Refonte du site des Cahiers de la Linguistique (clf.unige.ch)

Nouvelles Technologies d'informations et communications, 2011

Hervé Nindanga
nindang2@etu.unige.ch

26 mai 2011

Contents

1 Introduction

Le site des cahiers de linguistique française est un site web regroupant les revue de l'unité de linguistique française du département de linguistique de l'université de Genève.

Il a pour vocation de :

- mettre en valeur les recherches en linguistique française ;
- permettre aux chercheurs (doctorants) de publier leurs recherches ;
- publier les actes de colloques du Département ;
- permettre l'accès en ligne, et gratuit, aux articles des numéros antérieurs des CLF.

Pour ce projet, nous avons pour but de faire une refonte du site wue des cahiers de la linguistique française.

Le projet a été divisé en deux parties :

- La conception du site web ;
- l'indexation.

La partie du projet qui m'a été attribué est l'indexation des documents, c'est à dire rendre le plus facile possible la manipulation des documents des revues.

Pour réaliser ce travail, je l'ai regroupé en trois parties :

1. La conversion des documents
2. La gestion des mots-clés
3. La création d'un nuage de mots-clés

2 Conversion des documents

Pour pouvoir indexer les documents, il faut que ces documents puissent être analyser. Malheureusement, la majorité des documents sont des documents scannés et ne sont pas analysables.

Pour ce faire, il nous faut d'abord les convertir en utilisant des outils qui puissent recupérer du texte sur des documents scannés ou des images.

Ces logiciels sont communement appelés OCR (Optical Character Recognition).

Un logiciel OCR est un outil qui nous permet de pouvoir extraire du texte dans des documents scannés ou des images, et de le sauvegarder dans un fichier où le texte peut être exploité.

3 Indexation

Après avoir converti tous les documents en document texte, on peut maintenant rechercher les mots qui correspondent au mieux au contenu informationnel de chaque document. Ces mots sont communement appelés des mots-clés.

Pour retrouver ces mots-clés, nous devons :

- extraire tous les mots du document ;
- supprimer les mots vides (stopwords) ;
- determiner l'occurrence de chaque mot dans le document ;
- ordonner les mots selon leur occurrence ;
- les mots, ayant le plus d'occurrence, seront considérés comme mots-clés.

3.1 Stopwords

stopwords, ou mot vide, sont des mots tellement courant qu'ils ne véhiculent aucune information.

Parmi ces mots, on peut distinguer les mots tels que : le, la, un, des, ...

Pour pouvoir ignorer ces mots vides pendant la recherche des mots-clés :

- j'ai d'abord stocké la liste de ces mots dans un fichier texte (stopwords.txt) ;
- puis j'ai implémenté, en Java, une application qui me permettra :
  - de parcourir ce fichier texte (stopwords.txt) ;
  - et de sauvegarder ces mots dans la base de données.

3.2 Recherche de mots-clés

Pour la recherche des mots, j'ai impléementé une application en Java qui me permettra de :

1. Parcourir un document PDF et en extraire le texte, à l'aide de la librairie PDFBox d'Apache ;
2. Vérifier si un mot fait partie de la liste des mots vides :
   (a) si oui, le mot sera ignorée ;
   (b) si non, le mot sera enregistré dans une liste ainsi que l'occurrence de ce mot.
3. Enfin, la liste de ces mots, et occurrences corréspondants, sera stocké dans la base de données.

3.3 Ajout et supprésion manuel de mots-clés

Malheureusement, on peut se heurter sur des cas où les mots-clés trouvés ne définissent pas exactement le document ou que les mots-clés le definissant ne sont pas suffisament fréquents.

Pour remédier à ce problème :

- j'ai ajouté un champ, où on peut ajouter des mots-clés, dans le formulaire d'ajout de document ;
- j'ai ajouté une page où on peut supprimer les mots-clés indésirables.

