Project Title: Real Estate Document Classifier & Splitter

Objective:

Your assignment is to build a solution that takes a **merged PDF file** which contains multiple real estate-related documents combined into one and intelligently **splits it into individual documents**, labeling each one with the correct document type.

For this task, you'll focus specifically on identifying and separating **Deeds**, **Property Cards**, and **Tax Documents**.

Download dataset：
https://drive.google.com/drive/folders/1kNfi06fmmPe1mlJ4ZbVvboDpfyAiiQNa

---

Step-by-Step Guidance:

1. **Objective**:
   The goal is for your program to analyze a merged PDF and:
   - Determine the **type** of each document (e.g., Deed, Property Card, or Tax Document).
   - Assign an appropriate **file name** or label to each split document.
2. **Start Simple**:
   Initially, limit your scope to just the three document types mentioned above. Once you've successfully handled these, you can think about expanding the program to accommodate other document types found in **Sample Full Search PDFs**.
3. **Use the Sample Dataset**:
   We've shared a labeled sample dataset with you. Use this to:
   - Learn how each document type typically appears.
   - Train your model or develop rules or patterns while you are working on the solution as per your requirement.

Key Document Characteristics:

Deeds:

- **Newer Deeds** often have clear headings such as:
    - *"Title to Real Estate"*
    - *"Quitclaim Deed"*
    - *"Special Warranty Deed"*
    - *"Warranty Deed"*
    - *"Indenture"*
    - *"Conveyance"*
- **Older Deeds** may not have headings but instead contain indicative phrases in the body of the text, such as:
    - *“do grant, bargain, sell, and release…”*
    - *“do grant and convey…”*
- **Deed Book and Page Numbers**:
  These are typically found at the **beginning or end** of the document and can help you label the document correctly.

Recording Pages / Stamps:

- The **last page** of each document often includes a *recording stamp* or *recording information* (e.g., “Recorded on [Date]”) which signals the end of that document.
- If the document doesn’t have a clearly marked stamp, look for phrases like *“recorded on [date]”*, which may be handwritten or typed.

Deliverables:

You are expected to submit the following:

1. **A Functional Solution** that can:
    - **Split** a merged PDF into individual documents based on detected boundaries.
    - **Classify and rename** each document accurately using a consistent naming format.
      Example:
        - Deed_1.pdf, Deed_2.pdf
        - Property-Card_1.pdf
        - Tax-Document_1.pdf
2. **A Brief Write-Up** explaining your approach (1-2 page document):
    - Describe how you identified where each document starts and ends.
    - Explain the features, keywords, or logic you used to determine the document types.
    - Mention any machine learning models, rules, libraries, or tools that were part of your solution.
3. **A Short Video Walkthrough** (not more than 10 minutes):
    - Record your screen while explaining what you built.
    - Walk through the code and demonstrate the program in action.
    - Explain your thought process and key decisions as you go.
4. Create a Repository in your Git-hub profile and share the Codebase with us.

This task is designed to assess both your practical skills in document classification, NLP, and data handling and your ability to communicate your approach effectively. Think of it as building a small-scale intelligent document processor—similar to what real-world companies use to automate legal or real estate workflows.