Exercices : GLAWI et parsing SAX en Java

Franck Sajous (CLLE-ERSS, CNRS & Université de Toulouse 2)

Les données à télécharger sont accessibles à l'adresse : `http://fsajous.free.fr/`

1 Prise en main

Écrivez un programme qui lance l'analyse d'un fichier XML et affichez, par exemple, les balises ouvrantes. Puis les attributs, les portions de texte comprises entre une paire de balises donnée, etc. Comme d'habitude, commencez par tester votre programme sur un fichier de taille réduite.

2 Extractions à partir de GLAWI

1. Parsez un extrait de GLAWI et extrayez sa nomenclature.
2. Établissez une table de fréquence des parties du discours de votre extrait
3. Quels sont les différents *types* de marques lexicographiques (*linguistic labels*) dans votre extrait (avec leur fréquence)?
4. Quelles sont les 5 marques de domaine les plus fréquentes dans votre extrait ?
5. Pour le domaine le plus fréquent, construire un lexique composé des entrées :
   – dont au moins une définition d'une section POS porte cette marque
   – dont toutes les définitions d'une section POS porte cette marque
6. Extraire les entrées et les gloses associées qui portent une marque de néologie
7. Extraire les entrées ayant plusieurs gloses et dont seulement une glose porte une marque de néologie
8. Extraire les entrées importées du Littré
9. Extraire les entrées importées du Littré qui portent une marque de néologie
10. Extraire les entrées importées du Littré dont toutes les gloses d'une définition portent une marque de néologie

3 Inspecteur de structure

Écrire un programme qui permet de donner un aperçu de l'arborescence d'un document XML et testez-le sur votre extrait de GLAWI.
Format de sortie attendu :

```
1       split
23514   split/article
23514   split/article/meta
804     split/article/meta/category
1952    split/article/meta/import
767     split/article/meta/reference
126     split/article/meta/spellingVariation
23514   split/article/pageId
23514   split/article/text
2716    split/article/text/etymology
2840    split/article/text/etymology/etym
180     split/article/text/etymology/etym/labels
199     split/article/text/etymology/etym/labels/label
2840    split/article/text/etymology/etym/txt
2840    split/article/text/etymology/etym/xml
...
2298    split/article/text/etymology/etym/xml/i
...
```

```
24423   split/article/text/pos
24423   split/article/text/pos/definitions
30334   split/article/text/pos/definitions/definition
3807    split/article/text/pos/definitions/definition/example
14      split/article/text/pos/definitions/definition/example/labels
14      split/article/text/pos/definitions/definition/example/labels/label
...
```

Ou par ordre de fréquence :

```
41671   split/article/text/pos/definitions/definition/gloss/xml/innerLink
30334   split/article/text/pos/definitions/definition
30326   split/article/text/pos/definitions/definition/gloss/xml
30326   split/article/text/pos/definitions/definition/gloss/txt
30326   split/article/text/pos/definitions/definition/gloss
25487   split/article/text/pos/inflectionInfos/inflectedForm
25191   split/article/text/pos/definitions/definition/gloss/xml/i
24423   split/article/text/pos/definitions
24423   split/article/text/pos
23963   split/article/text/pos/pronunciations/pron
23877   split/article/text/pos/pronunciations
23514   split/article/title
23514   split/article/text
23514   split/article/pageId
23514   split/article/meta
23514   split/article
...
```

Idem en incluant les attributs :

```
1       split
23514   split/article
23514   split/article/meta
804     split/article/meta/category
...
126     split/article/meta/spellingVariation
126     split/article/meta/spellingVariation/@norm
...
180     split/article/text/etymology/etym/labels
199     split/article/text/etymology/etym/labels/label
199     split/article/text/etymology/etym/labels/label/@type
199     split/article/text/etymology/etym/labels/label/@value
...
66      split/article/text/etymology/etym/xml/date
525     split/article/text/etymology/etym/xml/foreignWord
525     split/article/text/etymology/etym/xml/foreignWord/@lang
158     split/article/text/etymology/etym/xml/foreignWord/@sense
127     split/article/text/etymology/etym/xml/foreignWord/@translit
24423   split/article/text/pos
14      split/article/text/pos/@demonym
21      split/article/text/pos/@equivFem
11      split/article/text/pos/@equivMasc
4414    split/article/text/pos/@gender
184     split/article/text/pos/@homoNb
24423   split/article/text/pos/@lemma
24423   split/article/text/pos/@locution
4511    split/article/text/pos/@number
24423   split/article/text/pos/@type
24423   split/article/text/pos/definitions
...
```