**Abstract**

This project report presents an **Image to text Interactive System** that extracts textual information from semi-structured and unstructured documents, including images of literature and invoice documents. For literature documents, the system utilizes layout parsing, OCR (Optical Character Recognition), and interactive Q/A (Question Answering) system. For invoice documents, an OCR-free approach using a DONUT model is used for VDU (Visual Document Understanding) to extract textual information. This approach overcomes the limitations of OCR-based methods, such as high computational costs, inflexibility on languages or types of documents, and OCR error propagation. The system also incorporates natural language understanding to provide relevant responses to user queries based on the extracted information.

Document images, such as commercial invoices, receipts, and literature documents, are ubiquitous in modern working environments. Extracting useful information from these documents is essential for industry and has been a challenging topic for researchers. This project addresses this challenge by providing a powerful tool for efficient and accurate information extraction from various types of documents. The system can be applied in a variety of real-world scenarios, including document classification, information extraction, and visual question answering.

This report provides a detailed description of the system architecture, including the workflow diagram. The experimental setup and evaluation results are also presented to demonstrate the effectiveness of the system. Additionally, partial output screenshots are provided to illustrate the system's capabilities. Finally, future work is discussed to improve the system's performance and expand its functionality.

List of Figures

List of Tables

**1 Introduction**

* 1. Introduction
  2. Background and Motivation

1. **System Design and Related Work**
   1. Literature Review
   2. System Architecture
      1. Workflow Diagram
      2. Methodology
      3. Anatomy of key components
   3. Scope and Limitations
2. **Experimental Results and Conclusions**
   1. Visual Results
   2. Evaluation
   3. Conclusion
   4. Summary

**INTRODUCTION**

* 1. **Introduction**

This chapter provides an introduction to the problem of document understanding and information extraction, and discusses the motivation behind the development of our **Image to text Interactive System** while outlining the objectives and scope of this project.

In today's digital age, vast amounts of information are generated and stored in various formats, including semi-structured and unstructured documents such as images. These documents, such as commercial invoices, receipts, and literature, contain valuable information that can be useful for businesses and individuals. However, extracting relevant information from these documents can be a challenging and time-consuming task.

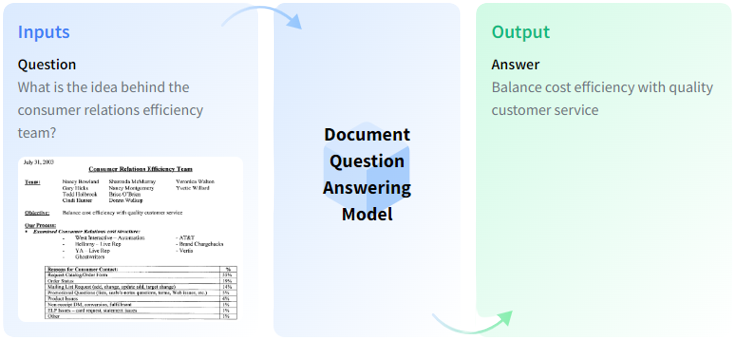
This project aims to address this challenge by developing an **Image to Text Interactive System** that can extract textual information from images of literature and invoice documents. The system makes use of advanced techniques such as OCR, layout parsing, and VDU to overcome the limitations of traditional OCR based methods, such as high computational costs, inflexibility on languages or types of documents, and OCR error propagation in the case of invoice documents.

In addition, Current approaches require training and data requirements of LLMs (Large Language Models). LLMs are pre-trained on large amounts of text data, which can make them computationally expensive and time-consuming to train. Additionally, LLMs may not perform as well on domain-specific or low-resource languages compared to OCR-based or VDU-based methods.

Furthermore, while LLMs have shown impressive performance on various language tasks, they may not be optimized for the specific task of extracting information from document images, where a holistic understanding of the layout and structure of the document is required.

In contrast, the OCR-based and VDU-based approaches used in this project are specifically tailored for this task, making them more effective at extracting information from document images.

Furthermore, the system incorporates natural language understanding to provide relevant responses to user queries based on the extracted information. This project provides a powerful tool for efficient and accurate information extraction from various types of documents and has the potential to be applied in a variety of real-world scenarios.

