|  |  |  |  |
| --- | --- | --- | --- |
| Title | Sentence Boundary Extraction from Scientific Literature of Electric Double Layer Capacitor Domain: Tools and Techniques. | | |
| Author(s) Name | Miah, M. S. U.; Sulaiman, J.; Sarwar, T. B.; Naseer, A.; Ashraf, F.; Zamli, K. Z.; and Jose, R | | |
| Contact Email(s) | saef@aiub.edu | | |
| Published Journal Name | Applied Sciences | | |
| Type of Publication | journal | | |
| Volume | 12 | Issue | 3 |
| Publisher | MDPI | | |
| Publication Date | Jan 2022 | | |
| ISSN |  | | |
| DOI | http://doi.org/10.3390/app12031352 | | |
| URL |  | | |
| Other Related Info. |  | | |
|  | | | |

|  |  |
| --- | --- |
| Abstract |  |
| Given the growth of scientific literature on the web, particularly material science, acquiring data precisely from the literature has become more significant. Material information systems, or chemical information systems, play an essential role in discovering data, materials, or synthesis processes using the existing scientific literature. Processing and understanding the natural language of scientific literature is the backbone of these systems, which depend heavily on appropriate textual content. Appropriate textual content means a complete, meaningful sentence from a large chunk of textual content. The process of detecting the beginning and end of a sentence and extracting them as correct sentences is called sentence boundary extraction. The accurate extraction of sentence boundaries from PDF documents is essential for readability and natural language processing. Therefore, this study provides a comparative analysis of different tools for extracting PDF documents into text, which are available as Python libraries or packages and are widely used by the research community. The main objective is to find the most suitable technique among the available techniques that can correctly extract sentences from PDF files as text. The performance of the used techniques Pypdf2, Pdfminer.six, Pymupdf, Pdftotext, Tika, and Grobid is presented in terms of precision, recall, f-1 score, run time, and memory consumption. NLTK, Spacy, and Gensim Natural Language Processing (NLP) tools are used to identify sentence boundaries. Of all the techniques studied, the Grobid PDF extraction package using the NLP tool Spacy achieved the highest f-1 score of 93% and consumed the least amount of memory at 46.13 MegaBytes. | |