MINI PROJECT REPORT

PYQ Analyser

Submitted in partial fulfilment of the requirements for the degree of

Bachelor of Engineering

in

Computer Science and Business Systems

Submitted By

Siddharth Aggarwal	102203375
Surya Rathee	102203232

Under the supervision of:

Dr. Arun Singh Pundir

Assistant Professor CSED TIET

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology

Patiala – 147004

November 2024

Table of Contents

1. Introduction 1

2. Background 3

3. Objectives 6

4. Methodology...8

5. Limitations 15

6. Conclusion and Future Work...16

6.1. Conclusions:..16

6.2. Future Works: 16

7. Bibliography and References...18

1. Introduction

In the dynamic landscape of educational technology and machine learning, our pioneering project emerges as a transformative solution to the complex challenge of automated question extraction and classification from diverse academic textbooks. As the educational ecosystem continues to evolve, demanding innovative approaches to knowledge digitization and analysis, our research presents a sophisticated AI-driven methodology that transcends traditional boundaries of document processing and educational content management.

The surge in digital learning resources and the increasing complexity of academic textbooks underscore the critical need for intelligent systems capable of navigating intricate document structures, extracting meaningful educational content, and enabling advanced analytical capabilities. Our project—a comprehensive Question Extraction and Classification System—represents a groundbreaking approach to addressing these multifaceted challenges, leveraging cutting-edge technologies in optical character recognition, natural language processing, and machine learning.

At the heart of our research lies a nuanced understanding of the inherent complexities surrounding question extraction from varied academic texts. Each textbook presents a unique landscape of formatting, chapter organization, and question presentation, rendering traditional extraction methods inadequate. Our system meticulously navigates these challenges through a sophisticated, adaptable pipeline that combines the robust optical character recognition capabilities of PaddleOCR with advanced natural language processing techniques and state-of-the-art BERT-based classification.

The core objectives of our project encompass a comprehensive spectrum of technological and educational innovation:

i. Intelligent Question Extraction:
Our system transcends conventional document parsing methodologies, employing advanced optical character recognition to excavate questions from diverse textbook formats. By developing specialized extraction scripts tailored to individual book structures, we ensure unprecedented accuracy and adaptability in question retrieval.

ii. Semantic Preprocessing and Enrichment:
Utilizing sophisticated NLTK preprocessing techniques, our pipeline transforms raw extracted text into semantically enriched, structured data. Through strategic stopword removal, tokenization, and contextual analysis, we prepare the foundation for sophisticated machine learning-driven classification.

iii. Contextual Question Classification:
Leveraging the powerful semantic understanding capabilities of BERT models, our system classifies extracted questions with remarkable precision, mapping them not just to keywords but to intricate contextual relationships within academic content.

The significance of our research extends far beyond mere technological demonstration. By creating a robust, adaptable system for automated question extraction and classification, we address critical challenges in educational content digitization, knowledge management, and learning analytics.Our approach offers educators and researchers a powerful tool for transforming traditionally static educational resources into dynamic, analyzable datasets.

In the subsequent sections of this report, we will embark on a comprehensive exploration of our methodology, dissecting each technological component with surgical precision. We will navigate through the intricacies of our optical character recognition strategies, delve into the nuanced preprocessing techniques, and illuminate the sophisticated machine learning models that form the backbone of our classification system.

Our journey represents more than a technological endeavor—it is a testament to the transformative potential of artificial intelligence in reshaping educational content analysis, offering a glimpse into a future where complex academic resources can be intelligently parsed, understood, and leveraged with unprecedented efficiency.

2. Background

In the dynamic landscape of educational technology and machine learning, the evolution of document processing and knowledge extraction has experienced a profound transformation. The proliferation of diverse academic resources, coupled with the increasing complexity of educational content, has created an urgent need for intelligent systems capable of navigating the intricate terrain of textbook analysis.

Our Question Extraction and Classification Project emerges from a critical understanding of the challenges inherent in traditional document processing approaches. The digital revolution in education has fundamentally reshaped how knowledge is captured, analyzed, and utilized, demanding innovative solutions that can bridge the gap between raw textual data and structured, meaningful insights.

Technological Foundations

The project draws upon a rich ecosystem of cutting-edge technologies that form its robust technological infrastructure:

Optical Character Recognition (PaddleOCR)

PaddleOCR emerged as a critical computational solution for our research, particularly addressing the computational challenges presented by extensive academic textbooks spanning approximately 1,000 pages. Leveraging its robust GPU acceleration capabilities, the toolkit transformed our text extraction process from a potentially weeks-long endeavor to a matter of hours. Developed by Baidu, this open-source OCR toolkit was strategically selected for its exceptional performance in high-volume document processing, offering unprecedented text extraction speed and accuracy.