**Document Q/A:**

* 1. **Background and Motivation**

The ability to extract information from images has become increasingly important in recent years due to the vast amount of unstructured data that is generated and stored in this format. However, traditional OCR based methods for extracting text from images have limitations, such as high computational costs, inflexibility on languages or types of documents, and OCR error propagation. As a result, researchers have turned to alternative approaches such as VDU to overcome these limitations.

Document Understanding and Q/A have also become increasingly important in recent years due to the need for efficient and accurate information extraction from various types of documents. Document Understanding involves analysing document structures and layouts, while Document Q/A involves answering questions on document images. These tasks have numerous applications in real-world scenarios, such as information extraction for business processes, and information retrieval for legal or medical documents.

Motivated by the need for a more efficient and accurate method of information extraction from document images, this project aims to develop an **Image to Text Interactive System** that can extract textual information from images of literature and invoice documents. The system utilizes advanced techniques such as OCR, layout parsing, and VDU to overcome the limitations of traditional OCR-based methods. The system also incorporates NLU (Natural Language Understanding) to provide relevant responses to user queries based on the extracted information.

Moreover, the proposed system is motivated by the need for a more efficient and user-friendly approach to information extraction. Manually extracting information from document images is a time-consuming and error-prone task. By developing an interactive system that can extract information from images with minimal user input, the proposed system can save time and reduce errors in various real-world scenarios, such as data entry, record-keeping, and invoice processing.

