Rapport sprint 5

Hamza Latrach, Marc Pelloux, Marieme Soda Mbeguere

November 30, 2023

Abstract

Dans le cadre de notre projet, nous avons développé un parseur dédié au traitement sélectif de fichiers PDF d'articles scientifiques. Lorsque le programme est lancé, un menu interactif permet à l'utilisateur de choisir spécifiquement les fichiers PDF à parser, offrant ainsi une approche ciblée plutôt que de traiter l'ensemble du répertoire.

Notre solution offre une sortie finale flexible, présentant les résultats sous deux formats distincts: un fichier texte (.txt) et un fichier XML (.xml). La structure du fichier XML est soigneusement conçue, incluant des sections telles que le nom du fichier d'origine, le titre du papier, la liste des auteurs avec leurs affiliations, le résumé de l'article, l'introduction, le développement, la conclusion, la discussion et les références bibliographiques.

Une fonctionnalité supplémentaire a été intégrée pour permettre à l'utilisateur de choisir le type de sortie souhaité, que ce soit en texte ou en XML, en utilisant des arguments spécifiques (-t — -x) lors de l'exécution du parseur.

En résumé, notre système offre une solution complète répondant aux besoins spécifiques de notre département. Il propose une flexibilité remarquable en permettant à l'utilisateur de choisir le convertisseur approprié et fournit des sorties textuelles et XML bien structurées, facilitant ainsi une analyse approfondie des articles scientifiques.

1 Methode

Le programme en langage C que nous avons développé offre plusieurs fonctionnalités essentielles pour la manipulation de fichiers PDF dans un répertoire spécifié. Pour la conversion des PDF en texte, la fonction PdfToText utilise la commande pdftotext. Les fichiers texte résultants sont regroupés dans un sous-répertoire appelé "resultat". Le programme intègre également diverses fonctions spécialisées dans l'extraction d'informations spécifiques, telles que trouver-titre, trouver-abstract, trouver-bibliographie, trouver-auteur, trouver-introduction, trouver-discussion, et trouver-conclusion, à partir des fichiers texte convertis.

La génération de fichiers de sortie est gérée en fonction de l'option spécifiée en ligne de commande (-t pour un fichier texte ou -x pour un fichier XML), produisant des fichiers structurés contenant les informations extraites. Dans la fonction principale (main), la boucle prend en compte les arguments en ligne de commande pour définir le dossier contenant les fichiers PDF. Les noms des fichiers texte convertis sont stockés temporairement dans un fichier appelé "namesOfFiles". Chaque fichier texte est ensuite analysé, et les informations pertinentes sont extraites en fonction de l'option de sortie choisie, en utilisant les fonctions dédiées énumérées précédemment.

En ce qui concerne la sécurité, il est important de noter que le programme recourt à des commandes système (system()) pour exécuter des commandes du terminal. De plus, l'allocation de mémoire est effectuée à l'aide de malloc. Un processus de nettoyage est mis en place pour supprimer les fichiers temporaires générés pendant l'exécution du programme, contribuant ainsi à maintenir l'intégrité du système.

2 resultat

la convertion txt compte les sections :

- Titre
- Auteur
- Abstract
- Bibliographie

la convertion xml quand à elle compte les sections supplémentaire :

- Preamble
- Introduction
- Conclusion
- Discussion

TEST PDF G1 :

Polibits_42_02.pdf
TXT :
Titre ok Auteur ok Abstract ok Biblio absent note 3/3
XML :
Preamble ok Titre ok Auteur ok Abstract ok Bibliographie absent Introduction ok Conclusion absent Discussion erroné note 5/7

LDA_resume.pdf
TXT :
Titre ok Auteur partiel Abstract ok Biblio absent note 2/3
XML :
Preamble ok Titre ok Auteur partiel Abstract ok Bibliographie absent Introduction trop Conclusion absent Discussion ok note 4/6

