Générer une base iTrameur

Mot d'introduction

Le but de cette feuille est de créer **un nouveau script** afin de formater votre corpus afin de le rendre lisible par iTrameur. iTrameur [1] est un outil d'exploration de corpus et de textométrie.

Il faudra créer divers fichiers au cours de ce TD, à ranger selon l'architecture de dossiers suivante, qui enrichit celle déjà existante évoquée dans les feuilles précédentes. En **bleu** les dossiers censés exister au début de cette feuille, en *rouge* les dossiers à créer pour la séance. Les noms de type `lang1-1.txt` reprennent le nom du fichier d'URL correspondant :

```
├── URLs
│   ├── lang1.txt
│   └── lang2.txt
├── ...
├── contextes
│   ├── lang1-1.txt
│   ├── lang1-2.txt
│   ├── ...
│   ├── lang2-1.txt
│   ├── lang2-2.txt
│   └── ...
├── dumps-text
│   ├── lang1-1.txt
│   ├── lang1-2.txt
│   ├── ...
│   ├── lang2-1.txt
│   ├── lang2-2.txt
│   └── ...
└── itrameur
    ├── contextes-lang1.txt
    ├── contextes-lang2.txt
    ├── dumps-text-lang1.txt
    └── dumps-text-lang2.txt
```

Note 1 Nous travaillerons ici sur les fichiers de dump et de contextes. Il ne sera pas demandé d'avoir une connexion internet.

Exercice 1 Exemples de fichier base iTrameur

Des fichiers d'exemple pour iTrameur (appelés bases iTrameur) sont disponibles sur le git dans le dossier `exercices/itrameur`. Vous pourrez vous en servir comme base pour créer la structure de votre fichier. Un exemple de fichier à créer est le suivant. Il y a deux choses à préciser. La première est qu'il s'agit d'un pseudo-XML : il n'y a pas forcément de racine unique [2] et il y a un attribut directement lié à la balise, deux choses interdites dans la norme XML. Vous pourrez noter également la présence de symboles "§" après les balises `page`, ces symboles sont importants car ils sont les marqueurs de contexte par défaut dans iTrameur (sinon, le contexte est uniquement la ligne), pensez à les mettre.

```
<lang="fr">
<page="fr-1">
<text>Ici, le contenu du fichier fr-1.</text>
</page> §
<page="fr-2">
<text>Là, le contenu du fichier fr-2.</text>
</page> §
</lang>
```

L'idée est de créer un fichier par langue afin d'effectuer des analyses quantitatives dessus pour trouver des éléments notables pour ensuite effectuer une analyse qualitative.

1. `http://www.tal.univ-paris3.fr/trameur/iTrameur/`
2. ici nous en avons une à des fins d'affichage dans l'outil, mais ce n'est pas requis

Exercice 2 Travail sur les dumps textuels

Créez avant de lancer les scripts suivants le dossier `itrameur`.

Créez un script `make_itrameur_corpus.sh`. Ce script prendra en argument :

— un dossier

— un nom de base qui correspond au nom du fichier URL sans son extension (`lang1` ou `lang2` dans l'exemple plus haut)

Le script devra créer une base iTrameur qui prendra la forme indiquée dans l'Exercice 1. Pour ce faire, il y aura deux étapes :

1. créer un fichier par langue
2. concaténer les différents fichiers pour obtenir une base globale pour toutes les langues

Pour 1., il faut itérer sur les différents fichiers dump (1 par URL), et les intégrer à une balise page et mettre le contenu textuel (dump) dans une balise text, comme indiqué plus haut. Pour le nom de base `lang1`, le fichier écrit aura alors le nom `dump-lang1.txt` et sera placé dans le dossier `itrameur`.

Pour 2., il suffit de concaténer l'ensemble des fichiers dans un fichier final appelé `dump.txt`

Important Le fichier iTrameur est un fichier XML, pour ne pas avoir de problème d'interprétation, il faut gérer les entités HTML/XML. On fera donc des substitutions avec la commande `sed`.

Si on souhaite garder une trace de ces éléments :

— & remplacé par &amp ;

— < remplacé par &lt ;

— > remplacé par &gt ;

On peut aussi décider de les supprimer avec la commande `td` : on supprime donc le jeu de caractères "&<>".

Exercice 3 Travail sur les contextes

Refaites les manipulations précédentes, mais en les appliquant aux fichiers de contexte. Ajoutez le tout dans le script `make_itrameur_corpus.sh`. Les fichiers créés devront avoir des noms de type `contexte-lang1.txt` et le fichier final sera `contexte.txt` dans le dossier `itrameur`.

Exercice 4 Travail sur iTrameur

En utilisant la feuille de travail sur iTrameur, chargez votre base sur le iTrameur et analysez votre corpus.

Une fiche de travail et un turotiel sont disponibles sur icampus dans la section `Ressources pour le cours`, rubrique `Documentation iTrameur`.

Exercice 5 Utiliser des scripts directement

L'un des intérêts d'iTrameur est de venir avec une interface graphique afin de rendre son utilisation plus simple. Cependant, certaines fonctionnalités peuvent être plus compliquées d'accès voire non-disponibles dans la version en ligne, comme :

— l'annotation morphosyntaxique

— la recherche par expression régulière

À cet effet, des scripts vous sont fournis sur icampus dans les `ressources supplémentaires` afin de palier ces difficultés. La documentation minimale est fournie dans le readme que vous pouvez ouvrir avec un lecteur en ligne fourni par l'UC Louvain `https://rsted.info.ucl.ac.be`. Les scripts qui ont également une aide spécifique chacun.