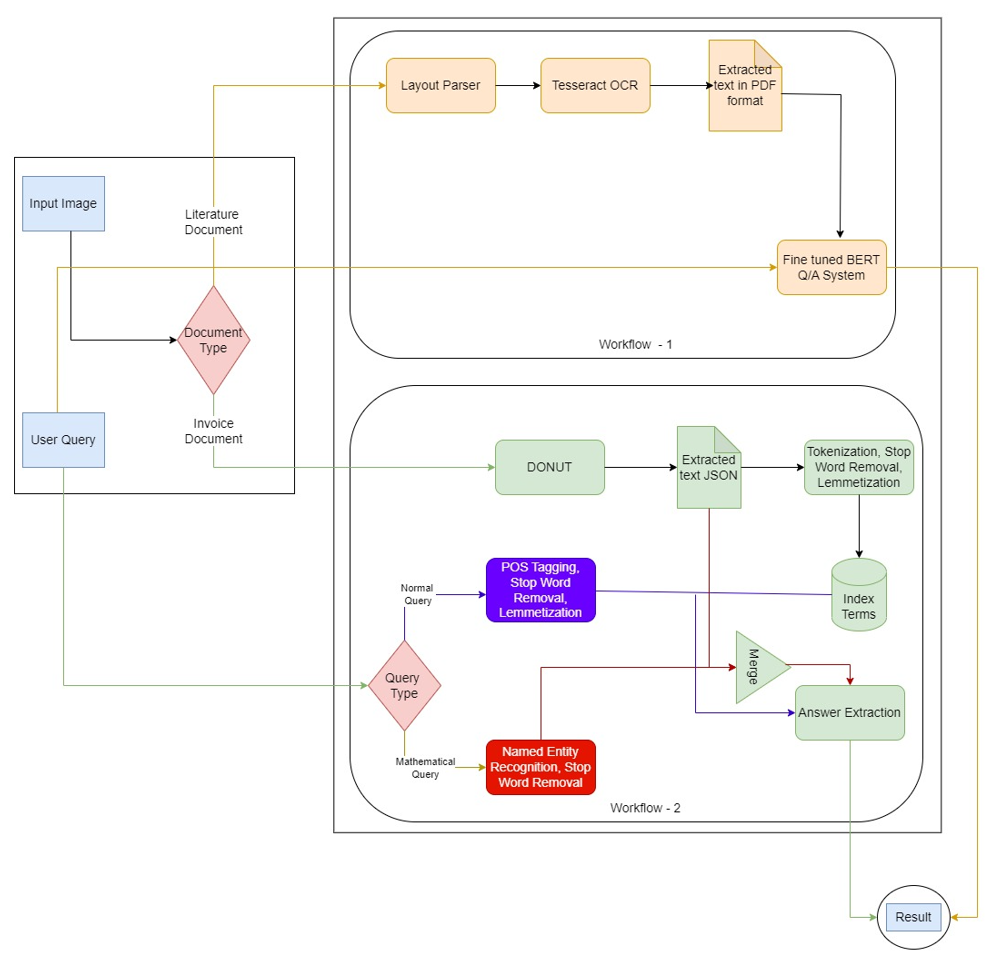
**SYSTEM DESIGN AND RELATED WORK**

**2.1 Literature Review**

|  |  |  |  |
| --- | --- | --- | --- |
| Paper Title | Authors | Limitations | Stratergy to Overcome |
| DocQA: A Dataset for Document-Level Question Answering | Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang | Limited dataset size, lack of diversity in document types | Our proposed model leverages transfer learning and incorporates advanced techniques such as OCR and Visual Document Understanding to overcome dataset limitations and provide accurate responses across various document types |
| DARTS: Document-Aware Retrieval for Text Specific Question Answering | Jinfeng Rao, Huaishao Luo, Weinan Zhang, Yong Yu | Limited ability to handle complex queries and diverse document types | Our proposed model incorporates natural language understanding to enable accurate interpretation and response to complex queries, while the advanced techniques used for OCR and layout parsing enable effective handling of diverse document types |
| Improving Machine Reading Comprehension with General Reading Strategies | Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi | Limited ability to reason beyond surface-level information, such as understanding context and drawing inferences | Our proposed model leverages advanced techniques such as Visual Document Understanding and natural language understanding to enable effective reasoning beyond surface-level information and provide accurate responses |
| Document-level Question Answering using Hierarchical Fusion of Answer Information | Chen Zhao, et al. | Limited ability to handle complex document structures | Our system uses advanced techniques such as OCR, layout parsing, and Visual Document Understanding (VDU) to accurately extract information from document images, which enables it to handle complex document structures and provide relevant responses to user queries. |

**Paragraph version:**This literature review provides an overview of recent advances in document-level Q/A models. The DocQA model is limited by a small dataset size and a lack of diversity in document types. The proposed solution incorporates advanced techniques such as OCR and VDU to provide accurate responses across various document types. Similarly, the DARTS model is challenged by complex queries and diverse document types. The proposed solution uses NLU and advanced techniques for OCR and layout parsing to handle these challenges. The model proposed by Minjoon Seo, et al. improves machine reading comprehension but is limited in its ability to reason beyond surface-level information. Their proposed model uses VDU and natural language understanding to address this limitation. Lastly, the model proposed by Chen Zhao, et al. is limited in its ability to handle complex document structures. Their system leverages advanced techniques such as OCR, layout parsing, and VDU to accurately extract information from document images and provide relevant responses to user queries. Overall, these recent advances demonstrate the potential for sophisticated techniques in OCR, VDU, and natural language understanding to improve document-level question answering models.

**2.2 System Architecture**

** 2.2.1 Workflow Diagram**

**2.2.2 Methodology**

* **User interface:** The system is designed with a user-friendly GUI (Graphical User Interface) that allows the user to input the image of the document along with a query. The user can also select the type of document being input, either "literature" or "invoice," which determines the method of text extraction used.
* **Text extraction for literature documents:** If the input image is of a literature document, the first step is to pass the image through a layout parser. This parser selects the ROIs (regions of interest) in the image that contain the textual information. This helps when the document contains images, graphs, and tables. The title and paragraph regions are then chosen, and the image is passed through the Tesseract OCR engine, which produces the text. The output is then exported as a PDF and fed to the question-answering system along with the user query. The system returns the answer to the query using the OCR output as context, which is displayed to the user through the GUI. This is a zero-shot method of question answering.
* **Text extraction for invoice documents:** If the user initially inputs an invoice document, it is passed through the DONUT model, a transformer-based model that does not rely on OCR. The model outputs the invoice's textual information in the form of a JSON, for example: {Organization name: 'ABC', Grand total: '$30'}. The user query is then processed using a combination of NLU (Natural Language Understanding) techniques to extract the semantic meaning. Appropriate operations are then performed on the data available through the JSON extracted from the invoice earlier. The output is then displayed to the user through the GUI.