Conversational_Networks_for_Automatic_Online_Moderation.pdf
TXT :
Titre erroné Auteur erroné Abstract ok Biblio absent note 1/2
XML :
Preamble ok Titre erroné Auteur erroné Abstract ok Bibliographie absent Introduction ok Conclusion absent Discussion absent note 3/4

Dynamical _Models_Explaining_Socia_Balance_and_Evolution_of_Cooperation.pdf
TXT :
Titre ok Auteur ok Abstract ok Biblio absent note 3/3
XML :
Preamble ok Titre ok Auteur ok Abstract ok Bibliographie absent Introduction trop Conclusion trop Discussion absent note 4/6

An_Improved_Branch-and-Cut_Code_for_the_Maximum_Balanced_Subgraph_of_a_Signed_Graph.pdf
TXT :
Titre ok Auteur ok Abstract ok Biblio absent note 3/3
XML :
Preamble ok Titre ok Auteur ok Abstract ok Bibliographie absent Introduction ok Conclusion absent Discussion absent note 5/5

AA_memetic_algorithm_for_community_detectionin_signed_networks.pdf
TXT :
Titre erroné Auteur absent Abstract ok Biblio absent note 1/2
XML :
Preamble ok Titre erroné Auteur absent Abstract ok Bibliographie absent Introduction ok Conclusion ok Discussion absent note 4/5

Exact_Clustering_via_Integer_Programming_and_Maximum_Satisfiability.pdf
TXT :
Titre erroné Auteur absent Abstract ok Biblio absent note 1/2
XML :
Preamble ok Titre erroné Auteur absent Abstract ok Bibliographie absent Introduction ok Conclusion trop Discussion absent note 3/5

Partitioning_large_signed_two-mode_networks:_Problems_and_prospects.pdf
TXT :
Titre ok Auteur erroné Abstract absent Biblio absent note 1/2
XML :
Preamble ok Titre ok Auteur erroné Abstract absent Bibliographie absent Introduction ok Conclusion erroné Discussion ok note 4/6

Cabrera_RESUMES_2019.pdf
TXT :
Titre ok Auteur erroné Abstract absent Biblio absent note 1/2
XML :
Preamble ok Titre ok Auteur erroné Abstract absent Bibliographie absent Introduction ok Conclusion trop Discussion ok note 4/6

3 Conclusion

La conclusion des résultats obtenus suite à l'analyse des fichiers PDF traités par notre parseur révèle plusieurs aspects significatifs. En premier lieu, nous avons évalué la performance du parseur en examinant divers articles scientifiques, chacun présentant des caractéristiques uniques et des structures variées.

Dans l'ensemble, la conversion vers le format texte a démontré une grande précision, en particulier pour des éléments tels que le titre, les auteurs, et les résumés. Cependant, des lacunes ont été identifiées, notamment dans la détection de la bibliographie. Ces lacunes pourraient être attribuées à des variations dans la structure des documents sources.

La conversion vers le format XML s'est avérée plus robuste, capturant efficacement des sections telles que le préambule, le titre, les auteurs, le résumé, la bibliographie, l'introduction, et la conclusion. Cependant, des problèmes ont été relevés dans la gestion de la discussion, avec des erreurs signalées dans plusieurs cas. Cette observation souligne la nécessité d'améliorer la détection de la discussion dans les futurs développements du parseur.

En ce qui concerne les fichiers PDF spécifiques, des incohérences ont été notées, notamment des titres erronés, des auteurs manquants, des sections manquantes, et des conclusions incorrectes. Ces problèmes soulignent l'importance de continuer à affiner notre parseur pour s'adapter à une variété de structures documentaires.

Malgré ces défis, notre parseur offre une solution flexible et ciblée, permettant à l'utilisateur de sélectionner spécifiquement les fichiers à parser. Les fonctionnalités supplémentaires, telles que le choix du format de sortie (texte ou XML), renforcent la polyvalence de notre solution.