The GPU acceleration of PaddleOCR became instrumental in our research methodology, enabling us to:

- Process massive document volumes efficiently
- Maintain high-resolution text extraction quality
- Minimize computational time and resource consumption
- Handle complex multi-page academic textbook layouts with remarkable precision

By utilizing GPU-powered OCR, we could simultaneously process multiple document pages, breaking through traditional computational bottlenecks. The toolkit's ability to distribute computational load across graphics processing units ensured that our 1,000-page document processing remained swift and reliable, transforming what could have been a weeks-long manual extraction process into a streamlined, automated workflow.

PaddleOCR's adaptability to different fonts, complex layouts, and varying image qualities made it uniquely suited to our research objectives, particularly when dealing with diverse academic textbook formats. The GPU acceleration not only improved processing speed but also maintained exceptional accuracy in text extraction, proving crucial for our comprehensive question classification system.

Natural Language Processing Ecosystem

The project leverages a comprehensive suite of Natural Language Processing (NLP) technologies:

- NLTK (Natural Language Toolkit): Providing advanced preprocessing capabilities
- Pandas: Enabling sophisticated data manipulation and cleaning
- Scikit-learn: Supporting advanced preprocessing and feature engineering techniques

Machine Learning Classification

At the heart of our approach lies the BERT (Bidirectional Encoder Representations from Transformers) model, a state-of-the-art machine learning architecture that brings unprecedented semantic understanding to question classification. BERT's ability to contextualize language nuances makes it an ideal choice for parsing complex educational content.

Deployment

Streamlit emerges as our primary deployment platform, offering a seamless bridge between complex machine learning models and user-friendly interfaces. This Python framework allows us to transform our sophisticated question extraction and classification system into an interactive, accessible tool. The deployment strategy emphasizes:

- Intuitive user interfaces
- Real-time processing capabilities
- Interactive visualization of extracted and classified questions

Data Visualization with Seaborn

Seaborn represents a sophisticated data visualization library that elevates our project's analytical capabilities by transforming raw data into insightful visual narratives. Unlike basic plotting libraries, Seaborn is specifically designed for statistical graphics, offering deep integration with Python's data structures and providing a more nuanced approach to data representation.

PyTorch: Advanced Machine Learning Framework

PyTorch emerges as a powerful deep learning framework that provides exceptional flexibility and dynamic computation capabilities for our question classification model.

PyTorch Advantages in Our Project

1. Dynamic Computational Graphs
    - Enables real-time modification of neural network architectures
    - Supports complex BERT model fine-tuning
    - Allows seamless experimentation with model configurations
2. GPU Acceleration
    - Optimizes training processes for BERT models
    - Enables faster computation of complex semantic embeddings
    - Supports distributed training strategies

Research Context

Our project is positioned at the intersection of several critical research domains:

- Educational Technology
- Machine Learning
- Natural Language Processing
- Document Analysis

The motivation stems from a fundamental challenge in modern education: the need to transform static, hard-to-analyze textbook content into dynamic, processable datasets. Traditional methods of question extraction and classification have been limited by:

- Inconsistent document formatting
- Lack of adaptable extraction techniques
- Limited semantic understanding of educational content

Technological Innovation

What distinguishes our approach is its holistic methodology. Unlike traditional document processing systems, our project:

- Develops book-specific extraction scripts
- Implements advanced semantic mapping techniques
- Utilizes machine learning for contextual understanding
- Provides a flexible framework for handling diverse educational resources

3. Objectives

The primary objectives of our Question Extraction and Classification Project are comprehensive and multifaceted, addressing critical challenges in educational content analysis and student learning support:

i. Comprehensive Question Extraction

To develop an advanced AI-driven system capable of intelligently extracting questions from diverse academic textbooks, overcoming the challenges of varied formatting, font styles, and document layouts. This objective aims to transform static textbook content into a dynamic, machine-readable dataset that can be systematically analyzed and utilized for educational insights.

ii. Semantic Classification and Mapping

To implement a sophisticated BERT-based classification mechanism that goes beyond surface-level keyword matching, providing deep semantic understanding of extracted questions. The goal is to create a nuanced mapping that:

- Identifies the contextual relationships between questions
- Categorizes questions based on their underlying conceptual complexity
- Provides insights into the thematic structure of academic content

iii. Examination Preparation Optimization

To develop an intelligent system that helps students identify and prioritize key topics critical for college examinations. By analyzing question patterns, frequency, and semantic importance, the project aims to:

- Highlight frequently examined topics across different textbooks
- Create targeted revision guides
- Provide strategic insights into examination-oriented content
- Reduce student study anxiety by offering data-driven preparation strategies

iv. Temporal and Thematic Content Analysis

To analyze the distribution of questions across different chapters, academic units, and conceptual domains. This objective seeks to:

- Identify topic weightage in academic resources
- Understand the evolving emphasis of educational content
- Provide insights into curriculum design and knowledge progression

v. Machine Learning Enhanced Question Difficulty Assessment

To develop an innovative approach for automatically assessing question complexity and difficulty levels. The system will:

- Categorize questions based on cognitive complexity
- Provide students with stratified learning resources
- Enable personalized study path recommendations
- Support adaptive learning strategies

vi. Cross-Textbook Comparative Analysis

To create a robust methodology for comparing question patterns across multiple academic textbooks, enabling:

- Identification of common educational themes
- Understanding variations in content presentation
- Supporting educators in curriculum alignment
- Providing insights into interdisciplinary knowledge connections

Strategic Significance

These objectives collectively represent a holistic approach to revolutionizing educational content analysis. By leveraging advanced machine learning techniques, our project aims to:

- Transform how students interact with academic resources
- Provide data-driven insights into learning materials
- Support more efficient and targeted study methodologies
- Bridge the gap between traditional textbook learning and modern technological approaches

4. Methodology

In the intricate world of educational technology, the methodology of our Question Extraction and Classification Project represents a sophisticated journey through the intersection of artificial intelligence, optical character recognition, and semantic understanding. Like an archaeologist carefully excavating layers of historical artifacts, our approach meticulously deconstructs the complex terrain of academic textbooks, transforming static educational resources into dynamic, analyzable knowledge repositories.

Conceptual Framework

Our methodology emerges from a fundamental reimagining of how educational content can be processed and understood. Traditional approaches to document analysis have been limited by rigid, manual extraction methods that fail to capture the nuanced semantic relationships inherent in academic texts. In contrast, our project develops a holistic, technology-driven pipeline that treats each textbook as a living, interconnected ecosystem of knowledge.

The journey begins with a critical observation: academic textbooks are far more than mere collections of words and questions. They are intricate knowledge landscapes, where each question represents a carefully constructed pathway to understanding. Our methodology is designed to not just extract these questions, but to understand their deeper contextual significance, their relationship to broader conceptual frameworks, and their potential value in student learning.

Technological Architecture

At the heart of our approach lies a carefully orchestrated technological symphony, where multiple advanced tools and techniques converge to create a powerful, intelligent system:

1. **Computational Preprocessing**: PaddleOCR-GPU serves as our primary text extraction mechanism, leveraging GPU acceleration to transform time-consuming manual processes into swift, accurate digital conversions. By processing extensive documents spanning hundreds of pages, we transcend traditional computational limitations.
2. **Semantic Intelligence**: The BERT (Bidirectional Encoder Representations from Transformers) model acts as our cognitive engine, providing deep contextual understanding that goes beyond surface-level text analysis. It doesn't just read questions; it comprehends their underlying semantic structures, their complexity, and their potential educational significance.
3. **Machine Learning Classification**: Our custom neural network architecture transforms raw extracted text into structured, meaningful datasets. By implementing advanced classification techniques, we create a system that can dynamically categorize questions, understand their relationships, and provide insights that were previously invisible.

Methodological Innovation

What distinguishes our approach is not just the technological sophistication, but the philosophical approach to educational content processing. We view each textbook not as a static document, but as a dynamic knowledge ecosystem waiting to be understood.

Our methodology can be conceptualized as a multi-stage transformation:

- Raw Document → Preprocessed Image
- Preprocessed Image → Extracted Text
- Extracted Text → Semantic Embedding
- Semantic Embedding → Intelligent Classification

This approach allows us to progressively enrich and understand educational content, creating layers of insight that can be leveraged by students, educators, and researchers.

Strategic Objectives

Beyond the technical implementation, our methodology is driven by several strategic objectives:

- Democratizing access to educational insights
- Providing data-driven learning strategies
- Supporting personalized educational experiences
- Creating reproducible research methodologies

In the subsequent sections, we will meticulously deconstruct each component of this methodology, illuminating the intricate processes, technological innovations, and strategic thinking that underpin our approach.

4.1. `question_extractor.ipynb`: (Comprehensive Data Preprocessing and Visualization Pipeline)

PaddleOCR-GPU Integration

- Leverages GPU acceleration for high-performance text extraction
- Handles large academic documents (1000+ pages) efficiently
- Supports multiple language and font configurations
- Implements advanced image preprocessing techniques