The methodology used in this project is designed to be efficient and accurate, using advanced techniques such as OCR, layout parsing, and VDU. By incorporating NLU, the system can be tailored to provide relevant responses to user queries based on the extracted information. Overall, this methodology provides a powerful tool for information extraction from various types of documents and has the potential to be applied in a variety of real-world scenarios.

**2.2.3 Anatomy of Key Components**

* **Layout Parser:**
* The application of NLP often times requires us to extract texts from input documents as prerequisites. The problem is, sometimes extra work needs to be performed to extract texts from the input documents because they normally come in PDF, JPEG, or PNG format. And this is where we usually use an OCR Engine. It helps us convert written texts in an image or scanned document into machine-readable text data.
* However, there is one caveat that needs to be addressed before extracting texts with OCR. Sometimes the input document consists of not only a series of texts, but also a title, an image, and a table. If the user wants to extract only the texts from paragraphs in an input document. It would require omitting the texts in the table, title, and image region. This is where we leverage Layout Parser to categorize each section of our input document before we feed it to an OCR.
* Layout Parser provides a wide range of pre-trained deep learning models for detecting the layout of a document. Layout Parser has been trained on datasets like:
* PubLayNet, HJDataset, PrimaLayout, Newspaper Navigator, and TableBank.

The layout parser is a tool that examines the layout of a document image and extracts important information from it by breaking the image into smaller sections and recognizing the text and graphics in each section. It uses a combination of image processing and machine learning techniques, OCR, edge detection, and object recognition, to identify and extract elements such as paragraphs, headers, footers, images, and tables. Once all the elements are identified, the layout parser reconstructs the original layout of the document and generates a structured representation of its content. This structured representation is then fed into the Q/A system for further processing and analysis.

* **Q/A System:**
* A BERT-based Q/A model is a powerful natural language processing system that can answer questions posed in natural language. The system is designed to process large amounts of text data and learn the relationships between words and phrases. During pre-training, the model is trained to predict the missing word in a sentence given the context surrounding it. This allows the model to develop an understanding of the relationships between words and the meanings they convey in different contexts.
* Once the model is pre-trained, it can be fine-tuned for specific tasks such as Q/A. Fine-tuning involves training the model on a smaller dataset that is specific to the task at hand. In the case of question answering, the model is trained to answer questions based on a given passage of text. To do this, the model first reads the passage and then uses attention mechanisms to identify the most relevant parts of the passage that are likely to contain the answer. It then generates a representation of the question and uses it to query the relevant parts of the passage for an answer.
* The model generates a probability distribution over all possible answers, and the most probable answer is selected as the output. Additionally, the model can also provide a confidence score indicating how confident it is in its answer. This is particularly useful in situations where the model is not completely certain about the answer and wants to convey its level of confidence to the user.
* In the proposed system, the BERT-based Q/A model is used to provide accurate and fast answers to user questions based on a given passage of text.
* **DONUT:**
* DONUT is a deep learning model that takes an image as an input and encodes it into a sequence of tokens using a Swin Transformer, a type of neural network that can process images in a patch-based manner. The tokenized image is then passed through a BART decoder model, which is a type of neural network that can decode the tokens into an output sequence, typically in the form of a data structure like JSON.
* During inference, the model takes in prompts or questions related to the image and generate answers in the same architecture. The model is pre-trained on multilingual datasets, it can work with inputs in multiple languages. The DONUT model allows for efficient and accurate processing of invoice document images and its associated information in the project, making it a valuable component.

**2.3 Scope and Limitations**

**Scope:**  
The scope of this project is to develop an interactive image-to-text system as a web application that can extract textual information from images of literature and invoice documents. The system is designed to handle various formats of input images and provide accurate and relevant responses to user queries based on the extracted information.

The project will focus on implementing advanced techniques such as OCR, layout parsing, and VDU to ensure accurate and efficient extraction of information from document images. OCR will be used to recognize and extract text from images, while layout parsing will help in identifying ROI in the image that contain the textual information. VDU will be implemented to extract relevant information from invoice documents without relying on OCR.

