

Projet de Developpement

Cadou Valentin, Diallo Boubacar, Gareche Amirouche, and Ngoma
Sita Dorian

cadou.e1800104@etud.univ-ubs.fr,
diallo.e1704214@etud.univ-ubs.fr,
gareche.e1704251@etud.univ-ubs.fr,
ngoma-sita.e182705@eutd.univ-ubs.fr.

Univerité Bretagne Sud, Vannes, France

1 Abstract

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 à avoir un système qui les présente un aperçu de l'article. Notre projet consiste donc à faire un programme qu'on a choisit de faire en python qui fait un resumé des sections de l'article PDF dans une version texte bien découpée(le titre, les auteurs, l'introduction, le développement,la conclusion etc.)

2 Methode

le projet a ete developpé suivant une methode agile composés de cinq sprints . Pour commencer Nous avons choisi l'outil pdftotext pour faire la conversion des aricles scientifiques en format texte et python comme langage de programmation pour faire notre programme parce qu'on le maitrise plutot bien et aussi pour son efficacité en terme de temps d'execution.Nous avons placés progressivement les versions intermédiaires et le systeme final sur Github. Le système prend comme entrée un dossier contenant les fichiers PDF et créer un sous-dossier pour déposer les sorties en plain texte avec le même noms que les PDF.nous avons ensuite rajouter une nouvelle sortie .xml. Ainsi, au moment de lancer le parseur, nous avons ajouter un argument (-t — -x) pour bien choisir le type de sortie : soit des fichiers dans une version texte (-t) , soit au format XML (-x). En sortie nous avons un repertoire composé de fichier resumé des differents pdf en .txt ou .xml ayant les memes noms que les fichiers pdf.Ces fichiers doivent etre composés du Titre,des auteurs et de leur affiliation, de l'abstract, de l'introduction, du corps

du document, de la conclusion, de la partie discussion et des references. Pour terminer nous avons utilisés un nouveau corpus composés de 10 fichiers pdf pour nous permettre de mesurer la precision de notre systeme.

3 Resultats

-fichier du corpus en format texte

```
pdftotext -layout CorpusTEST//IPM1481.pdf xml CorpusTEST//IPM1481.xml
ubs@ubs-Latitude-E5430-non-vPro:~/Téléchargements/Parser1.0-main$ python3 parser10.py -t CorpusTEST/
Liste des fichiers contenus dans le répertoire :
L18-1504.pdf
IPM1481.pdf
C14-1212.pdf
txtCorpusTest.zip
b0e5c43edf116ce2909ae009cc27a1546f09.pdf
On the Morality of Artificial Intelligence.pdf
BLESS.pdf
Guy.pdf
infoEmbeddings.pdf
surveyTermExtraction.pdf
acl2012.pdf
Veuillez indiquer quel(s) fichier(s) analyser en écrivant le(s) nom(s), utilisez une virgule comme séparateur entre chaque nom (indiquez all pour analyser tous les fichiers) :
all
pdftotext -layout CorpusTEST//L18-1504.pdf txt CorpusTEST//L18-1504.txt
pdftotext -layout CorpusTEST//IPM1481.pdf txt CorpusTEST//IPM1481.txt
pdftotext -layout CorpusTEST//C14-1212.pdf txt CorpusTEST//C14-1212.txt
pdftotext -layout CorpusTEST//b0e5c43edf116ce2909ae009cc27a1546f09.pdf txt CorpusTEST//b0e5c43edf116ce2909ae009cc27a1546f09.txt
pdftotext -layout CorpusTEST//On the Morality of Artificial Intelligence.pdf txt CorpusTEST//On the Morality of Artificial Intelligence.txt
pdftotext -layout CorpusTEST//BLESS.pdf txt CorpusTEST//BLESS.txt
pdftotext -layout CorpusTEST//Guy.pdf txt CorpusTEST//Guy.txt
pdftotext -layout CorpusTEST//infoEmbeddings.pdf txt CorpusTEST//infoEmbeddings.txt
pdftotext -layout CorpusTEST//surveyTermExtraction.pdf txt CorpusTEST//surveyTermExtraction.txt
pdftotext -layout CorpusTEST//acl2012.pdf txt CorpusTEST//acl2012.txt
ubs@ubs-Latitude-E5430-non-vPro:~/Téléchargements/Parser1.0-main$
```