3.4 Création du nuage de mots-clés

Le nuage de mots-clés (tag cloud en anglais) est une représentation visuelle des mots-clé. Il nous permet d'afficher les mots-clés dans des polices de caractères différentes selon l'importance de chaque mot-clé.

Pour créer ce nuage de mots-clés :

- On récupère les mots-clés ajoutés manuellement ;
- On recupère les mots-clés du document ;
- On ajoute ces mots-clés dans un fichier XML ;
- On appelle l'application FLASH qui affichera les mots-clés stocker dans le fichier XML.

La création de ce nuage de mots-clés (en 3D) a été réalisé à l'aide d'un module FLASH développé pour les utilisateurs de WordPress.

4 Statistiques

Pour cette partie, j'ai installé un outil, trouvé sur le net, me permettant de mesurer quelques données statistiques tels que :

- le nombre de visites ;
- les pages les plus visités ;
- les pages visités par visiteurs ;
- etc...

Cet outil statistique se nomme PhpMyVisite.

5 Outils utilisés

5.1 OCR

Avec un bon nombre de logiciels OCR disponibles, j'ai choisit d'utiliser le logiciel "ABBYY Fine Reader".

Mon choix s'est porté sur ce logiciel car :

- Il est disponible pour toutes les plateformes à la différence des autres qui ne proposent pas de version pour Mac OS X ;
- Il a une interface utilisateur facile à utiliser ;
- Il reconstruit exactement la même mise en page que la page scannée ;
- Il est plus stable lors du traitement de documents volumineux ;
- Il est rapide et très éfficace.

5.2 Java

Pour la recherche et l'ajout des mots-clés, j'ai choisi d'utiliser le langage pour deux raisons :

1. Il m'a été plus facile de trouver la librairie (PDFBox) me permettant d'analyser un document PDF en Java que pour PHP ;
2. on a fait un travail similaire dans un autre cours avec le langage Java.

6 Conclusion

Ce projet m'a apporté de riches expériences tant sur la conception que sur la réalisation. Nous avons converti les documents à l'aide de logiciel OCR. Nous avons implémenté l'application avec PHP/MySQL, des outils très utilisés en exploitation d'applications web dynamiques, en y ajoutant des applications en Java. Malgré certaines difficultés durant le développement, surtout lié au problème d'encodage, nous avons pu obtenir une application qu'on peut déjà utilisé et intéger à l'applicatio finale.

En bref, ce cours sur le NTIC nous a permis d'améliorer notre connaissance et savoir faire en matière de développement d'application, de gestion et de manipulation de base de données.

7 Annexe

Modèle conceptuel des données

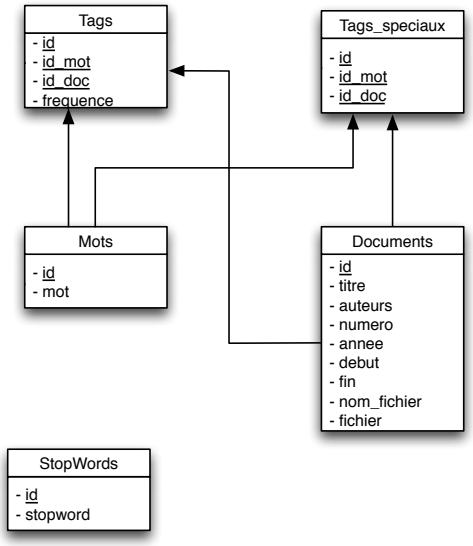
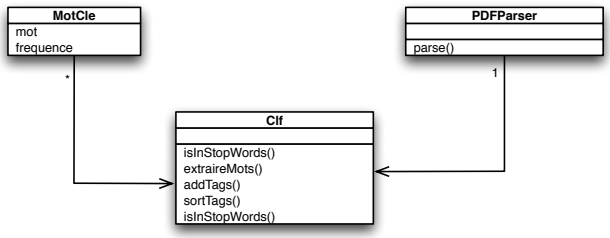


Diagramme de classes de l'archive "clf.jar"



Workflow du site