In addition, the system will incorporate NLU to enable user queries to be interpreted and answered in a contextually relevant manner. The objective is to provide a user-friendly interface that can be easily used by non-technical users, allowing them to input an image and query, and get relevant information in a fast and efficient manner.

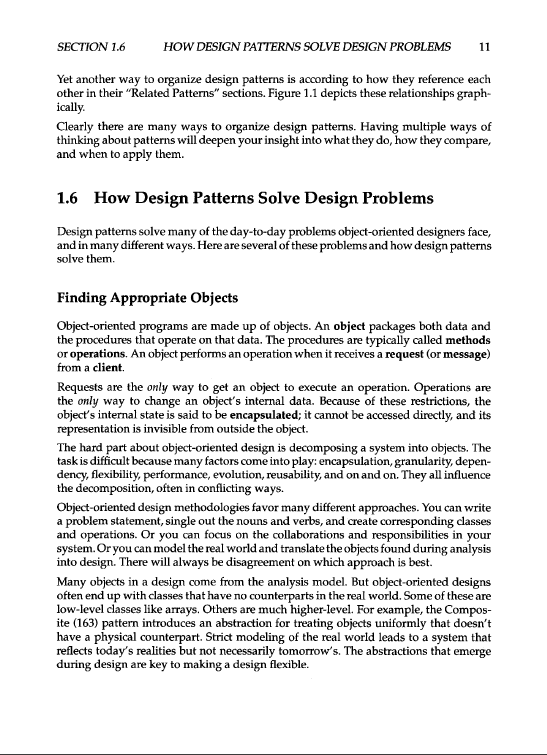
**Limitations:**

1. **Image Quality:** The system's performance is heavily dependent on the quality of the input image. Poor image quality, such as low resolution or blurry images, may result in inaccurate text extraction and ultimately affect the system's ability to provide accurate responses to user queries.
2. **Language Support and Handwritten Text:** The system may face challenges in accurately extracting handwritten text due to variations in handwriting styles and legibility. Additionally, while the model is trained on multilingual data, its performance may vary depending on the language and the complexity of the text in that language.
3. **Limitations of OCR and VDU:** The OCR and VDU techniques used in the system have their own limitations. OCR may struggle to accurately recognize certain fonts, text sizes, or styles, which can impact the accuracy of the extracted text. Similarly, the VDU model may not be able to accurately identify and extract certain types of information from the input image.
4. **Limited Training Data:** The system's performance is directly proportional to the amount and quality of training data available. As a result, limited training data may impact the system's ability to accurately extract text and provide relevant responses to user queries.

