# TP Projet de développement Scrum 2020-2021

## Carnet produit sprint 1


## Membres du groupe :

-CADOU Valentin
-DIALLO Boubacar
-GARECHE Amirouche
-NGOMA SITA Dorian


## Sujet :

Les chercheurs de l'IRISA pour avancer dans leurs recherches, doivent lire les articles scientifiques publiés dans le monde. Malheureusement ils n'ont pas le temps de tout lire et voudraient avoir un système qui présente un aperçu de l'article permettant ainsi de gagner du temps. Ils préfèrent donc un format texte des articles à un format PDF pour l'utilisation d'un système de Traitement Automatique de Langues. D'où l'idée de faire une transformation de PDF à texte.

Pour cela, il existe deux outils libres qui sont pdftotext et pdf2txt avec des options de lancement différentes que l'on aimerait justement tester.

Lequel des deux logiciels est le plus adapté? Quelles options? Nous l'ignorons pour l'instant.

Testons ces deux outils pour convertir les PDF du corpus d'apprentissage en plein texte. Ensuite, nous évaluerons qualitativement les sorties afin de décider lequel utiliser: les frontières des phrases sont-elles bien respectées? Les mots sont-ils mal découpés? Les lignes (ou phrases) sont-elles entremêlées par rapport au PDF? etc.

L'étude suivante s'appuie sur une partie des documents contenus dans le corpus et se borne à certaines options des deux convertisseurs(à savoir celles qui nous seront utiles pour la réalisation de ce projet).

# pdftotext :

## 1. pdftotext sans options sur le fichier Boudin-Torres-2006.pdf:

- On remarque que la séparation des pages en deux colonnes n'a pas été prise en compte et celles-ci sont entremêlées.
- Non prises en charge également des caractères spéciaux et mathématiques.
- Le tableau a été mal converti.

## 2. pdftotext -layout sur le fichier Boudin-Torres-2006.pdf:

- La séparation des pages en deux colonnes a bien été faite hormis un mauvais espacement de certaines lignes et mots.
- Non prise en charge de certains caractères et symboles mais les formats des équations mathématiques sont à peu près corrects.

$$\mathrm{JW}_e(s,Q) = \frac{1}{|Q|} \cdot \sum_{q \in Q} \max_{m \in S'} \mathrm{JW}(q,m) \quad (1)$$

fichier PDF

```
J We (s, Q) =  1 X  ·   max J W(q, m)        (1)
              |Q|     m∈S 0
                   q∈Q
```

fichier Texte

- Le format tableau a bien été fait sans conserver la séparation des lignes et colonnes et la mise en évidence de la 7éme ligne.

| | 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 |
| $cosine + \mathrm{JW}_e$ | 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 |

fichier PDF

| | ROUGE -1 | ROUGE -2 | ROUGE - SU 4 |
|---|---|---|---|
| Baseline | 0.26232 | 0.04543 | 0.08247 |
| 3rd system | 0.35715 | 0.09622 | 0.13245 |
| 2nd system | 0.36965 | 0.09851 | 0.13509 |
| cosine + J We | 0.35905 | 0.10161 | 0.13701 |
| NR | 0.36207 | 0.10042 | 0.13781 |
| S MMR | 0.36323 | 0.10223 | 0.13886 |
| 1st system | 0.37032 | 0.11189 | 0.14306 |
| Worst human | 0.40497 | 0.10511 | 0.14779 |

fichier Texte

## 3. pdftotext -layout sur le fichier Kessler94715.pdf:

- Les figures sont mâles converties et les images ne sont pas du tout affichées.


Fig. 1. figure describing the context of the project

fichier PDF

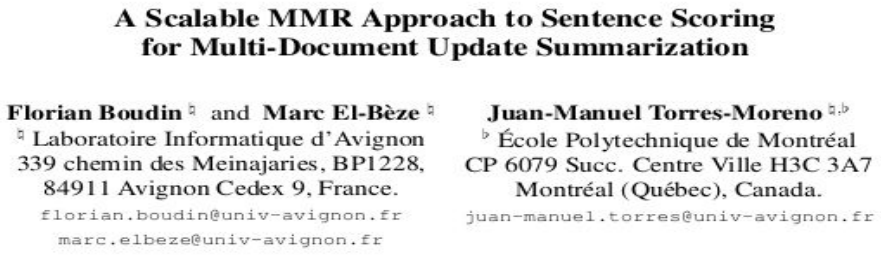
Fig. 1. figure describing the context of the project

fichier Texte

# pdf2txt :

## 1. pdf2txt sans options sur le fichier Boudin-Torres-2006.pdf:

- Le pied de page du fichier pdf se retrouve au début de la page du fichier Texte.
- Les caractères spéciaux et symboles mathématiques ne sont pas reconnus.
- Les phrases, lignes et paragraphes sont mal découpés ainsi que le non espacement des mots.