Visualization Strategies

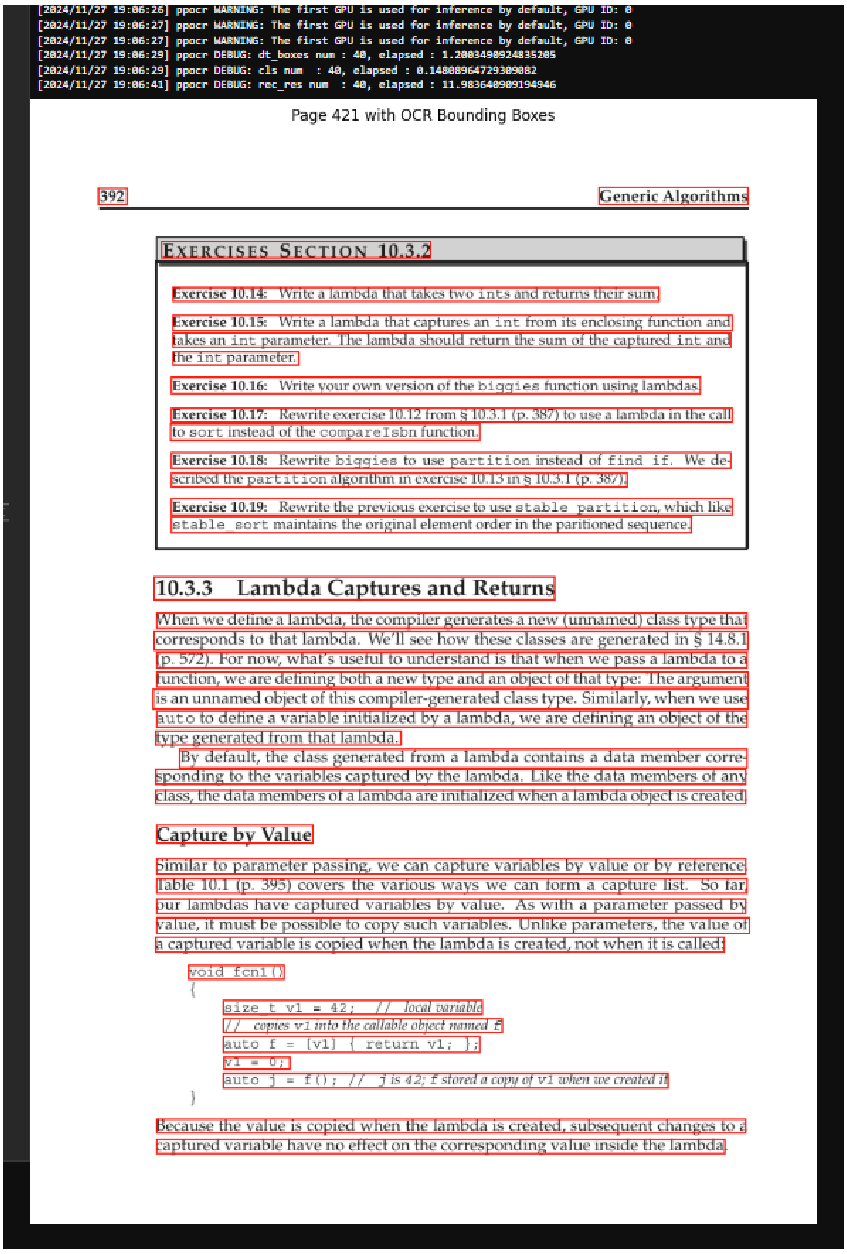
- **Matplotlib Bounding Box Visualization**

1. PDF Document Loading:
    - Uses PyMuPDF (`fitz`) for efficient PDF processing
    - Allows high-quality rendering of PDF pages
    - Supports various rendering resolutions

2. Image Conversion:
    - Converts PDF page to high-resolution pixel map
    - Uses a scaling matrix to control image quality
    - Transforms PDF content into a NumPy array for processing

3. Bounding Box Extraction
    - Retrieves text blocks with precise location information
    - Differentiates between text, images, and other elements
    - Provides granular control over visualization

4. Visualization Techniques
    - Matplotlib used for sophisticated plotting
    - Adds colored rectangles to represent text locations
    - Allows customization of rectangle properties



```
[2024/11/27 19:06:26] ppocr WARNING: The first GPU is used for inference by default, GPU ID: 0
[2024/11/27 19:06:27] ppocr WARNING: The first GPU is used for inference by default, GPU ID: 0
[2024/11/27 19:06:27] ppocr WARNING: The first GPU is used for inference by default, GPU ID: 0
[2024/11/27 19:06:29] ppocr DEBUG: dt_boxes num : 40, elapsed : 1.2003490924835205
[2024/11/27 19:06:29] ppocr DEBUG: cls num  : 40, elapsed : 0.14808964729309082
[2024/11/27 19:06:41] ppocr DEBUG: rec_res num  : 40, elapsed : 11.983640909194946
```

Page 421 with OCR Bounding Boxes

392 Generic Algorithms

EXERCISES SECTION 10.3.2

Exercise 10.14: Write a lambda that takes two ints and returns their sum.

Exercise 10.15: Write a lambda that captures an int from its enclosing function and takes an int parameter. The lambda should return the sum of the captured int and the int parameter.

Exercise 10.16: Write your own version of the biggies function using lambdas.

Exercise 10.17: Rewrite exercise 10.12 from § 10.3.1 (p. 387) to use a lambda in the call to sort instead of the compareIsbn function.

Exercise 10.18: Rewrite biggies to use partition instead of find_if. We described the partition algorithm in exercise 10.13 in § 10.3.1 (p. 387).

Exercise 10.19: Rewrite the previous exercise to use stable_partition, which like stable_sort maintains the original element order in the paritioned sequence.

