TD 8

XML type « base de données » et « document/mixte »

Les Bibliothèques

xml.dom, Beautiful soup, xml.etree, lxml

Bibliothèque xml.etree.ElementTree

Exercice

Fonctions : `ET.parse(file_name).getroot()` et `ET.tostring(node, encoding='unicode')`

Charger et afficher 2002-01-02.xml avec la bibliothèque .

Le XML affiché est-il le même que celui du fichier ?

Exercice

Fonction : `ET.register_namespace(namespace_id, namespace_name)`

Ajouter un namespace par défaut (namespace_id = chaîne vide).

Exercice

Fonction : `ET.tostringlist(node, encoding='unicode' method='xml')`

Afficher le XML sous forme de liste, un élément de la liste par ligne.

Exercice

Fonctions : `node.find(xpath_expression)`, *node*`.findall(xpath_expression)`, attributs `node.tag`, `node.text` et `node.attrib`

Trouver le bloc availability, et afficher son nom, son texte et ses attributs.

Trouver le bloc p du bloc availability, et afficher son nom, son texte et ses attributs.

Trouver les blocs change, et afficher leur nom, leur texte et leurs attributs.

Trouver le bloc title, et afficher son nom, son texte et ses attributs.

Documents XML « document/mixte » (*vs* « base de données »), comment accéder au texte ?

Exercice − tostringlist et itertext

Affichez tout le texte du bloc *title* :

- Avec `ET.tostringlist(node, encoding='unicode' method='xml')`
  - Attention, n'affichez que le texte !
- Avec `node.itertext()`

Exercice − tail

Utiliser les attributs `node.text` et `node.tail` pour ajouter des [crochets] autour de chaque nœud texte du document.

Avec la bibliothèque xml, peut-on facilement trouver le tag de l'élément parent de chaque nœud texte ?

Exercice − Lemmatisation

Utilisez le lexique créé au début du TD 7 pour afficher (sans modifier le XML) une version POS-lemmatisée du texte contenu dans les balises ***p***.

Fin de la sortie attendue :

```
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
```

Exercice

Affichez cette fois le document XML (en gardant les balises), où tous les blocs ***p*** auront été lemmatisés dans le format [*forme lemma1_pos1|lemma2_pos2*] [*forme lemma1_pos1|lemma2_pos2*] [*forme lemma1_pos1|lemma2_pos2*], etc.

La fin du document doit ressembler à :

```
<head>LE SORTILEGE DU SCORPION DE JADE</head>
<p>[Film FILM_Nc] [présenté PRÉSENTER_Ge|PRÉSENTÉ_Ag|PRÉSENTÉ_Nc] [en EN_Pp|
EN_S] [version VERSION_Nc] [originale ORIGINAL_Ag|ORIGINAL_Nc|ORIGINALE_Nc]
[américaine AMÉRICAIN_Ag|AMÉRICAIN_Nc|AMÉRICAINE_Nc] [sous UNK] [titrée
TITRER_Ge|TITRÉ_Ag]</p>
<p>[New µNEW_Np] [York µYORK_Np] [1940 UNK] [Betty UNK] [Ann UNK] [a A_Nc|
AVOIR_Vvc] [été ÉTÉ_Nc|ÊTRE_Ge|ÊTRE_Vvc] [engagée ENGAGER_Ge|ENGAGÉ_Ag|
ENGAGÉ_Nc] [pour UNK] [moderniser UNK] [les IL_Pp|LE_Da] [assurances UNK] [North
µNORTH_Np] [Coast UNK] [L CINQUANTE_Mc|L_Nc|L_Xi] [énergique ÉNERGIQUE_Ag] [Miss
MISS_Nc] [Fitzgerald UNK] [affiche AFFICHE_Nc|AFFICHER_Vvc] [d CENT_Mc|CINQ_Mc|
D_Nc] [emblée EMBLER_Vvc] [ses SON_Ds] [ambitions AMBITION_Nc] [et ET_Cc] [une
UN_Mo] [rationalité RATIONALITÉ_Nc] [à À_S] [tout UNK] [crin CRIN_Nc] [en EN_Pp|
EN_S] [déclarant DÉCLARANT_Ag|DÉCLARANT_Nc|DÉCLARER_Ga] [la IL_Pp|LA_Nc|LE_Da]
[guerre GUERRE_Nc] [aux UNK] [intuitions INTUITER_Vvc|INTUITION_Nc] [géniales
GÉNIAL_Ag] [et ET_Cc] [aux UNK] [méthodes MÉTHODE_Nc] [obsolètes OBSOLÈTE_Ag]
[de DE_Di|DE_S|DU_Dp|UN_Dn] [CW UNK] [Briggs UNK] [l CINQUANTE_Mc|L_Nc|L_Xi]
[enquêteur ENQUÊTEUR_Ag|ENQUÊTEUR_Nc] [vedette VEDETTE_Nc] [de DE_Di|DE_S|DU_Dp|
UN_Dn] [la IL_Pp|LA_Nc|LE_Da] [compagnie COMPAGNIE_Nc] [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]</p>
<p>[A A_Nc|AVOIR_Vvc] [Epinal UNK] [au UNK] [Palace PALACE_Nc]</p>
</div>
</div>
</div>
</div>
```

</body>
</text>
</TEI>