Recovering Damaged Documents to Improve Information Retrieval Processes

Angel Luis Garrido[1*], Alvaro Peiró[2]

[1] SID Research Group, University of Zaragoza, Zaragoza, Spain
[2] InSynergy Consulting S.A., Madrid, Spain

* garrido@unizar.es

Abstract

Computer forensics is a specialty within forensic science disciplines. The goal of computer forensics is the search for evidence in digital data, involving the preservation, identification, storage media extraction, documentation and interpretation [1]. Although computer forensics is frequently related to the investigation of computer crimes, it can also be used in civil procedures, being the information recovery one of the main scenarios where this discipline works. A use case is the information retrieval from damaged documents, where words have undergone alterations, either accidentally or intentionally. Correcting texts, also known as "text curation", is a well-know task in the field Natural Language Processing and it is applied in many applications [2]. When facing a high number of lengthy documents, some type of automation is advisable to complete the task in a reasonable time.

In this paper, we present a new tool able to retrieve information from large volumes of documents whose words have been damaged. We have designed a new approach to recover the original words. This recovery process is composed of two main steps: 1) a *Text Cleaning Filter*, able to remove non relevant information, as heading, page numbers, stamps, etc., and 2) a *Text Correction Unit*, which combines a general purpose spell-check with a domain N-Gram based spell checker built specifically for the domain of the documents we are dealing with. The benefits of using this combined approach are two-fold: on the one hand, the general spell checker allows us to leverage all the general purpose techniques that are usually used to perform the corrections; on the other hand, the use of an N-gram-based model allows us to adapt them to the particular domain we are tackling exploiting text regularities detected in successfully processed domain documents. The result of the recovery allows us to improve automatic information retrieval tasks of from the texts. We have implemented our proposal on AIS [3], an information system within the domain of legal texts, where the information retrieval process is carried out guided by a specific domain ontology for the typology of the document. We have tested it using a real data set, in collaboration with the company InSynergy Consulting, with very promising results

References

[1] W. G. Kruse II, J. G. Heiser, *Computer forensics: incident response essentials*. Pearson Education (2001).

[2] C.D. Manning, H. Schütze, *Foundations of statistical natural language processing*. MIT Press (1999).

[3] M.G. Buey, A.L., Garrido, C. Bobed, S. Ilarri, *The AIS Project: Boosting Information Extraction from Legal Documents by using Ontologies*. Proceedings of the 8th International Conference on Agents and Artificial Intelligence (2016) 438-445.

Acknowledgements

This research work has been supported by projects TIN2013-46238-C4-4-R, TIN2016-78011-C4-3-R (AEI/FEDER, UE), and DGA/FEDER. Thanks to Dr. Carlos Bobed.