10.3.3 Lambda Captures and Returns

When we define a lambda, the compiler generates a new (unnamed) class type that corresponds to that lambda. We'll see how these classes are generated in § 14.8.1 (p. 572). For now, what's useful to understand is that when we pass a lambda to a function, we are defining both a new type and an object of that type: The argument is an unnamed object of this compiler-generated class type. Similarly, when we use auto to define a variable initialized by a lambda, we are defining an object of the type generated from that lambda.

By default, the class generated from a lambda contains a data member corresponding to the variables captured by the lambda. Like the data members of any class, the data members of a lambda are initialized when a lambda object is created.

Capture by Value

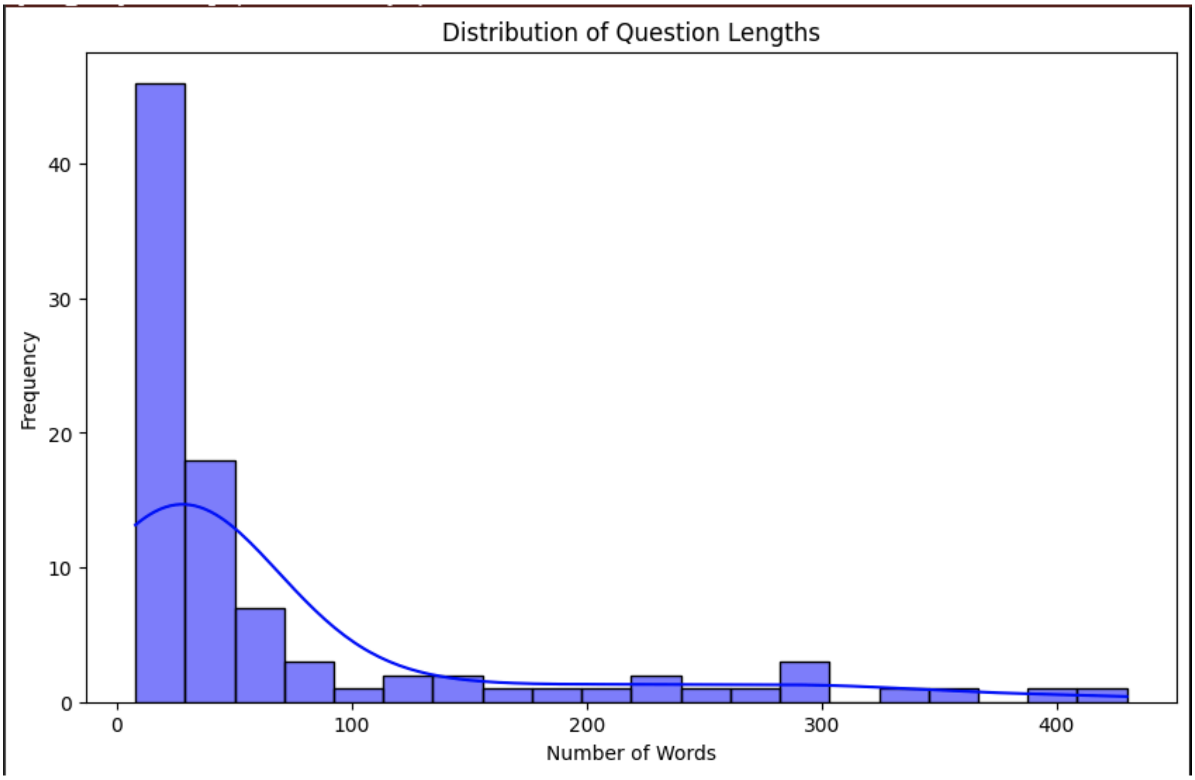
Similar to parameter passing, we can capture variables by value or by reference. Table 10.1 (p. 395) covers the various ways we can form a capture list. So far, our lambdas have captured variables by value. As with a parameter passed by value, it must be possible to copy such variables. Unlike parameters, the value of a captured variable is copied when the lambda is created, not when it is called:

