UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

C E N T R E
D'ENSEIGNEMENT
ET DE RECHERCHE
EN INFORMATIQUE

Licence Informatique
ILSEN
UE Génie Logiciel

>>> Projet SCRUM: Séance 3

Groupe 2

Enzo GUENY
Adam SERGHINI
Jarod DURET

Encadrement
Juan Manuel TORRES MORENO

CERI - LIA
339 chemin des Meinajariès
BP 1228
84911 AVIGNON Cedex 9
France

Tél. +33 (0)4 90 84 35 00
Fax +33 (0)4 90 84 35 01

http://ceri.univ-avignon.fr

Sommaire

1 Introduction

Ce document est un rapport sur l'avancement du projet et et les étapes réalisées lors de la séance du 12 déc. sur le projet **UAPV TP de Génie logiciel - Scrum**.

SCRUM Master (DURET Jarod).
Dépot GitHub : -> https://github.com/Team-Rocket-CERI/SCRUM_Project

2 Rétrospective du Sprint

Suite au sprint précédent, nous avons parser les fichiers pdf afin de pouvoir récuperer les éléments suivant :

- Le nom du fichier d'origine
- Le titre de l'article
- Le résumé de l'article (Abstract)

Étant satisfait du résultat, obtenue grâce au python et au rust, nous avons décidé de continué sur cette lancée pour les prochaines étapes.

3 La mêlée quotidienne

Lors de la mêlée du jour, nous avons décidé de continué sur notre voie et de rajouté les éléments nécessaire à la nouvelle étape ; on souhaite obtenir en sortie un résultat contenant toutes les informations extraites sous formes de balises(<titre>, <auteur>, <abstract>, <biblio> ...). Afin d'exploiter les informations extraites, nous utiliserons le langage python avec les librairies etree et LXML. Ce changement de cap brutal nous a obligés à réflechir et à prendre des décisions lors du brainstorming. Il a fallu alors revenir sur des décisions prises auparavant tels que les programmes et langages choisis pour réaliser le parseur.

4 Le sprint

Reprogrammation du parseur avec les informations suivantes à extraire (sous cette forme) :

- <**article**>
  - <**preamble**> Le nom du fichier d'origine </**preamble**>
  - <**titre**> Le titre du papier </**titre**>
  - <**auteur**> La section auteurs et leur adresse </**auteur**>
  - <**abstract**> Le résumé de l'article </**abstract**>
  - <**biblio**> Les références bibliographiques du papier </**biblio**>
- </**article**>

La majorité du code sera désormais rédigée en Python, nous conservons cependant le programme pdf_to_text utile dans le main codé en rust. Ce script, préalablement réalisé, permet de parser une première fois le document dans un .txt provisoire. Le script python "parser.py" sera ensuite chargé d'extraire les informations utiles et les organiser dans un nouveau fichier XML.

5 Rétrospective du sprint

Nous avons rencontrés des problèmes dû a l'encodage des caractères car l'xml n'accepte que l'unicode ou l'ASCII.

6 Conclusion

Avec tout les outils mentionnés plus tôt, nous avons maintenant parser les documents pdf en respectant les nouvelles conditions.
Résultat final :


```
<xml><article><preambule>Torres-Moreno_2012_Artex is another text summarizer.pdf</preambule><titre>Artex is another text summarizer</titre><auteur>Torres-Moreno</auteur><abstract>Abstract
This paper describes Artex, another algorithm for Automatic Text Summarization.
In order to rank sentences, a simple inner product is calculated between each sentence, a
document vector (text topic) and a lexical vector (vocabulary used by a sentence). Summaries are then generated by assembling the highest ranked sentences. No ruled-based
linguistic post-processing is necessary in order to obtain summaries. Tests over several
datasets (coming from Document Understanding Conferences (DUC), Text Analysis Conference (TAC), evaluation campaigns, etc.) in French, English and Spanish have shown
that Artex summarizer achieves interesting results.

[9] H.P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and
Development, 2(2):159&#8211;165, 1958.
[10] I. Mani and M. Mayburi. Advances in Automatic Text Summarization. MIT Press, Cambridge,
1999.
[11] Eric SanJuan, Patrice Bellot, V&#233;ronique Moriceau, and Xavier Tannier. Overview of the INEX
2010 Question Answering Track (QA@INEX). In Shlomo Geva, Jaap Kamps, Ralf Schenkel, and
Andrew Trotman, editors, Comparative Evaluation of Focused Retrieval, volume 6932 of Lecture
Notes in Computer Science, pages 269&#8211;281. Springer Berlin / Heidelberg, 2011.

</abstract><biblio>References
[1] Iria da Cunha, Silvia Fern&#225;ndez, Patricia Vel&#225;zquez-Morales, Jorge Vivaldi, Eric SanJuan, and
Juan Manuel Torres-Moreno. A new hybrid summarizer based on vector space model, statistical
physics and linguistics. In Proceedings of the 6th Mexican International Conference on Advances in
Artificial Intelligence (MICAI&#8217;07), pages 872&#8211;882, Aguascalientes, Mexico, 2007. Springer-Verlag.
[2] Harold Daum&#233; III. Practical structured learning techniques for natural language processing. PhD
thesis, Los Angeles, CA, 2006.
[3] H. P. Edmundson. New Methods in Automatic Extraction. Journal of the Association for Computing Machinery, 16(2):264&#8211;285, 1969.
[4] B. Favre, F. B&#233;chet, P. Bellot, F. Boudin, M. El-B&#232;ze, L. Gillard, G. Lapalme, and J-M. TorresMoreno. The LIA-Thales summarization system at DUC-2006. In Proceedings of the Docume
[5] Silvia Fern&#225;ndez, Eric SanJuan, and Juan-Manuel Torres-Moreno. Textual Energy of Associative Memories: performants applications of Enertex algorithm in text summarization and topic
segmentation. In Proceedings of the Mexican International Conference on Artificial Intelligence
(MICAI&#8217;07), pages 861&#8211;871, Aguascalientes, Mexico, 2007. Springer-Verlag.
[6] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In Proceedings of the
18th Conference ACM Special Interest Group on Information Retrieval (SIGIR&#8217;95), pages 68&#8211;73,
Seattle, WA, United States, 1995. ACM Press, New York.
[7] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine
Moens and Stan Szpakowicz, editors, Proceedings of the Workshop Text Summarization Branches
Out (ACL&#8217;04), pages 74&#8211;81, Barcelone, Spain, july 2004. ACL.
[8] Annie Louis and Ani Nenkova. Automatic Summary Evaluation without Human Models. In First
Text Analysis Conference (TAC&#8217;08), Gaithersburg, MD, United States, 17-19 November 2008.
[9] H.P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and
Development, 2(2):159&#8211;165, 1958.
[10] I. Mani and M. Mayburi. Advances in Automatic Text Summarization. MIT Press, Cambridge,
1999.
[11] Eric SanJuan, Patrice Bellot, V&#233;ronique Moriceau, and Xavier Tannier. Overview of the INEX
2010 Question Answering Track (QA@INEX). In Shlomo Geva, Jaap Kamps, Ralf Schenkel, and
Andrew Trotman, editors, Comparative Evaluation of Focused Retrieval, volume 6932 of Lecture
Notes in Computer Science, pages 269&#8211;281. Springer Berlin / Heidelberg, 2011.

</biblio></article></xml>
```