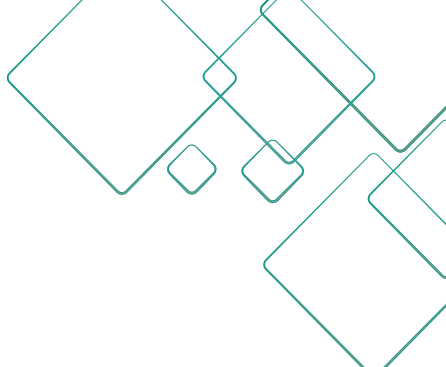




Research Relevance and Potential Impact

Distilling operational information from the web is a wealth and a foundation for a knowledge economy. Resources dealing with Arabic text on the web (news, information, tweets, blogs, etc.) are scarce, if non-existent. Information from such a rich and abundant data is not readily available for policy makers, social scientists, linguists, and the general public. This dissertation provides an effective solution to Arabic text reuse. It builds a framework and develops an automated system for the identification of Arabic text reuse on the web. By itself, the Arabic language presents numerous natural language processing challenges. A model and algorithms are elaborated to address these challenges in the context of a heterogeneous web. The results demonstrate the soundness and the efficiency of the model.

The conceptual model and its software implementation constitute a direct support for a wider use of Arabic-related applications, such as data mining, text analysis, information retrieval. They provide also an operational platform for researchers from diverse fields to collect corpora and analyze socio-linguistic parameters.

Relevant Publications

- Anas Boubas, Leena Lulu, Boumediene Belkhouche, Saad Harous, "GENESTEM: A Novel Approach for an Arabic Stemmer Using Genetic Algorithms", IIT2011
- Anas Boubas, Leena Lulu, Boumediene Belkhouche, Saad Harous, "A Genetic-Based Extensible Stemmer for Arabic Verbs" (2014), Journal of Linguistica Communication.
- Leena Lulu, Boumediene Belkhouche, Saad Harous, "Candidate Document Retrieval for Arabic-based Text Reuse Detection on the Web", IT2016.
- Leena Lulu, Boumediene Belkhouche, Saad Harous, "Overview of Fingerprinting Methods for Local Text Reuse Detection", IT 2016.
- Leena M. Lulu, Boumediene Belkhouche and Saad Harous, 2016. A Local Text Reuse Detection Method for Arabic-Based Documents. ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2016), 17 pages. (under second revision)