

"AI-Powered PDF Abstract Extraction and Retrieval Using .NET and TensorFlow"

The extraction and retrieval of relevant research abstracts from PDF documents can be a challenging task due to the diverse formats and structures of academic papers. This paper proposes an AI-powered solution for PDF abstract extraction and retrieval using the .NET framework in combination with TensorFlow for deep learning. The proposed system aims to automate the extraction of abstracts from PDF documents and use machine learning models to enhance the retrieval process. The system starts by utilizing the iTextSharp library to extract raw text from PDF files. This text is then processed through an NLP pipeline, including steps such as tokenization, sentence segmentation, and part-of-speech tagging. For abstract extraction, the system applies a convolutional neural network (CNN) model trained using TensorFlow to detect and extract abstract sections based on structural patterns and context. The model is fine-tuned on a large dataset of academic PDFs, enabling it to identify the most relevant text sections. Once the abstracts are extracted, the system uses deep learning-based retrieval techniques, such as semantic search with embeddings generated by TensorFlow's pre-trained models, to return relevant abstracts based on user queries. The retrieval process is further enhanced by using contextual word representations, allowing the system to understand the meaning behind queries and match them to the abstracts accurately. The system is deployed as a web application using ASP.NET Core, offering researchers an intuitive interface to upload PDFs, search for abstracts, and receive ranked results. Evaluation results demonstrate that the AI-powered system significantly improves the accuracy of abstract extraction and retrieval compared to traditional rule-based approaches, making it a valuable tool for academic research.