"A .NET-Based Framework for Intelligent Text Extraction and Semantic Analysis of PDF Abstracts"

In recent years, the sheer volume of research papers has increased exponentially, making it difficult to manually extract and analyze information from research abstracts. Efficient systems for text extraction and analysis are essential for academic and scientific research in order to automate the process and provide meaningful insights. This paper proposes a .NET-based framework for intelligent text extraction and semantic analysis of PDF abstracts. The proposed framework utilizes the .NET Core platform and various natural language processing (NLP) techniques to automatically extract relevant information from academic PDF documents, focusing on research abstracts. The first step in the framework involves parsing and extracting the raw text from the PDF files using the iTextSharp library. The extracted text is then processed using several NLP methods, including tokenization, named entity recognition (NER), part-of-speech tagging, and lemmatization, to structure the data and identify key components such as research focus, methods, and results. Once the text has been processed, semantic analysis techniques are applied to understand the context of the research and its relevance to a given query. The system employs a combination of vector space models and machine learning-based algorithms, such as Latent Semantic Analysis (LSA) and Word2Vec, to generate semantic representations of the text and perform document retrieval and classification based on user queries. The framework also incorporates user feedback, enabling continuous learning and improvement of the extraction and retrieval processes. Additionally, the proposed system provides a user-friendly web interface built using ASP.NET Core, where researchers can upload PDFs, query for relevant abstracts, and view categorized results based on semantic relevance. The framework's effectiveness is demonstrated through a series of experiments with real-world academic datasets, showcasing its ability to provide accurate and efficient retrieval of relevant research abstracts.