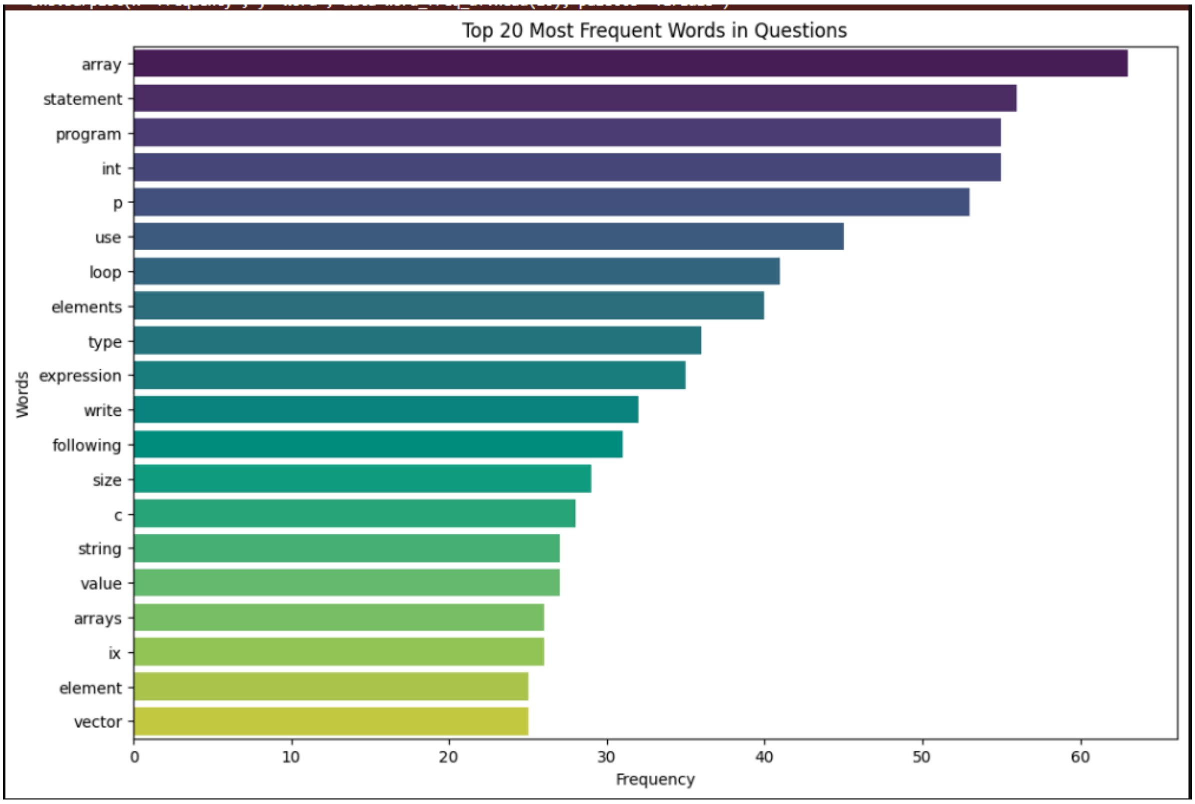
```
void fcn1()
{
    size_t v1 = 42;  // local variable
    // copies v1 into the callable object named f
    auto f = [v1] { return v1; };
    v1 = 0;
    auto j = f(); // j is 42; f stored a copy of v1 when we created it
}
```

Because the value is copied when the lambda is created, subsequent changes to a captured variable have no effect on the corresponding value inside the lambda.

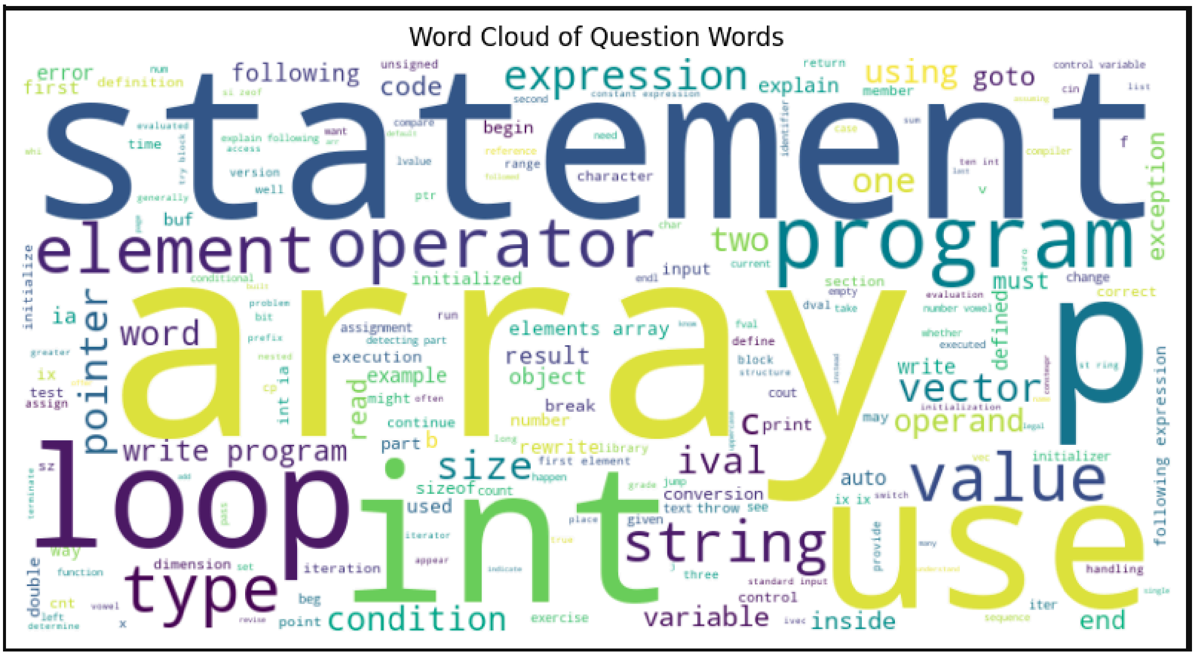
1.1 Image showing bounding boxes showing text extraction from the pages of the book

