

Objective:

Participants will develop a program that can extract structured information from different types of documents related to a request for proposals (RFP) process. They will use Language Models to process and interpret the data.

Assignment Details:

1. Task Description:

- The participants will be provided with several documents, including HTML files and PDFs containing RFP details and addendums.
- The goal is to extract structured information similar to the one provided as following structure:

Fields	Value
Bid Number	
Title	
Due Date	
Bid Submission Type	
Term of Bid	
Pre Bid Meeting	
Installation	
Bid Bond Requirement	
Delivery Date	
Payment Terms	
Any Additional Documentation Required	
MFG for Registration	
Contract or Cooperative to use	
Model_no	
Part_no	
Product	
contact_info	
company_name	
Bid Summary	
Product Specification	

1. Documents Provided:

- HTML and PDF file with information on the Bid.

2. Steps to Complete:

- **File Parsing:** The participants must develop or use libraries to parse HTML and PDF documents.

- **Text Extraction:** Extract relevant text from different sections of the documents.
- **Information Structuring:** Use LLMs, RAG or NLP techniques to identify and structure the information in a predefined format.
- **Data Mapping:** Map the extracted information to the corresponding fields as shown in the example.

3. Requirements:

- The program should be able to handle different document formats (HTML and PDF).
- Ensure that the structured data extracted is accurate and maps correctly to the specified fields.
- Provide a method to output the structured data in a JSON format similar to the example provided.

4. Evaluation Criteria:

- **Accuracy:** How accurately the information is extracted and structured.
- **Robustness:** The ability of the program to handle various document structures and content.
- **Code Quality:** Clean, well-documented code with clear instructions for use.
- **Use of LLMs/NLP/RAG:** Effective use of Language Models, Retrieval Augmented Generation techniques or Natural Language Processing techniques in extracting and structuring data.

Deliverables:

- The Python script or notebook used to extract and structure the information.
- A README file with instructions on how to run the code and any dependencies required.
- A JSON file containing the structured information extracted from each provided document.