Projet de développement

Sprint 4

Equipe : Romain Le Talour, Corentin Leconte, Thibaut Quilleré.

Pour ce projet de développement, la méthodologie de travail est basée sur l'agile avec une approche SCRUM.

Maitre SCRUM : Romain Le Talour

Lien du GitHub : https://github.com/genielog-ue/pdftotxt.git

Dans ce cinquième et dernier sprint nous nous sommes penchés sur l'évaluation de l'efficacité de notre parseur. En effet, nous avons dû calculer la précision avec un corpus de test contenant 10 PDF de type scientifique. Pour ce faire nous avons employé la formule :

Précision = SectionsCorrectesTrouvéesParLeSystème / SectionsVéritables

Les sections véritables proviennent de la conversion manuelle que nous avons dû effectuer. Nous avons utilisé le modèle suivant pour établir un corpus de référence :

<article>

<preamble> Le nom du fichier d'origine </preamble>

<titre> Le titre du papier </titre>

<auteurs>

<auteur> L'auteur A et son adresse courriel</auteur>

<affiliation> L'affiliation de l'auteur A et son adresse</affiliation>

<auteur> L'auteur B et son adresse courriel</auteur>

<affiliation> L'affiliation de l'auteur B et son adresse</affiliation>

...

<auteurs/>

<abstract> Le résumé de l'article</abstract>

<introduction> La introduction</introduction>

<corps> Le développement du papier</corps>

<conclusion> La conclusion du papier</conclusion>

<discussion> La discussion du papier</discussion>

<biblio> Les références bibliographiques du papier</biblio>

</article>

Nous avons extrait des PDF les sections correspondantes et ensuite avons comparé avec ce que nous donne notre parseur.

Voici donc les résultats obtenus :

b0e5 : Titre : 100% 1 Auteurs : 100%, 98, 98, 3 Emails : 100% 1 Abstract : 100% 1 Introduction :93% 1 Corps : 24% 0 Conclusion : 27% 0 Biblio : 94.7% 1 8/10 0.8	**BLESS :** Titre : 100% 1 Auteurs : 100% 100% 2 Emails : 100% 100% 2 Abstract : 99% 1 Introduction : 92% 1 Corps : 93.9% 1 Conclusion : 99% 1 Biblio : 19% 0 9/10 0.9	**C14-1212 :** Titre : 100% 1 Auteurs : 100% 100% 100% 100% 100% 5 Emails : 100% 1 Abstract : 93.82% 1 Introduction : 58% 0 Corps : 92.83% 1 Conclusion : 99.35% 1 Biblio : 99.88 % 1 11/12 0.91
Guy : Titre : 97.87% 1 Auteurs : 40% 40% 0% 0 Emails : 100% 100% 100% 3 Abstract : 76%0 Introduction : 98% 1 Corps : 88% 0 Conclusion : 99% 1 Biblio : 90% 1 7/12 0.58	**infoEmbeddings :** Titre : 98% 1 Auteurs : 0% 0 Emails : none Abstract : 72%0 Introduction : 74% 0 Corps : 91% 1 Conclusion : 98% 1 Biblio :97.03% 1 4/7 0.57	**IPM1481 :** Titre : 0% 0 Auteurs : 100% 100% 100% 0% 100% 4 Emails : 100 100 40 100 3 Abstract : 2% 0 Introduction :48% 0 Corps : 0% 0 Conclusion : 19% 0 Biblio : 93% 1 8/13 0.6
L18-1504 : Titre : 97% 1 Auteurs : 100% 0% 0% 0% 1 Emails : 100% 0% 0% 0% 1 Abstract : 99.75% 1 Introduction : 97% 1 Corps : 0% 0 Conclusion : 95% 1 Biblio : 22% 0 6/14 0.4	**acl2012 :** Titre : 84.8% 0 Emails : 100% 1 Auteurs : 63.15%+54.54%+38%+53.84+55.55% = 53.016% Abstract : 98.88% 1 Introduction : 91.051567239636% 1 Corps : 95.61% 1 Conclusion : 97% 1 Discussion : 92.36% 1 Biblio : 64% 0 6/13 0.46	**On The Morality of Artificial Intelligence** Titre : 66% 0 Auteurs : 0% 0% 0 Emails : None Abstract : 40%0 Introduction : None Corps : 87% 0 Conclusion : 71% 0 Biblio : 96% 1 1/6 0.16
SurveyTermExtraction Titre : 100% 1 Auteurs : 0% 0% 0% 0 Emails : 100% 100% 100% 3 Abstract : 100% 1 Introduction : 15% 0 Corps : 34% 0 Conclusion : 4% 0 Biblio : 97% 1 6/12 0.5		

Moyenne générale :

(0.46+0.8+0.9+0.91+0.58+0.57+0.6+0.4+0.16+0.5) / 10 = 0.548 = 54.8%

Pour finir nos avons édité un compte rendu en LaTeX structuré comme les PDF des corpus.

Nous pensons avoir obtenu des résultats tout à fait satisfaisant et le cahier des charges est rempli.