AVIGNON UNIVERSITÉ

Parseur d'Articles Scientifiques

Groupe 3

Mejai
Chementel
Ouahsouni
Jahid
Boudount

8 décembre 2023

Licence d'Informatique
Ingénierie Logicielle
UE Génie logiciel

UFR
SCIENCES
TECHNOLOGIES
SANTÉ

AVIGNON UNIVERSITÉ
CERI

CENTRE
D'**ENSEIGNEMENT**
ET DE **RECHERCHE**
EN **INFORMATIQUE**
ceri.univ-avignon.fr

Sommaire

1 Objectif

Le but de ce projet est de créer un programme capable de convertir des articles scientifiques au format PDF en fichiers texte et XML. Le programme identifiera les différentes sections d'un article, telles que le nom du fichier, le titre, l'abstract, l'introduction, le corps du texte, les discussions, la conclusion et la bibliographie. Le langage de programmation utilisé pour ce projet sera Python. Le résultat sera un répertoire nommé "Resultat" contenant tous les fichiers générés.

2 Méthode

2.1 Introduction

Le développement du parseur d'articles scientifiques a été guidé par une approche méthodique, intégrant Python et des scripts Shell pour traiter efficacement les fichiers PDF. Chaque fonction du programme a été conçue avec une stratégie spécifique pour optimiser l'extraction et la transformation des données.

2.2 Stratégies Fonctionnelles

- **Fonction findext :**
  Objectif : Identifier tous les fichiers PDF dans un dossier donné.
  Stratégie : Utilisation de glob pour filtrer les fichiers par extension, garantissant que seuls les fichiers PDF sont traités.
  Impact : Assure une initialisation efficace et ciblée du processus de parsing.
- **Fonction pdftotxt :**
  Objectif : Convertir les fichiers PDF en texte.
  Stratégie : Lancement d'un script Shell exécutant pdftotext, permettant de préserver la mise en forme et la structure du contenu.
  Impact : Préparation des fichiers pour une analyse textuelle approfondie.
- **Fonction findparagraph :**
  Objectif : Extraire des sections spécifiques comme l'Introduction, la Conclusion, et l'Abstract.
  Stratégie : Recherche de mots-clés et critères d'arrêt pour identifier les limites de chaque section.
  Impact : Extraction précise des sections clés pour une analyse détaillée.
- **Fonctions finddiscussion et findreference :**
  Objectif : Extraire les sections Discussion et Références.
  Stratégie : Identification de ces sections via des titres spécifiques et utilisation de balises de fin.
  Impact : Fournit une vue complète et structurée des articles analysés.
- **Fonction findcorps :**
  Objectif : Isoler le corps principal de l'article.
  Stratégie : Début de l'extraction après l'Introduction et terminaison avant la Discussion ou la Conclusion.
  Impact : Capture intégrale du contenu principal pour une analyse complète.
- **Fonctions parserfiletotxt et parserfiletoxml :**
  Objectif : Générer des sorties en formats texte et XML.
  Stratégie : Mise en forme des données extraites selon le format choisi.
  Impact : Fournit des résultats dans des formats adaptés à diverses applications.
- **Fonction main et selectpdffiles :**

Objectif : Offrir une interface utilisateur pour la sélection des fichiers et du format de sortie.
Stratégie : Menu interactif avec des options claires pour une expérience utilisateur optimisée.
Impact : Facilite la sélection des fichiers et améliore l'accessibilité du script.

2.3 Conclusion

Le parseur d'articles scientifiques, avec sa structure méthodique et ses stratégies fonctionnelles ciblées, représente une solution complète pour le traitement et l'analyse de documents PDF. Chaque fonction joue un rôle crucial dans la réalisation de l'objectif final : fournir une analyse structurée et détaillée des articles scientifiques. Cette approche méthodique assure une extraction précise des données, une conversion efficace et une présentation adaptée des résultats.

3 Resultat

3.1 Les fichiers utilisés

Numero	Fichiers
1	A Benders Decomposition Approach to Correlation Clustering
2	A memetic algorithm for community detectionin signed networks
3	An Improved Branch-and-Cut Code for the Maximum Balanced Subgraph of a Signed Graph
4	Cabrera RESUMES 2019
5	Conversational Networks for Automatic Online Moderation
6	Dynamical Models Explaining Socia Balance and Evolution of Cooperation
7	Exact Clustering via Integer Programming and Maximum Satisfiability
8	LDA resume
9	Partitioning large signed two-mode networks : Problems and prospects
10	Polibits 4202

3.2 Calcul de precision Strict

Fichiers	1	2	3	4	5	6	7	8	9	10
Frontières véritables	6	6	6	6	6	6	6	7	7	7
Frontières trouvées	6	6	6	6	6	6	6	7	6	7
Frontières correctes	6	5	6	5	5	4	5	7	4	6
Frontières incorrectes	0	1	0	0	1	2	0	0	2	1
Frontières non détectées	0	0	0	1	0	0	0	0	1	0
Precision stricte	1	0.83	1	0.83	0.83	0.66	0.83	1	0.66	0.85

Fichiers	1	2	3	4	5	6	7	8	9	10
Frontières véritables	6	6	6	6	6	6	6	7	7	7
Frontières trouvées	6	5	6	6	5	6	6	7	6	7
Frontières correctes	6	5	6	6	5	4	6	7	4	6
Frontières incorrectes	0	1	0	0	1	2	0	0	2	1
Frontières non détectées	0	0	0	0	0	0	0	0	1	0
Precision stricte	1	0.83	1	1	0.83	0.66	1	1	0.66	0.85

3.3 Calcul de precision Souple

3.4 calcul de la moyenne des precisions

3.4.1 Moyenne des precisions strictes

3.4.2 Moyenne des precisions strictes

4 Conclusion

En conclusion, l'utilisation de la méthodologie Scrum a été un succès pour le développement du parseur. Cette méthodologie nous a permis d'être flexibles et de nous adapter aux besoins du groupe. Elle a également renforcé l'esprit d'équipe, car nous avons travaillé ensemble sur des tâches communes.

L'utilisation de la méthodologie Scrum pour le développement du parseur a permis de nombreux avantages. Tout d'abord, elle a permis d'être flexibles et de nous adapter aux besoins du groupe. En effet, les sprints courts nous ont permis de réagir rapidement aux changements ou aux besoins émergents. Par exemple, si le groupe avait besoin de nouvelles fonctionnalités ou de modifications, nous pouvions les intégrer rapidement dans le projet.

Deuxièmement, l'utilisation de la méthodologie Scrum a renforcé l'esprit d'équipe. En effet, les sprints nous ont obligés à travailler ensemble pour atteindre des objectifs communs. Cela a créé un sentiment de cohésion et de collaboration au sein de l'équipe.