## PDF Content Extraction Projects

This repository contains Python projects for extracting content from PDFs and other sources.

### Project List

#### 1. **Basic PDF Reader**

- Read PDF file from folder

- Write content to output file

#### 2. Recursive PDF Reader

- Traverse subfolder tree

- Extract PDF content

#### 3. Single Page PDF Reader

- Take page number input

- Extract specific page content

#### 4. Regex Based PDF Reader

- Match content via regex

- Config driven extraction

#### 5. PDF to Database

- Store questions in database

- MySQL based storage

#### 6. Load Questions by Chapter

- Take chapter as input

- Load questions into console

#### 7. Multithreaded RSS Parser and web scraping

- Parse RSS feed

- Launch parallel requests

## Project Descriptions

### Basic PDF Reader

#### Requirements:

Store PDF file in `/content` folder

Read PDF content

Write to `output.txt`

#### Error Handling:

Folder/file availability

File permissions

### Recursive PDF Reader

#### Requirements

Nested subfolder structure

Extract PDFs from all subfolders

Write output per folder

#### Error Handling

- Folder/file availability

- File permissions

### Single Page PDF Reader

#### \*\*Requirements:\*\*

- Take page number input

- Extract specific page content

- Write to output file

#### \*\*Error Handling\*\*

- Invalid page number

- Usability focused

### ## Prerequisites

- Python 3.6+

- pip

- virtualenv

#### Running Projects

# Set up venv

virtualenv venv

source venv/bin/activate

pip install -r requirements.txt

# Run desired project

python project1.py

Overall, the goal is to provide modular PDF parsing projects for learning purposes with ample error handling.

Let me know if you have any suggestions for improving