Corpus : aller plus loin

Mot d'introduction

L'objectif de cette séance est de passer d'un travail individuel à un travail en groupe et d'enrichir nos tableaux avec de nouvelles colonnes.

Il faudra pour cela commencer à télécharger les documents HTML correspondant à vos URL et prétraiter le texte qu'ils contiennent.

Avant tout, il nous faut établir une arborescence de dossiers pour organiser clairement notre travail.

Il faudra créer divers fichiers au cours de ce TD, à ranger selon l'architecture de dossiers suivante. Les noms de type `lang1-1.txt` ou `lang1-1.html` reprennent le nom du fichier d'URL correspondant (le nom du fichier doit indiquer la langue traitée) :

```
├── URLs
│   ├── lang1.txt
│   └── lang2.txt
├── aspirations
│   ├── lang1-1.html
│   ├── lang1-2.html
│   ├── ...
│   ├── lang2-1.html
│   ├── lang2-2.html
│   └── ...
├── 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
│   └── ...
├── programmes
│   ├── script1.sh
│   ├── script2.sh
│   └── ...
├── tableaux
│   ├── lang1.html
│   └── lang2.html
└── concordances
    ├── lang1-1.html
    ├── ...
    ├── lang2-1.html
    └── lang2-1.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 Création du git et de l'arborescence du projet

— Si ce n'est pas déjà fait, créer un dépôt git pour votre groupe. La personne qui crée le dépôt doit s'assurer d'ajouter ses camarades en "collaborateurs" (menu *settings*).
— Vous pouvez ensuite reprendre l'un de vos scripts de mini-projets individuels comme point de départ, ou combiner vos solutions pour créer un nouveau script. Celui-ci sera à sauvegarder dans le dossier `programmes/`.
— La personne propriétaire du dépôt peut alors faire un **commit** et le **push** afin que les autres membres du groupe puissent simplement le **pull**.

Exercice 2 Ne traiter que les URL correctes

Il est possible que certaines de vos pages renvoient un code HTTP d'erreur. On ne souhaite pas intégrer ces pages à notre corpus. Lorsqu'on rencontre une page pour laquelle une requête renvoie un code d'erreur, on souhaite afficher un message d'erreur indiquant la page concernée et ne pas la traiter. On écrit dans le fichier HTML l'absence de résultat à chaque colonne de la ligne et on passe à l'URL suivante. On écrit également dans l'erreur standard un message indiquant que l'URL ne sera pas traitée.

Exercice 3 Sauvegarder la page aspirée et le dump textuel

Jusqu'à présent, les pages html et leur contenu textuel 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` (revoir les diapos de la séance 04-web) ;
— ajouter les liens correspondants dans deux nouvelles colonnes de vos tableaux.
Suivre les conventions de nommage indiquées dans l'introduction.

Assurez vous d'être bien synchronisés avec votre git et entre membres du projet.

Exercice 4 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 5 Contexte

Toujours sur le dump textuel, récupérer des contextes d'apparition de votre mot dans le contexte

On cherchera à isoler les occurrences de votre mot avec une ou deux lignes précédentes et suivantes. (consulter la documentation de **grep** pour vous aider).

Il faudra sauvegarder ces contextes dans le dossier `contextes`. Et ajouter une colonne à votre tableau.

Exercice 6 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 gauche et droit.

Vous trouverez un exemple d'un tel tableau à cette adresse :
`https://pmagistry.github.io/PPE1-2024/exemple_conc.html`