Data Science Assignment: Invoice Data Extraction

Background

You are tasked with developing a solution to extract data from invoices. The invoices are in PDF format, which may include a mixture of regular PDFs, scanned documents, and PDFs containing both text and images.

Objective

Your goal is to implement the most cost-effective approach while maximizing accuracy in data extraction. Accuracy takes priority, with an expected accuracy rate of >90%.

Requirements

1. Data Extraction:
    - Develop a system to extract relevant information from invoice PDFs.
    - The system should handle various PDF types (regular, scanned, mixed text/image).
2. Accuracy Check and Trust Determination:
    - Implement an accuracy check logic for each input.
    - The system must provide an accuracy assessment for every extracted data point.
    - In 99% of cases, the system MUST be able to determine whether the extracted data can be trusted or not. This is a critical requirement.
3. Cost-Effectiveness vs. Accuracy:
    - Analyze and implement the most cost-effective approach while prioritizing accuracy.
    - If an approach is 20% more expensive but provides a 5% increase in accuracy, prefer the more accurate approach.
4. Performance Metrics:
    - Achieve an overall accuracy rate of >90%.
    - Provide a detailed breakdown of accuracy rates for different types of invoice data (e.g., invoice number, date, total amount, line items).
5. Scalability and Efficiency:
    - The solution should be scalable to handle large volumes of invoices.
    - Optimize for processing speed without compromising accuracy.
6. Error Handling and Reporting:
    - Implement robust error handling mechanisms.

- Provide clear reports on extraction failures, accuracy issues, and data reliability.

Deliverables

1. Source Code:
    - Well-documented code for the invoice data extraction system.
    - Include all necessary scripts, modules, and dependencies.
2. Technical Documentation:
    - Detailed explanation of the approach and algorithms used.
    - Justification for chosen methods, especially regarding the balance between cost-effectiveness and accuracy.
    - Specific explanation of the method used to achieve the 99% trust determination requirement.
3. Accuracy and Trust Assessment Report:
    - Comprehensive report on the accuracy of the system.
    - Detailed analysis of the system's ability to determine data trustworthiness in 99% of cases.
    - Breakdown of accuracy by invoice type and data field.
    - Explanation of the accuracy check and trust determination logic implemented.
4. Performance Analysis:
    - Analysis of system performance, including processing speed and resource utilization.
    - Comparison of different approaches tested, including cost-benefit analysis.

Evaluation Criteria

1. Accuracy of data extraction (40%)
2. Implementation of effective accuracy check and trust determination logic (30%)
3. Cost-effectiveness of the solution (15%)
4. Code quality and documentation (10%)
5. Scalability and performance optimization (5%)

Note to Candidates

You will be provided with a set of sample PDFs for testing your solution. These samples will include a variety of invoice types to ensure robust testing of your system. Your ability to handle different PDF formats, maintain high accuracy across all types, and especially your capability to determine data trustworthiness in 99% of cases will be crucial in the evaluation.

Good luck!