Corpus : aller plus loin

Mot d'introduction

Le but de cette feuille est d'enrichir le script de la feuille précédente (Validation et construction du corpus depuis des URL) afin d'obtenir des informations supplémentaires pour analyser votre corpus.

Il faudra créer divers fichiers au cours de ce TD, à ranger selon l'architecture de dossiers suivante. Les noms de type `fich1-1.txt` ou `fich1-1.html` reprennent le nom du fichier d'URL correspondant :

```
├── URLs
│   ├── fich1.txt
│   └── fich2.txt
├── aspirations
│   ├── fich1-1.html
│   ├── fich1-2.html
│   ├── ...
│   ├── fich2-1.html
│   ├── fich2-2.html
│   └── ...
├── contextes
│   ├── fich1-1.txt
│   ├── fich1-2.txt
│   ├── ...
│   ├── fich2-1.txt
│   ├── fich2-2.txt
│   └── ...
├── dumps-text
│   ├── fich1-1.txt
│   ├── fich1-2.txt
│   ├── ...
│   ├── fich2-1.txt
│   ├── fich2-2.txt
│   └── ...
├── programmes
│   ├── script1.sh
│   ├── script2.sh
│   └── ...
└── tableaux
    ├── fich1.html
    └── fich2.html
```

Note 1 Vous avez la possibilité de créer ces dossiers avant de les remplir, vous pourrez les laisser vides. Par défaut, un dossier vide n'est pas suivi sur git. Il est possible de suivre un dossier vide en introduisant un fichier nommé `.gitkeep` (sans contenu) à l'intérieur d'un dossier vide. Git suivra alors le dossier.

Exercice 1 Télécharger le script de correction type au besoin

En cas de besoin, télécharger le dossier `debut_seance8` et lancez-le sur un de vos fichiers d'URL afin de vous assurer qu'il fonctionne bien et vous permet d'obtenir un tableau attendus.

Exercice 2 Sauvegarder la page aspirée et le dump textuel

Dans la feuille précédente, ces deux données n'étaient pas stockées, c'est le moment de changer cela :
— stocker les pages aspirées par cURL dans le dossier `aspirations` ;
— stocker les dumps textuels récupérés avec Lynx dans le dossier `dumps-text`.
Il est conseillé de suivre les conventions de nommage indiquées dans l'introduction.

Exercice 3 Compter les occurrences du mot étudié

Une fois le dump textuel de votre page effectué, comptez le nombre d'occurrences de votre mot d'étude sur chacune des pages. Cette information est à ajouter à la suite de votre tableau dans une colonne "compte".

Exercice 4 Contexte

Toujours sur le dump textuel, récupérer des contextes d'apparition de votre mot dans le contexte (lignes précédentes et suivantes).

Il faudra sauvegarder ces contextes dans le dossier `contextes`.

Exercice 5 Concordances

À l'aide de la commande sed, transformer le texte afin d'obtenir un concordancier autour de votre mot d'étude. Un concordancier, en HTML, sera un tableau à trois colonnes dont la colonne centrale sera votre mot, et les colonnes de gauche et de droite respectivement les contextes gauches et droit.

Vous trouverez un exemple d'un tel tableau à cette adresse :
`https://yoanndupont.github.io/PPE/exemple_conc.html`