- **Seaborn Statistical Visualizations**
  - Generates complex statistical graphics
  - Analyzes question distribution across chapters





- **WordCloud Generation**
  - Creates visual representations of most frequent terms
  - Helps understand thematic content of extracted questions
  - Provides quick semantic overview of document contents



4.2. `BERT.ipynb`: (Comprehensive Model Development Process)

1.Model Initialization

- Loads pre-trained BERT base uncased model
- Configures custom classification layers
- Prepares for domain-specific fine-tuning

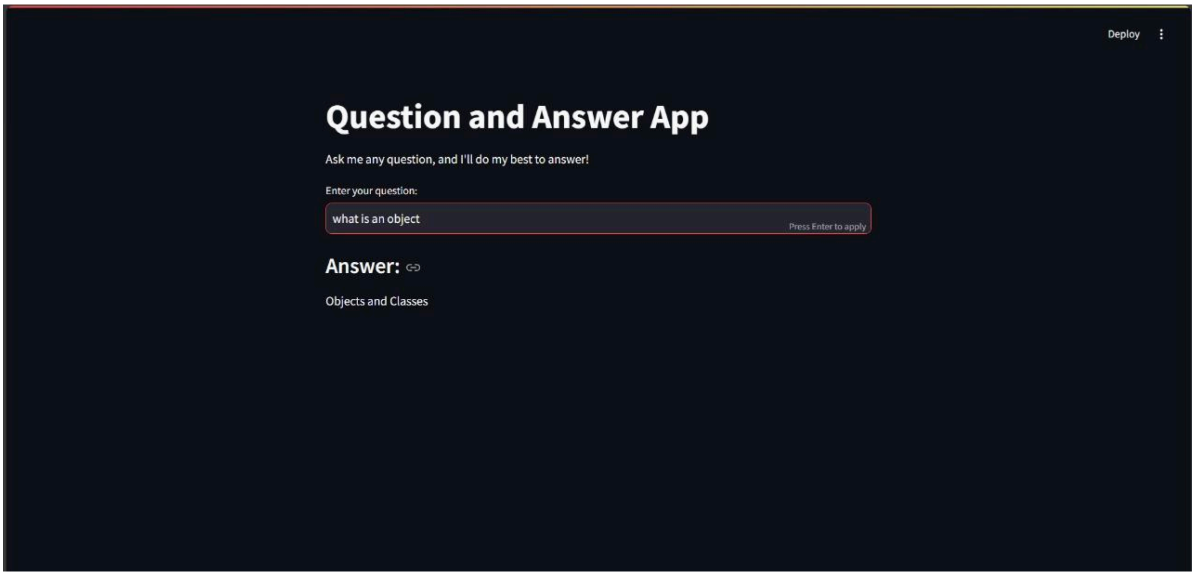
2.Training Strategy

- Implements advanced training techniques
- Uses cross-entropy-loss
- Applies learning rate scheduling
- Incorporates early stopping mechanisms

3.Model Saving

- Saves trained model to `model_complete.pth`
- Preserves entire model architecture
- Enables easy model deployment and reproduction

4.3. web-app.py:
(Deployment Architecture)



Question and Answer App

Ask me any question, and I'll do my best to answer!

Enter your question:

what is an object

Answer:

Objects and Classes

Interface Features

- Interactive question input
- Real-time classification
- Visualization of classification probabilities
- Error handling and user guidance

4. final_questions.csv: Structured Dataset

Dataset Characteristics

- Two-column structure:
    1. question_text: Extracted question content
    2. chapter_name: Corresponding book chapter

5. Limitations

While the analysis conducted by the Group Discussion Analyzer yields valuable insights, it is crucial to recognize the inherent limitations that may impact the interpretation of findings.

i. OCR Accuracy and Data Quality:

The accuracy of the extracted questions depends heavily on the quality of the OCR (Optical Character Recognition) process. Factors such as poor scan quality, faded text, or complex formatting in PDFs can result in misinterpretations or missing data. Despite preprocessing efforts, OCR errors may lead to incorrect question extraction, impacting the overall reliability of the dataset.

ii. Incomplete or Inconsistent Data:

The completeness of the dataset is critical for meaningful analysis. Missing pages, incomplete questions, or inconsistent formatting across different coursebooks and question papers can introduce gaps in the data. These omissions may result in an incomplete understanding of question trends and could skew the insights derived from the analysis.

iii. Contextual Limitations in Classification:

While the BERT model is effective for classifying questions based on predefined categories, it primarily focuses on the textual content. It may not fully capture the context, tone, or implied difficulty of a question. Complex or ambiguous questions might be misclassified, especially if the training data lacks sufficient diversity to cover all potential variations in question phrasing.

iv. Sample Size and Representativeness:

The analysis is contingent on the available dataset, which might not comprehensively represent all previous year question papers or coursebooks. Limited sample size or bias towards specific subjects or institutions could result in findings that are not generalizable across different educational contexts. Variability in question formats and content standards further complicates the ability to draw universal conclusions.

v. Dynamic Trends Over Time:

The educational curriculum and question patterns may evolve over time. Historical question data might not always reflect current or future trends. Without continuous updates and validation against recent question papers, the insights generated may become outdated, reducing their practical relevance.

