**Report on Developing the Voter Information Extraction Solution**

**Introduction**

The goal was to accurately capture details such as the voter's name, relative's name, relation type, and house number from extracted text. This report outlines the process I followed, including the challenges I encountered and how I addressed them.

**Process Overview**

1. **Initial Analysis**:
   * I started by examining the text data extracted from pdf. The text varied significantly in format, which posed a challenge for creating consistent extraction patterns.
   * Key elements to extract included names, relation types (like "Father" or "Husband"), and house numbers.
2. **Developing Initial Regex Patterns**:
   * I crafted regex patterns to identify and extract the relevant information. The initial patterns focused on:
     + **Names**: Capturing text following keywords such as "Name" or "Husband's Name."
     + **Relations**: Extracting specific relation types like "Father" or "Husband."
     + **House Numbers**: Extracting numeric sequences often labeled as "House No."
3. **Challenges and Solutions**:

* **Libraries not able to read pdf:** The normal libraries that are used generally, were not able to read any text from the given pdf.
* **Solution:** To overcome this I used ‘pdf2image’ library to pdf to images and then I used ‘pytesseract’ to extract text from those images.
  + **Inconsistent Formats**: The text data included various formats for names and relations. For example, some names were preceded by titles or additional descriptors.
    - **Solution**: I refined the regex patterns to handle these variations. By using optional groups and alternatives in the patterns, I improved their flexibility.
  + **Missing Data**: Some house numbers and relations were not captured due to varied formats.
    - **Solution**: I adjusted the patterns to be more inclusive, ensuring they could handle different ways house numbers were presented and different labels for relations.

1. **Refining Regex Patterns**:
   * **Names**: I modified the patterns to better accommodate different ways names could appear, including cases where names were not directly after keywords.
   * **Relations**: I focused the patterns to extract only "Father" and "Husband," ensuring that unrelated terms were excluded.
   * **House Numbers**: I improved patterns to reliably capture numeric sequences and handle various formatting issues.
2. **Testing and Validation**:
   * I tested the refined regex patterns with a variety of text samples to check their accuracy and consistency.
   * I verified that each pattern correctly extracted the required data and formatted it appropriately.
3. **Final Implementation**:
   * I integrated the final regex patterns into the extraction process, ensuring data was organized into separate columns for each piece of information: voter's name, relative's name, relation type, and house number.
   * I also implemented additional steps to clean and validate the extracted data.

**Conclusion**

Developing the voter information extraction solution involved analyzing diverse text formats, refining regex patterns, and testing thoroughly to ensure accuracy. Even though the final output had a not very good accuracy, this project enhanced my skills in text processing and pattern matching, which are valuable in various data extraction and analysis tasks.