Pipeline to extract contents of pdfs and docx to mkdocs .

Introduction:

This project is a pipeline to extract the text and images from any pdfs and docx file and parse the contents into mkdocs static website.

This project uses Java's libraries like Apache POI and Apache PdfBox to do this task. Apache POI is used for pdfs and Apache PdfBox is for word documents or docx file.

Project Flow:

Step 1: This project takes all the text from the pdf and docx and parse in **index.md** of mkdocs till the image is encountered.

Step 2: As soon as the image is encountered the image is first saved in a folder named **imagesforpdf** for pdf 's image and **imagesfordocx** for docx's images.

Step 3: Then the hyperlink of the image is inserted in index.md in the location where the image is encountered in the pdf.

Step 4: This cycle repeats till all the text and images are parsed in the index.md of mkdocs.

Tools Used:

Language Used: Java.

Libraries Used: Apache POI and Apache PdfBox.

IDE used: IntelliJ

To Install all dependencies:

Add jar files dependencies in IntelliJ.

Steps to add dependencies in IntelliJ.

1.Files->Project Structure->Modules->Dependencies->Below Module Sdk there is a + icon

-> add Jar and Directories-> Add the jar file.

Download Jar files through this link:
https://drive.google.com/drive/folders/1khb4VsPifhXlToNLwldWPfVbBVmGsjjE?usp=drive_link

Or

Download: poi-bin-5.2.3 folder from internet.

RUN:

Please paste the relevent paths of the pdfs and docx file. And also the path of index.md.

Press Run icon or run from terminal like normal Java file run.