Maître Scrum : Théau HUTEAU Propriétaire du produit : laboratoire IRISA et M. Kessler
Equipe de développement : Antoine JAMELOT, Sofiane BEN MASSAOUD, Baptiste COLAS

Rapport

L'objectif de ce projet est de produire pour l'IRISA un parseur d'article scientifique permettant de leur résumer les informations principales d'articles scientifiques

Objectif de ce sprint :

Pour faire cet outil, il faut d'abord permettre la conversion des articles au format pdf au format texte sur une machine GNU Linux. Deux outils nous étaient proposés : pdf2text et pdftotext.
L'objectif de ce sprint était de tester ses deux outils avec les différentes options disponibles et lequel était le mieux adapté aux différentes contraintes du projet.

Notre choix :

Nous avons choisi pdftotext pour différentes raisons que nous allons vous détailler dans ce rapport en vous présentant les avantages et inconvénients des deux outils.

Particularités des articles :

Les articles du corpus qui nous a été fourni comporte différentes parties qu'il nous faudra extraire et résumer par la suite. Les titres et auteurs ne sont pas toujours disposés de la même façon ce qui peut rendre le travail délicat. Donc nous ne pouvons pas nous fier à l'ordre. Ces articles pour la plupart sont présentés avec deux colonnes par pages, mais pas toujours. Ils comportent aussi des tableaux, des équations, ainsi qu'un en-tête ou un pied de page (parfois les deux). Ceci peut entraîner des difficultés, c'est pour cela qu'il est important de prendre cela en compte dans le choix de l'outil de conversion du format 'PDF' vers 'TXT'.

Exemples :

	ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline	0.26232	0.04543	0.08247
3^{rd} system	0.35715	0.09622	0.13245
2^{nd} system	0.36965	0.09851	0.13509
$cosine + \text{JW}_e$	0.35905	0.10161	0.13701
NR	0.36207	0.10042	0.13781
SMMR	**0.36323**	**0.10223**	**0.13886**
1^{st} system	0.37032	0.11189	0.14306
Worst human	0.40497	0.10511	0.14779

Table 1: ROUGE average recall scores computed on the DUC 2007 update corpus.

$$P(s \in \mathcal{S} \mid F_1, F_2, ..F_k) = \frac{\prod_{i=1}^{k} P(F_i \mid s \in \mathcal{S}) \cdot P(s \in \mathcal{S})}{\prod_{i=1}^{k} P(F_i)}$$

(cid:81)k
i=1 P (Fi | s ∈ S) · P (s ∈ S)

(cid:81)k

i=1 P (Fi)

P (s ∈ S | F1, F2, ..Fk) =

(1)

Article Boudin_Torres-2006.pdf
Article Das_Martin.pdf

L'ensemble de ses exemples montrent les particularités des articles scientifiques. Étant donné que l'outil demander doit seulement résumer le contenu des différentes parties des articles, nous pensons plutôt valoriser le texte des articles aux tableaux et équations mathématiques ; qui problèmes aux deux outils.

Pdf2text :

Nous avons testé cet outil avec différentes options et sur différents articles.

Maître Scrum : Théau HUTEAU Propriétaire du produit : laboratoire IRISA et M. Kessler
Equipe de développement : Antoine JAMELOT, Sofiane BEN MASSAOUD, Baptiste COLAS

Pour le rapport détaillé voir sur https://github.com/TheauH/Parser-IRISA/tree/main/test, partie « conversionpdf » le document « rapport_pdf2text ». Petit condensé du rapport : L'outil sans option donne quelque chose d'illisible. L'option -A est préconisé dans son utilisation mais elle ne résout pas tous les problèmes.

Avantages :

- Le texte est lu qu'il soit scindé ou en bloc
- L'élément : '' informe du changement de page
- L'option -A permet un meilleur rendu du texte en le forçant

Inconvénients :

- Titre souvent sur 2 lignes distinctes
- Tables et formules retranscrite plus ou moins bien
- Accent non retranscrit sur la lettre mais à coté
- Police en gras ou spéciale rend un texte avec des de saut de lignes pour une même phrase ce qui la phrase difficile à lire et à « parser »
- Ordre du texte du fait de certains pieds de pages
- Option « -A » non uniforme sur tous les articles du corpus

Pdftotext :

L'outil pdftotext est similaire au précédent mais il comporte d'autres avantages et inconvénients dû à ses options qui sont plus riches que pdf2text. Pour l'intégralité du rapport voir sur https://github.com/TheauH/Parser-IRISA/tree/main/test, partie « conversionpdf » document « rapport_pdftotext ».Nous allons vous donner notre conclusion sur l'utilisation de cet outil sur les articles du corpus.

Avantages :

- Avec l'option « -layout » le texte garde son format initial a deux colonnes
- Différentes parties des articles bien délimiter

Inconvénients :

- Certains articles sont mal converti en utilisant « -layout »
- Champs d'en-tête problématique

L'option semblant la plus intéressante semble être « -layout ». Mais, il faut tout de même faire attention à ce que le texte comporte bien 2 colonnes. Sur l'article Das_Martins.pdf cela ne semble pas être le cas. C'est à notre avis pour cette raison que cela donne un résultat aussi peu fidèle à l'article initiale (cependant aucun des 2 outils ne semblent être adapter à l'article).

Conclusion :

Les deux outils ont leur particularité. Pdf2text semble rendre un résultat plus aéré pour le texte et ses différentes parties mais, son traitement des pieds de page peut parfois poser problèmes alors que pdftotext avec l'option « -layout » semble donner un résultat cohérent lorsque le texte a 2 colonnes, même sur les pieds de pages et rend un article dans l'ordre avec un titre des différentes sections fidèle au document d'origine. C'est pour cela que nous préconisons une utilisation de pdftotext plutôt que pdf2text.