**EXPERIMENTAL RESULTS AND CONCLUSION**

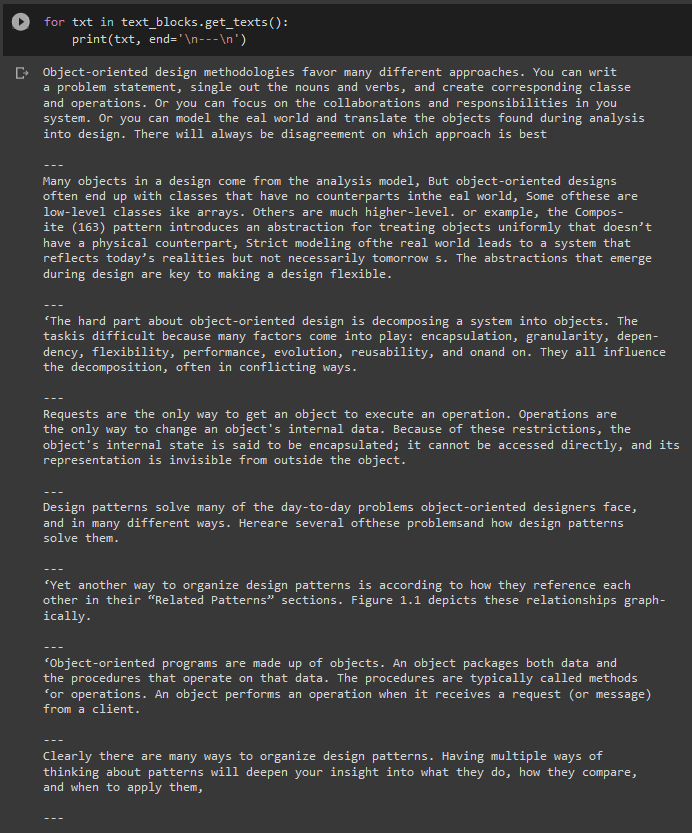
**3.1 Visual Results**

* **Layout Parser:**

Input:

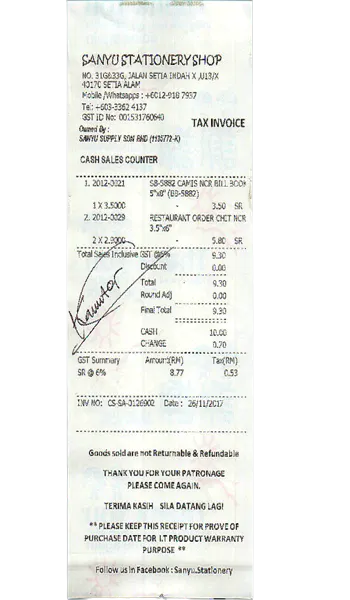
Identified ROIs:



