE_Nc
d   CENT_Mc|CINQ_Mc|D_Nc
un  UNK
collaborateur    COLLABORATEUR_Nc
de  DE_Di|DE_S|DU_Dp|UN_Dn
CW  UNK
Biggs   UNK
Au  UNK
cours   COUR_Nc|COURIR_Vvc|COURS_Nc
de  DE_Di|DE_S|DU_Dp|UN_Dn
la  IL_Pp|LA_Nc|LE_Da
soirée  SOIRÉE_Nc
le  IL_Pp|LE_Da
magicien    MAGICIEN_Nc
Voltan  UNK
hyptonise   UNK
CW  UNK
Briggs  UNK
et  ET_Cc
Miss    MISS_Nc
Fitzgerald  UNK
à   À_S
l   CINQUANTE_Mc|L_Nc|L_Xi
aide    AIDE_Nc|AIDER_Vvc
du  UNK
scorpion    SCORPION_Nc
de  DE_Di|DE_S|DU_Dp|UN_Dn
jade    JADE_Ag|JADE_Nc
dont    DONT_Pr|DONT_S
le  IL_Pp|LE_Da
sortilège   SORTILÈGE_Nc
entraîne    ENTRAÎNER_Vvc
les IL_Pp|LE_Da
ennemis ENNEMI_Ag|ENNEMI_Nc
jurés   JURER_Ge|JURÉ_Ag|JURÉ_Nc
dans    DANS_S
de  DE_Di|DE_S|DU_Dp|UN_Dn
rocambolesques   ROCAMBOLESQUE_Ag
aventures   AVENTURE_Nc|AVENTURER_Vvc
1   UNK
h   H_Nc
42  UNK
A   A_Nc|AVOIR_Vvc
Epinal  UNK
au  UNK
Palace  PALACE_Nc
```