A hadoop based platform for natural language processing of web pages and documents
5. Conclusions and future work

In this paper, a distributed system for crawling web documents and extracting keywords and keyphrases has been presented. The parallel architecture is provided by implementing the Apache Hadoop platform, while text annotation and key features extraction rely on the NLP opens source GATE platform. The main contributions offered by our work is the capability of executing general purpose GATE applications (including a wide range of NLP activities) in a distributed design (exploiting the benefits of scaling performances, especially for very large text corpora) with minimal code update and without the need for programmers to care about parallel computing constraints, such as task decomposition, mapping and synchronization issues. Evaluating processing performances on different cluster configurations (from 2 to 5 nodes) has showed a nearly linear scalability of the system, which is an encouraging result for future assessments on even larger datasets and cluster configurations. These actually represents open issues for future work. Moreover, it could be interesting to implement our keywords/keyphrases extraction module on other parallel computing environment, such as the cited Spark. Furthermore, in order to improve the quality of key features extraction, external knowledge resources could be used, especially Semantic repositories and frameworks, to allow the annotation of semantic features and relations.