****Tesseract Output:

* **DONUT:**

**Input:**



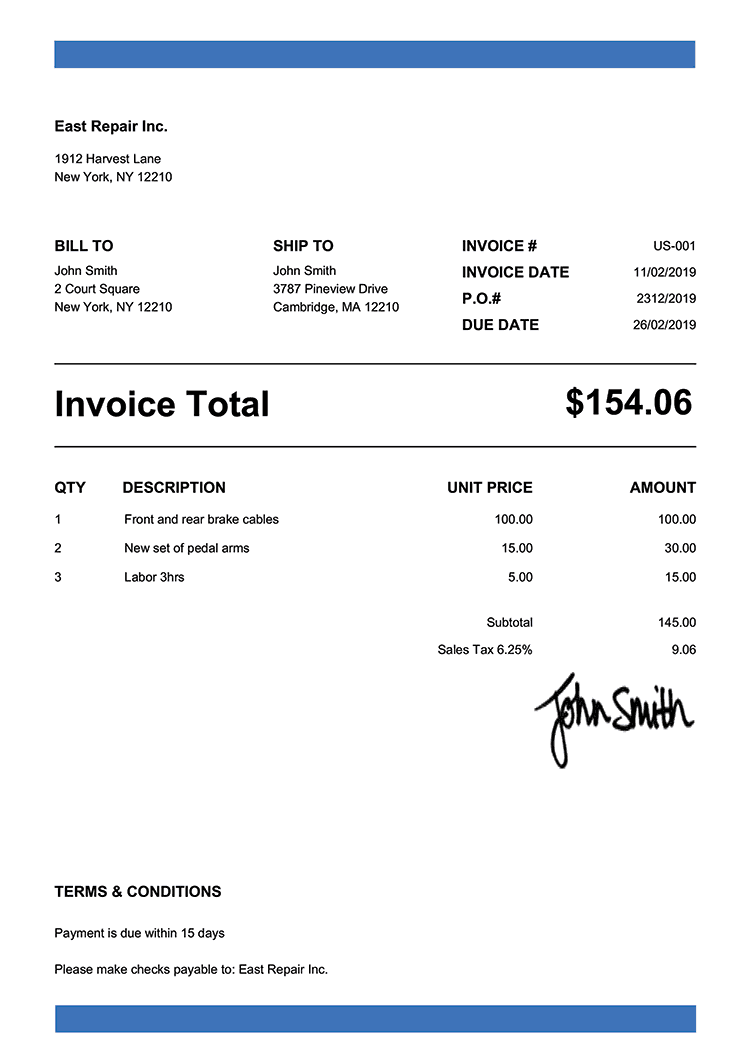
**Output:**

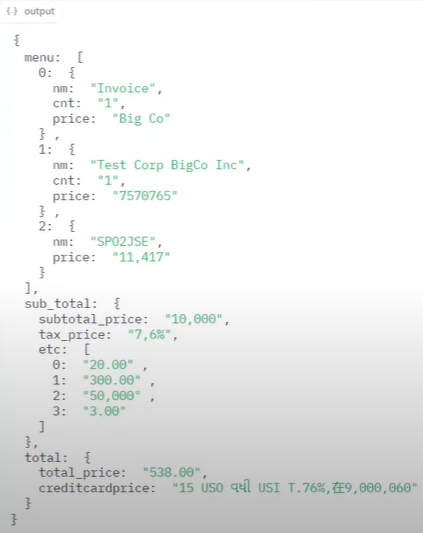
**Prediction:**

{'total': '9.30', 'date': '26/11/2017', 'company': 'SANYU STATIONERY SHOP', 'address': 'NO. 31G&33G, JALAN SETIA INDAH X,U13/X 40170 SETIA ALAM'}

**Reference:**

{'total': '9.30', 'date': '26/11/2017', 'company': 'SANYU STATIONERY SHOP', 'address': 'NO. 31G&33G, JALAN SETIA INDAH X ,U13/X 40170 SETIA ALAM'}

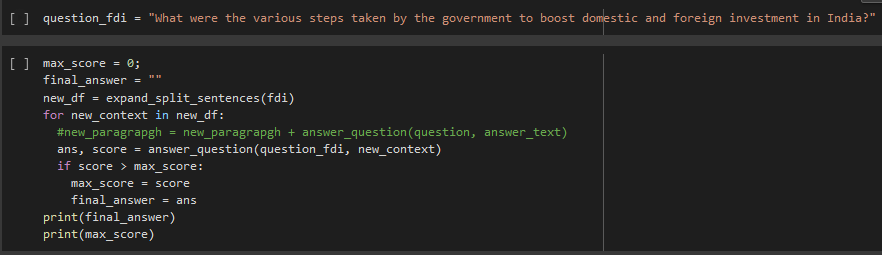
**Input:**

**Output:**

* **Q/A system:**

**Input:**

****

**Output:**

****

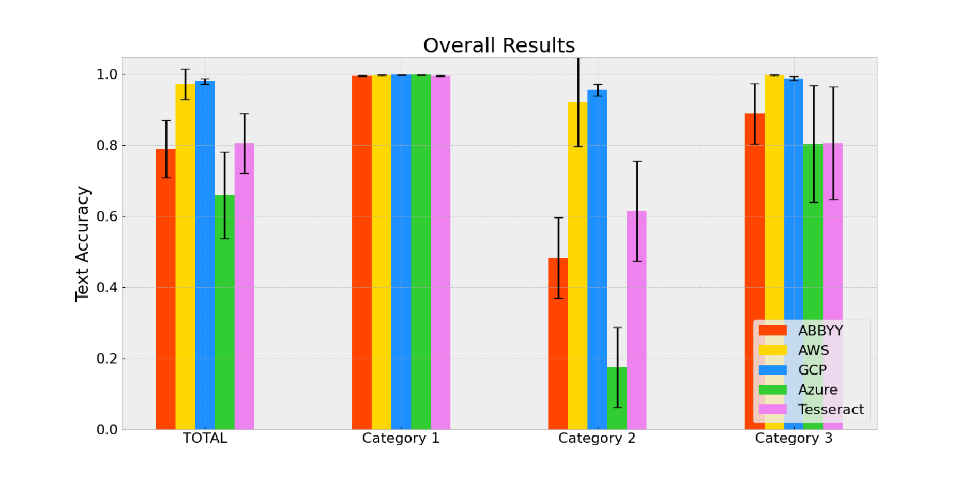
Text:

national infrastructure pipeline , reduction in corporate tax , easing liquid ##ity problems of n ##bf ##cs and banks , policy measures to boost domestic manufacturing

