Parseur d'articles scientifiques en format xml

Donovan Casagrande [*1], **Benoît Maurice**[1], **Arthur Gros**[1], and **Benjamin Haurogné**[1]

[1]Université Bretagne Sud, Vannes, France

[*]Scrum Master

Table des matières

Résumé

1 Abstract

Le projet consiste à créer un parseur d'articles scientifiques pour simplifier la lecture des chercheurs de l'IRISA. Souvent, les articles scientifiques sont en format PDF qui est loin d'être portable et loin d'être facile à analyser par les systèmes de Traitement Automatique de Langues (TAL).

Notre Parser a donc pour bute de généré une version plus portable des articles en un autre format plus portable ce qui permet aux systèmes TAL de travailler correctement. D'où l'idée de faire une transformation de PDF.

2 Introduction

Les chercheurs de l'Institut de Recherche en Informatique et Systèmes Aléatoires sont en constantes explorations et ne cessent de lire et d'analyser le contenu de nombreux articles scientifiques et revues académiques. Cette tâche, fastidieuse et chronophage, nous a amené à créer une solution simple afin d'automatiser l'analyse des articles scientifiques dans le but d'optimiser le temps des chercheurs.

Le projet vise à répondre à ce besoin critique en développant un analyseur d'articles scientifiques, capable d'extraire rapidement et efficacement les informations essentielles des articles au format PDF et de les convertir en un format texte brut ou XML. Initié par l'Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), nous nous inscrivons dans une volonté d'améliorer l'accessibilité et l'analyse des articles scientifiques pour les chercheurs.

Notre équipe a rencontré de nombreuses difficultés dans la transformation des articles au format PDF en format texte et XML. Le format PDF est largement utilisé dans la publication d'articles scientifiques. Ce format présente plusieurs obstacles à franchir comme la gestion de textes sur deux colonnes, les formules mathématiques, les figures (tableaux et images), etc. D'autre part, l'analyse automatique de ces documents par les systèmes de Traitement Automatique de Langues (TAL) est rendue complexe en raison d'une absence d'une structure universelle comprenant des standards.

En fournissant un aperçu clair des articles, notre analyseur permettra aux chercheurs de gagner du temps dans leur exploration de la littérature scientifique, tout en offrant de nouvelles opportunités pour des analyses approfondies et des découvertes innovantes.

Dans ce contexte, notre équipe propose un analyseur d'articles scientifiques pour faciliter l'analyse et l'exploitation de ceux-ci. Notre article détaillera en détail son approche pour résoudre les différents défis rencontrés, les outils utilisés ainsi que les résultats obtenus lors de l'évaluation de notre analyseur sur un ensemble de documents scientifiques.

3 Méthodologie

L'organisation de ce projet ce fait sous forme de sprint en 4 temps.

1. Reception des demandes du clients
2. Réunion avec le client
3. Réalisation du sprint sur les demandes et le projet
4. Contrendu de sprint avec le client

mail regex chercher position des mots clées coups de hache exctraction

4 Résultats

Les résultats obtenus ont été évalués sur un ensemble diversifié d'articles scientifiques provenant de différentes disciplines. Nous discuterons plus en détail de ces résultats dans la section suivante. L'évaluation s'est penché sur les sections clés telles que le résumé, l'introduction, la discussion, la conclusion et la bibliographie. Le titre, les auteurs, leur courrier électronique et leur affiliation ont, eux aussi, été pris en compte. La moyenne pour chaque article est supérieure à ...%.
Nous avons obtenu 100% sur l'ensemble des articles en ce qui concerne l'extraction du ... et de ... . Nonobstant, les parties ayant les plus faibles résultats sont ... . Ces faibles résultats sont à relativiser, car la moyenne de l'ensemble des articles est de ...%.

5 Discussion

La discussion se concentrera sur l'efficacité et la pertinence de notre analyseur d'articles scientifiques.
Nos performances sont comparées aux résultats d'autres équipes travaillant sur ce projet. Toutes les équipes sont soumises à une même contrainte : Les résultats sont comparés en fonction d'un site regroupant les articles au format XML. Le site compare donc notre résultat par rapport à la référence.
Nous pouvons souligner que la limite de notre analyseur d'articles scientifiques est la langue dans laquelle le chercheur va écrire son article. En effet, nous sommes restés dans le contexte des langues latines ainsi que l'anglais. Bien que la langue principale dans le monde de la recherche soit l'anglais, nous devons considérer que certains chercheurs pourraient l'écrire en mandarin ou en arabe par exemple. Ces langues ayant de nombreuses règles spécifiques, notre analyseur pourrait avoir des difficultés à correctement extraire les informations clés.

6 Conclusion

En conclusion, notre projet a abouti au développement d'un analyseur d'articles scientifiques efficace et simple à utiliser, capable d'extraire automatiquement les informations

essentielles des articles au format PDF. Cette solution offre aux chercheurs un moyen rapide et précis d'accéder et d'analyser des articles scientifiques, tout en réduisant considérablement le temps nécessaire à cette tâche. Grâce à son potentiel d'automatisation et d'optimisation des processus d'analyse, notre analyseur couvre l'ensemble des demandes du laboratoire et ouvre la voie à une meilleure exploration des connaissances du monde scientifique.

7 References

7.1 Sous-Partie 1

Lorem Ipsum.