**A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization**

**Florian Boudin** [a] and **Marc El-Bèze** [a]
[a] Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, BP1228,
84911 Avignon Cedex 9, France.
florian.boudin@univ-avignon.fr
marc.elbeze@univ-avignon.fr

**Juan-Manuel Torres-Moreno** [a,b]
[b] École Polytechnique de Montréal
CP 6079 Succ. Centre Ville H3C 3A7
Montréal (Québec), Canada.
juan-manuel.torres@univ-avignon.fr

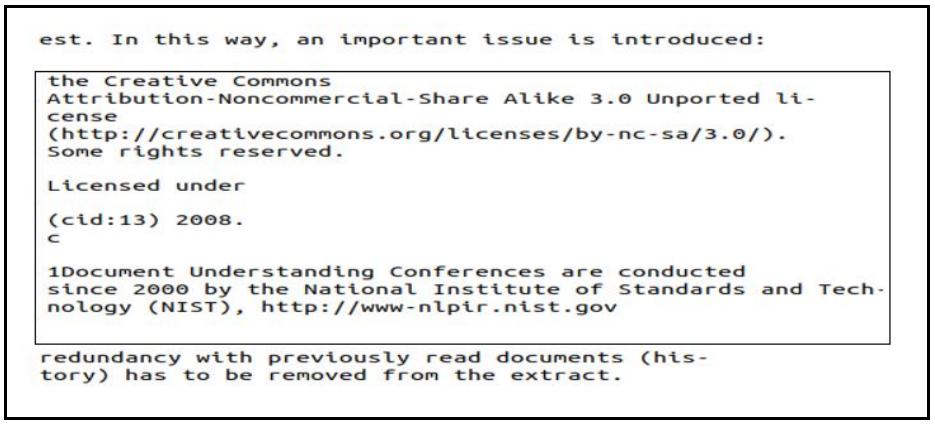
fichier PDF

```
AScalableMMRApproachtoSentenceScoringforMulti-DocumentUpdateSummarizationFlorianBoudin\andMarcEl-B`eze\-
\LaboratoireInformatiqued'Avignon339chemindesMeinajaries,BP1228,84911AvignonCedex9,France.florian.boudin@univ-avignon.frmarc.elbeze@univ-avignon.frJuan-ManuelTorres-Moreno\,-
[[´EcolePolytechniquedeMontr´ealCP6079Succ.CentreVilleH3C3A7Montr´eal(Qu´ebec),Canada.juan-manuel.torres@univ-avignon.frAbstractWepresentSMMR,ascalablesentencescoringmethodforquery-orientedup-
```

fichier Texte

## 2. pdf2txt -A sur le fichier Boudin-Torres-2006.pdf:

- Tout comme la conversion pdf2txt sans options, le pied de page se retrouve au début.
- Les mots, lignes et paragraphes sont bien espacés. Par contre les pages ne sont pas affichées sur deux colonnes et mauvais passage à la deuxième colonne(les définitions devant être en pied de page).



```
est. In this way, an important issue is introduced:

the Creative Commons
Attribution-Noncommercial-Share Alike 3.0 Unported li-
cense
(http://creativecommons.org/licenses/by-nc-sa/3.0/).
Some rights reserved.

Licensed under

(cid:13) 2008.
c

1Document Understanding Conferences are conducted
since 2000 by the National Institute of Standards and Tech-
nology (NIST), http://www-nlpir.nist.gov

redundancy with previously read documents (his-
tory) has to be removed from the extract.
```

# Conclusion :

Après analyses et tests des deux outils sur les documents précédents et plus largement l'ensemble du corpus, nous avons choisi le convertisseur **pdftotext** avec l'option **-layout** (ce convertisseur offre également d'autres options intéressantes que nous n'avons cependant pas retenues dans le cadre de ce projet) étant le plus adapté des deux même s'il est loin d'être parfait comme constaté au cours de cette étude.