Convertisseur de fichier PDF en fichier TXT ou XML

Romain ALLEGRE
Enzo LLORCA
Dylan GELY

Abstract

Dans cette article, nous allons présenter notre système, ses fonctionnalités et sa méthode de fonctionnement, puis nous présenterons les résultats obtenus sur un corpus de test contenant des fichiers pdf et enfin nous conclurons sur l'efficacité du systeme.

1. Méthode

Le systeme est composé du fichier python convert_txt.py.

Voici l'ensemble des fonctions du convertisseur :

1. **afficher_liste_fichiers(dossier):**
   - Affiche la liste des fichiers dans le dossier spécifié.
   - Retourne la liste des fichiers.
2. **demander_fichiers_a_convertir():**
   - Demande à l'utilisateur d'entrer le numéro du fichier à convertir.
   - Retourne la liste des fichiers sélectionnés.
3. **get_txt_files_in_folder(input_folder):**
   - Renvoie la liste des fichiers texte (.txt) dans le dossier spécifié.
4. **convert_txt(input_folder, uniq=None):**
   - Convertit les fichiers PDF en fichiers texte (.txt) et les place dans un sous-dossier appelé "filetxt".
   - Si l'argument uniq est spécifié, il convertit uniquement le fichier ou la liste de fichiers spécifiés.
5. **extract_title_from_file(file_path):**
   - Extrait le titre du document à partir du fichier texte.
6. **extract_abstract_from_file(file_path):**
   - Extrait l'abstract du document à partir du fichier texte.
7. **extract_abstract_author(folder_path):**

- Extrait l'auteur du document à partir du fichier PDF.

8. extract_biblio_from_file(file_path):

- Extrait la bibliographie du document à partir du fichier texte.

9. extract_corps_from_file(file_path):

- Extrait le corps du document à partir du fichier texte.

10.extract_intro_from_file(file_path):

- Extrait l'introduction du document à partir du fichier texte.

11.extract_conclusion_from_file(file_path):

- Extrait la conclusion du document à partir du fichier texte.

12.extract_discussion_from_file(file_path):

- Extrait la discussion du document à partir du fichier texte.

13.donne_txt(input_folder):

- Crée un dossier "output_txt" et génère des fichiers texte (.txt) avec le nom du fichier d'origine, le titre, et l'abstract.

14.donne_xml(input_folder):

- Crée un dossier "output_xml" et génère des fichiers XML avec différentes sections du document (titre, auteur, abstract, etc.).

15.get_txt_files_in_folder(input_folder):

- Renvoie la liste de tous les fichiers dans le dossier spécifié.

16.main:

- Vérifie que le programme est appelé avec les arguments appropriés.
- Affiche la liste des fichiers dans le dossier spécifié.
- Demande à l'utilisateur de sélectionner les fichiers à convertir.
- Convertit les fichiers en texte et génère des fichiers supplémentaires au format souhaité (txt ou xml) en fonction de l'option spécifiée.

Lors de l'écriture de sa commande d'execution, le choix de générer un fichier xml ou txt pour contenir le texte parsé du pdf est possible :

-t : pour générer une sortie texte
-x : pour générer une sortie XML

La commande entière à executer est la suivante :

python convert_txt.py <chemin_dossier_fichiers_pdf> [-t | -x]

Apres avoir lancé la commande, les numéros des fichiers que l'on veut convertir sont demandés.

Voici un exemple:
choissisez un fichier: 1
choissisez un fichier: 3
choissisez un fichier: fin

2. Résultats

Précision de chaque pdf :

A Benders Decomposition Approach toCorrelation Clustering : 75%

A_memetic_algorithm_for_community_detectionin_signed_networks : 88%

An_Improved_Branch-and-Cut_Code_for_the_Maximum_Balanced_Subgraph_of_a_Signed_Graph : 86%

Cabrera_RESUMES_2019 : 77%

Conversational_Networks_for_Automatic_Online_Moderation : 77%

Dynamical_Models_Explaining_Socia_Balance_and_Evolution_of_Cooperation : 75%

Exact_Clustering_via_Integer_Programming_and_Maximum_Satisfiability : 66%

LDA_resume : 63%

Partitioning_large_signed_two-mode_networks:_Problems_and_prospects : 66%

Polibits_42_02 : 63%

3. Conclusion

La moyenne de précision du systeme est de 67,3 %.