PROJECT PROPOSAL

Date of proposal: 02/10/2024

Project Title:
Text auto detection with summarization

Group ID (As Enrolled in Canvas Class Groups): 10

Group Members (name , Student ID):

DING YI - A0295756J

YANG RUNZHI – A0297296H

LOU SHENGXIN - A0297330A

SHI HAOCHENG - A0296265R

LIU LIHAO - A0296992A

Sponsor/Client: *(Company Name, Address and Contact Name, Email, if any)*

Background/Aims/Objectives:

Background

With the rapid growth of digital documents in various formats such as PDFs, scanned images, and academic papers, efficient processing and extraction of relevant information from these documents have become a significant challenge. Traditional methods for handling document layout analysis and text recognition are often manual, time-consuming, and prone to errors. Recent advancements in deep learning and computer vision, particularly the YOLO (You Only Look Once) model, have enabled efficient and accurate detection of objects within images, which can be applied to document layouts to identify various elements such as headings, paragraphs, tables, and images. Additionally, combining these techniques with Optical Character Recognition (OCR) and Natural Language Processing (NLP) opens up the possibility of not only extracting text but also summarizing document content.

The need for automated tools to streamline document processing is critical in academia, business, and legal industries, where handling large volumes of documents requires efficiency and precision. By leveraging YOLOv10-Document-Layout-Analysis(YOLOv10), followed by OCR recognition, and integrating summarization techniques, this project aims to provide a comprehensive solution. Furthermore, users can simply find the related information by Wikipedia API , the project will enable users to quickly locate relevant information across numerous documents.

Aims

The aim of this project is to develop an automated system that first utilizes YOLOv10 for document layout analysis, followed by Tesseract for OCR text extraction. Subsequently, scibert will be used to extract key knowledge points from the text, and finally, the system will query those knowledge points using Wikipedia's online API, enabling enhanced document retrieval and contextual searchability.

Objectives

Document Layout Analysis: Use YOLOv10 to detect various document layout elements such as headers, paragraphs, tables, and images, organizing the document structure for further processing.

Text Recognition: Utilize Tesseract OCR technology to accurately extract text from the layout regions identified by YOLOv10, converting scanned or digital documents into editable and searchable text.

Knowledge Point Extraction: Employ SpaCy, a Natural Language Processing (NLP) tool, to extract key knowledge points from the recognized text. SpaCy will identify relevant entities, concepts, and relationships, creating a structured representation of the important content.

Knowledge Point Querying: Integrate Wikipedia's online API to query relevant contextual information based on the extracted knowledge points. This step provides additional insights and related information, enhancing the understanding of the document's content.

Performance Optimization: Ensure high accuracy and efficiency of YOLOv10 for layout detection, Tesseract for OCR text extraction, and SpaCy for knowledge extraction. Optimize the workflow for scalability and reliability in academic, legal, and business document processing.

The system will begin by collecting a diverse dataset of document formats, including PDFs, scanned documents, and images. These documents will be annotated to identify various layout elements such as headers, paragraphs, tables, and images. YOLOv10 will then be trained to detect these layout elements, outputting bounding boxes and labels for each detected element. Once the layout analysis is complete, an OCR engine, Tesseract, will be integrated to extract text from the identified text regions.

The recognized text will then be processed by SpaCy, to generate high-quality summaries of the document's content, capturing key ideas, conclusions, and important points. Claude will also be leveraged to enhance semantic analysis techniques, offering contextual expansions and insights by providing related background information based on the extracted content.

After the summarization step, the extracted and summarized content will be structured into a document knowledge base. This knowledge base will organize the extracted information, allowing for efficient indexing and classification based on keywords, topics, and fields of research. Claude will further assist in optimizing the search functionality, enabling users to query the knowledge base and retrieve relevant documents or information through advanced semantic search techniques. With Claude's natural language understanding capabilities, users will be able to make queries in natural language, and Claude will provide precise and contextually relevant search results.

Users will be able to quickly find and access key points across a large volume of documents, improving their ability to extract useful insights from the collected data. The final system will be designed as a user-friendly application where users can upload documents for automatic layout analysis, text extraction, summarization, and searchability. The project will also involve optimizing both YOLOv10 and the OCR engine to ensure high accuracy and efficiency, making the system suitable for real-world applications in academic, legal, and business document processing.

References:

[1]Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. arXiv preprint arXiv:2405.14458, 2024.

[2]Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text[J]. arXiv preprint arXiv:1903.10676, 2019.

[3]Smith R. An overview of the Tesseract OCR engine[C]//Ninth international conference on document analysis and recognition (ICDAR 2007). IEEE, 2007, 2: 629-633.