6. Conclusion and Future Work

6.1. Conclusions

The **PYQ Analyser** has successfully demonstrated its capability to automate the extraction, classification, and analysis of previous year question (PYQ) papers, providing valuable insights into exam trends and question patterns. Key findings include:

i. Automated Question Extraction:
The OCR-based system effectively extracted questions from scanned coursebooks and question papers, converting them into structured datasets. Despite variations in text quality, the tool consistently identified and digitized relevant content, ensuring a comprehensive database for analysis.

ii. BERT-based Classification:
The integration of the BERT model for classifying questions by topic, chapter, and difficulty level provided accurate categorization. This classification facilitated a deeper understanding of the distribution and emphasis of different topics across various exam years.

iii. Trend Analysis and Visualization:
Temporal analysis revealed consistent patterns in question frequency and topic recurrence. Certain topics appeared more frequently in specific years or chapters, indicating areas of focus within the curriculum. The visualizations created using Seaborn and Matplotlib, such as heatmaps and bar charts, provided intuitive representations of these trends.

iv. Keyword Analysis and Themes:
Keyword frequency analysis highlighted common themes and topics prevalent in the question papers. The generated word clouds and keyword visualizations offered quick insights into the core subjects emphasized over time.

v. Difficulty Distribution Insights:
Classification of questions into difficulty levels showed that certain chapters or topics consistently presented more challenging questions. This information can help educators and students focus their efforts more strategically.

vi. Contextual Challenges:
While the BERT model provided strong performance in classification, it occasionally struggled with contextually nuanced questions, particularly those requiring deeper semantic understanding or domain-specific knowledge.

6.2. Future Works

i. Enhanced Contextual Understanding:
Future iterations of the **PYQ Analyser** could incorporate advanced NLP techniques such as semantic analysis and context-aware models. This would improve the system's ability to understand complex questions and differentiate between similar topics with subtle variations.

ii. Dynamic and Interactive Visualizations:
Expanding the visualization toolkit to include dynamic and interactive visualizations would enhance the user experience. Implementing dashboards with features like interactive timelines, topic heatmaps, and question trend graphs will facilitate more intuitive data exploration.

iii. Continuous Model Training:
To capture evolving exam patterns and curriculum changes, the BERT model should be continuously trained on newly extracted datasets. This would ensure the classification system remains up-to-date and relevant to current educational standards.

iv. Sentiment and Difficulty Analysis:
Incorporating sentiment analysis to determine the perceived difficulty of questions based on historical data and student performance reviews could add another layer of insight. This would help educators identify particularly challenging or misunderstood topics.

v. Enhanced Data Quality and Preprocessing:
Improving the OCR pipeline with additional preprocessing steps (e.g., noise reduction, text normalization) and leveraging more robust OCR models will enhance the accuracy of question extraction, especially for low-quality scans.

vi. Ethical Considerations and Data Privacy:
Future work should emphasize strict adherence to data privacy regulations, particularly when handling educational content from various institutions. Implementing encryption and anonymization techniques will ensure that sensitive information is protected.

vii. User Feedback Mechanisms:
Incorporating user feedback loops would help refine the tool's accuracy and usability. Educators and students could provide input on misclassified questions or suggest improvements, contributing to continuous system enhancement.

viii. Integration with Educational Platforms:
Integrating the **PYQ Analyser** with existing educational platforms or learning management systems (LMS) would provide seamless access to insights. This could include features like personalized learning paths based on historical question trends.

ix. Advanced Machine Learning Techniques:
Exploring advanced techniques such as topic modeling, question generation, and predictive analytics could offer deeper insights into emerging trends. Predictive models could forecast potential future exam questions based on historical data patterns.

x. Development of Proprietary AI Tools:
Developing customized AI tools tailored to educational data processing, similar to models like Firefly.ai, could enhance analysis speed and accuracy. These tools could be optimized for question pattern recognition and curriculum-specific analyses.

7. Bibliography and References

1. Pytorch Tutorial : https://pytorch.org/tutorials/beginner/basics/intro.html

2. PaddleOCR : https://github.com/PaddlePaddle/PaddleOCR

3. BERT LLM : https://en.wikipedia.org/wiki/BERT_(language_model)

4. Streamlit Resources : https://docs.streamlit.io/get-started

5. Pytorch Documentation : https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

6. BERT and Tensorflow tutorial : https://www.youtube.com/watch?v=7kLi8u2dJz0