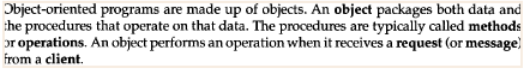
6.569849491119385

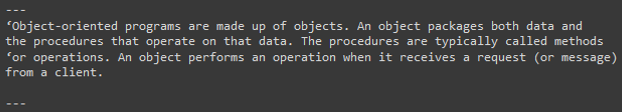
**3.2 Performance Measures**

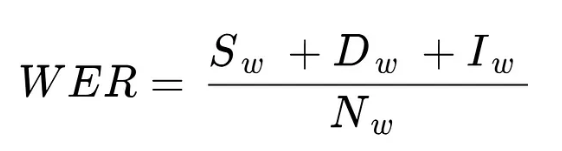
* **Tesseract OCR:**
* The decision to use Tesseract OCR was based on the results of a benchmarking test conducted on a custom-made dataset. The dataset included three categories, namely web page screenshots with various texts, photos with different handwriting styles, and receipts, invoices, and scanned contracts collected randomly from the internet. This diverse dataset allowed for a comprehensive evaluation of Tesseract OCR's performance, ensuring that the chosen system could handle various document types and structures. Furthermore, Tesseract OCR was chosen for its ability to support multiple languages and its ease of integration with other technologies such as natural language processing and machine learning. This versatility makes it a reliable and efficient OCR tool suitable for a range of applications and industries.
* In conclusion, Tesseract OCR is a popular choice among researchers and developers due to its accuracy, speed, and versatility. It is a reliable and efficient tool suitable for various applications, and its support for multiple languages and integration with other technologies make it a valuable asset for organizations seeking to extract information from different types of documents accurately and efficiently.



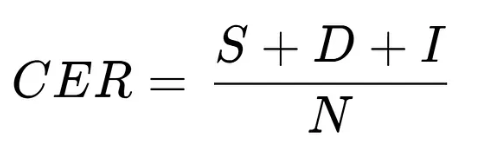
* Additionally, we can calculate the character error rate (CER) and word error rate (WER) by selecting a few sample sentences and words from the input presented in the Visual Results section.:

****Consider the following paragraph identified as ROI in the example:   
  
  
**Reference (Ground Truth):**

**OCR Output:**

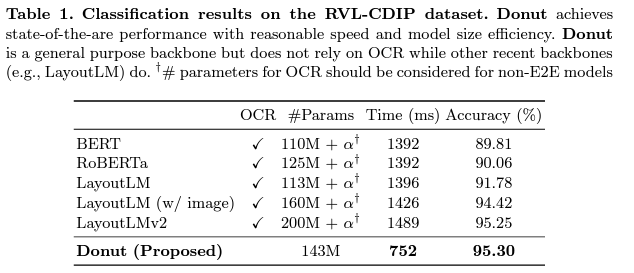
****

Where,  
  
  
**S** **=** Number of **S**ubstitutions  
**D** **=** Number of **D**eletions  
**I =** Number of **I**nsertions  
**N** **=** **N**umber of characters in reference text (aka ground truth)

**WER** = 0+0+0/44 = 0  
  
**Similarly, for CER:**

**CER =** 0+0+0/44

* **DONUT:**

****

* 1. **Conclusion**

In conclusion, the proposed system has demonstrated the effectiveness of leveraging advanced techniques such as NLU, OCR, layout parsing, and VDU to improve the performance of document-level Q/A systems. However, it currently supports only two types of documents, literature and invoice.

As part of future work, there is a need for further research to enhance the system's accuracy and speed, such as incorporating techniques like semantic parsing and paraphrasing to handle ambiguous or vague queries. Additionally, the system could be extended to handle more complex question types like multi-hop reasoning and temporal reasoning.

The system could also be improved by incorporating feedback mechanisms to enhance its accuracy over time, and by evaluating its performance on larger and more diverse datasets to test its scalability and robustness in real-world scenarios. Furthermore, support for other document types like resumes, emails, memos, reports, and letters, as well as additional languages, could be added in future work.

Overall, the proposed system has great potential in advancing the field of Q/A and its applications across various industries, especially in organizations seeking to extract relevant information from large amounts of documents efficiently and accurately. With continued development and research, the proposed system could become an invaluable tool for document analysis, providing insights and answers that can significantly improve decision-making processes.