Programmation et projet encadré - L7TI005

Lancement du projet de groupe

Yoann Dupont yoann.dupont@sorbonne-nouvelle.fr
Pierre Magistry pierre.magistry@inalco.fr

2022-2023

Université Sorbonne-Nouvelle
INALCO
Université Paris-Nanterre

Un petit mot avant de commencer

Certains services down

Les services suivants hébergés à USN sont *down* :

- plurital.org
- iTrameur (qu'on utilisera normalement la semaine prochaine)

On vous tient au courant dès qu'on en sait plus !

Ce qu'on attend aujourd'hui

Feuille d'exercices

Une feuille d'exercices avec tout ce qu'il y aura à faire vous sera donné. On va résumer ici ce qu'il y a dedans et ce qu'on va faire en plus.

Chaque exercice va demander d'ajouter une colonne au tableau existant.

Préparer le texte 1 : récupérer les sources de données

Jusqu'à présent, on a juste récupéré des informations pour voir si on arrivait à communiquer avec les différents sites web.

Préparer le texte 1 : récupérer les sources de données

Jusqu'à présent, on a juste récupéré des informations pour voir si on arrivait à communiquer avec les différents sites web.

À partir de là, on va commencer les vrais traitements. Mais il faut déjà avoir du matériau :

1. les pages HTML
2. les contenus textuels des pages HTML

Préparer le texte 1 : récupérer les sources de données

Jusqu'à présent, on a juste récupéré des informations pour voir si on arrivait à communiquer avec les différents sites web.

À partir de là, on va commencer les vrais traitements. Mais il faut déjà avoir du matériau :

1. les pages HTML
2. les contenus textuels des pages HTML

À partir des outils déjà vus, on va récupérer le tout et le stocker.

Préparer le texte 2 : gérer l'encodage

On a vu que les pages pouvaient avoir des encodages différents. Quand on veut faire des traitements, on normalise certains aspects du texte. Le premier aspect à normaliser est l'encodage : on voudra que tout soit en UTF-8.

Préparer le texte 2 : gérer l'encodage

On a vu que les pages pouvaient avoir des encodages différents. Quand on veut faire des traitements, on normalise certains aspects du texte. Le premier aspect à normaliser est l'encodage : on voudra que tout soit en UTF-8.

En partant de ce qu'on avait la semaine dernière :

1. vérifier l'encodage d'une page web
2. convertir le texte en UTF-8 si son encodage est différent (demandera peut-être un peu de recherche)

Préparer le texte 2 : gérer l'encodage

On a vu que les pages pouvaient avoir des encodages différents. Quand on veut faire des traitements, on normalise certains aspects du texte. Le premier aspect à normaliser est l'encodage : on voudra que tout soit en UTF-8.

En partant de ce qu'on avait la semaine dernière :

1. vérifier l'encodage d'une page web
2. convertir le texte en UTF-8 si son encodage est différent (demandera peut-être un peu de recherche)

À la fin, tous les textes qu'on va traiter seront en UTF-8.

Premiers traitements : compter (et compter encore)

Une fois le texte préparé, on va compter les occurrences de chaque mot dans une page.

On fera ensuite les comptages des bigrammes.

Premiers traitements : les contextes d'apparition

Tout simplement, récupérer un empan de texte qui contient notre mot d'intérêt.
On prendra comme contexte les lignes qui sont autour de notre mot.

Premiers traitements : un concordancier

Un concordancier est un moyen pratique de regarder précisément les usages d'un mot donné. Il s'agit d'une version plus lisible de l'exercice précédent.

