Spark Pdf

Spark-Pdf is a library for processing documents using Apache Spark.

It includes the following features:

- Load PDF documents/Images
- Extract text from PDF documents/Images
- Extract images from PDF documents
- OCR Images/PDF documents
- Run NER on text extracted from PDF documents/Images
- Visualize NER results

Installation

Requirements

- Python 3.11
- Apache Spark 3.5 or higher
- Java 8
- Tesseract 5.0 or higher

```
pip install spark-pdf
```