-fichier du corpus en format xml

```
ubs@ubs-Latitude-E5430-non-vPro:~/Téléchargements/Parser1.0-main$ python3 parser10.py -x CorpusTEST/
Liste des fichiers contenus dans le répertoire :
L18-1504.pdf
IPM1481.pdf
C14-1212.pdf
txtCorpusTest.zip
b0e5c43edf116ce2909ae009cc27a1546f09.pdf
On the Morality of Artificial Intelligence.pdf
BLESS.pdf
Guy.pdf
infoEmbeddings.pdf
surveyTermExtraction.pdf
acl2012.pdf
Veuillez indiquer quel(s) fichier(s) analyser en écrivant le(s) nom(s), utilisez une virgule comme séparateur entre chaque nom (indiquez all pour analyser tous les fichiers) :
all
pdftotext -layout CorpusTEST//L18-1504.pdf xml CorpusTEST//L18-1504.xml
pdftotext -layout CorpusTEST//IPM1481.pdf xml CorpusTEST//IPM1481.xml
pdftotext -layout CorpusTEST//C14-1212.pdf xml CorpusTEST//C14-1212.xml
pdftotext -layout CorpusTEST//b0e5c43edf116ce2909ae009cc27a1546f09.pdf xml CorpusTEST//b0e5c43edf116ce2909ae009cc27a1546f09.xml
pdftotext -layout CorpusTEST//On the Morality of Artificial Intelligence.pdf xml CorpusTEST//On the Morality of Artificial Intelligence.xml
pdftotext -layout CorpusTEST//BLESS.pdf xml CorpusTEST//BLESS.xml
pdftotext -layout CorpusTEST//Guy.pdf xml CorpusTEST//Guy.xml
pdftotext -layout CorpusTEST//infoEmbeddings.pdf xml CorpusTEST//infoEmbeddings.xml
pdftotext -layout CorpusTEST//surveyTermExtraction.pdf xml CorpusTEST//surveyTermExtraction.xml
pdftotext -layout CorpusTEST//acl2012.pdf xml CorpusTEST//acl2012.xml
ubs@ubs-Latitude-E5430-non-vPro:~/Téléchargements/Parser1.0-main$
```

-exemples de calcul de precision

```
ubs@ubs-Latitude-E5430-non-vPro:~/Téléchargements/Parser1.0-main$ python3 parser_precision.py xml_CorpusTEST/L18-1504.xml Resume_corpus_test/L18-1504.xml
L18-1504.xml precision :
titre : 100.0
auteurs : 98.94179894179894
abstract : 100.11547344110853
conclusion : 98.94291754756871

ubs@ubs-Latitude-E5430-non-vPro:~/Téléchargements/Parser1.0-main$ python3 parser_precision.py xml_CorpusTEST/surveyTermExtraction.xml Resume_corpus_test/surveyTeamExtraction.xml
surveyTermExtraction.xml precision :
titre : 100.0
auteurs : 97.8102189781022
abstract : 99.84984984984985
conclusion : 101.669449081803
```

4 Conclusion

En somme quatorze semaine et cinq sprints nous ont été nécessaire pour la réalisation de ce projet. Nous avons utilisé principalement deux ordinateurs équipés du système linux.

Nous avons également rencontré différents problèmes, tout d'abord lors du sprint 2 au niveau de la structure des documents qui était différente d'un document à l'autre on a donc dû prendre en compte toutes ses différences. Le deuxième problème rencontré a été lors du sprint 5 puisqu'il nous fallait faire évoluer notre parseur afin de convertir un nouveau corpus sans modifier les résultats obtenus avec le premier corpus.

Nous avons fait face à des bugs tout au long de la réalisation mais que nous avons réussi à corriger lors de nos réunions.

Finalement nous avons répondu aux attentes des chercheurs de L'IRIZA puisque notre parseur est écrit en langage python et donc particulièrement adapté pour ce projet qui repose sur le traitement des langues, de plus le temps d'exécution est relativement rapide bien que le corpus comporte un nombre important de documents. Deux formats de sortie sont possibles (format texte et xml) répondant à une fonctionnalité très attendue, pour finir nous obtenons un bon niveau de précision, plus de 90 pourcent pour chaque document, ce qui montre le bon découpage des fonctions.