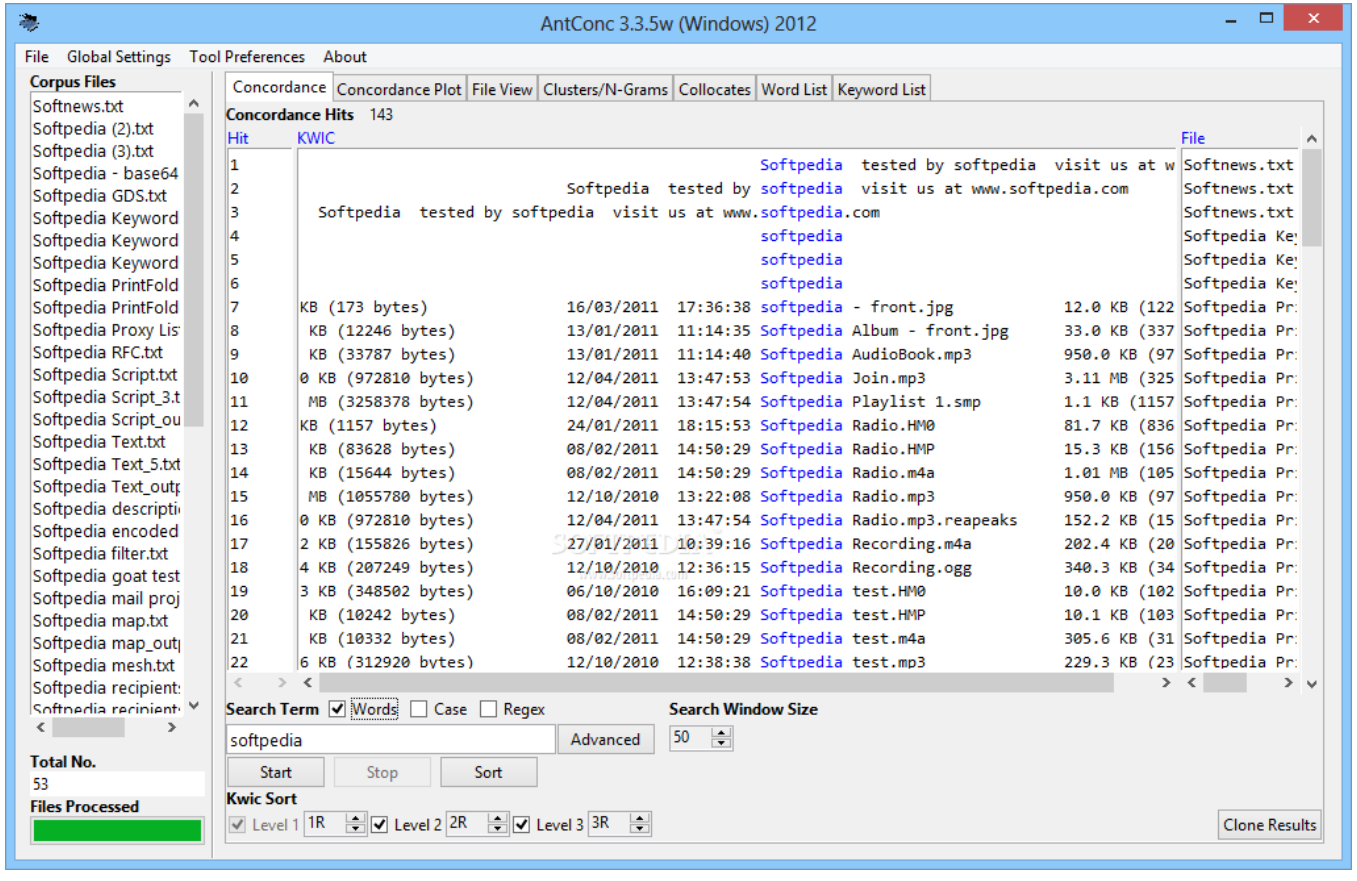

Figure 1: Un exemple de concordancier avec AntConc (source de l'image : https://www.softpaz.com

Point projet

Discutons !

On va passer vous voir pour discuter un peu de vos projets, des vos choix de mots et de langue.

Le rendu

Le rendu sera en groupe (faire au moins un rendu par groupe) :

Le rendu

Le rendu sera en groupe (faire au moins un rendu par groupe) :

1. page d'accueil du projet liée aux tableaux (il faudra mettre en commun vos scripts, sorties et styles)

Le rendu

Le rendu sera en groupe (faire au moins un rendu par groupe) :

1. page d'accueil du projet liée aux tableaux (il faudra mettre en commun vos scripts, sorties et styles)
2. vérifiez bien que les liens marchent aussi sur internet (pas uniquement votre ordinateur)

Le rendu

Le rendu sera en groupe (faire au moins un rendu par groupe) :

1. page d'accueil du projet liée aux tableaux (il faudra mettre en commun vos scripts, sorties et styles)
2. vérifiez bien que les liens marchent aussi sur internet (pas uniquement votre ordinateur)
3. c'est en général à ce moment qu'on a souvent des problèmes de mise en commun. Toujours pull avant commit et push.

Le rendu

Le rendu sera en groupe (faire au moins un rendu par groupe) :

1. page d'accueil du projet liée aux tableaux (il faudra mettre en commun vos scripts, sorties et styles)
2. vérifiez bien que les liens marchent aussi sur internet (pas uniquement votre ordinateur)
3. c'est en général à ce moment qu'on a souvent des problèmes de mise en commun. Toujours pull avant commit et push.

Une fois le tout terminé, on créé le *tag* `projet1`. Le rendu sur icampus devra contenir le lien vers le tag du projet, ainsi que les noms et github des différents membres du groupe.

Il faut continuer à avoir un journal, mais on va passer à un journal de groupe. On indiquera qui écrit l'entrée du journal pour plus de clarté.