Document Processing using Amazon Textract OCR

Client: William Muir, Paqt.

PDF Reader

So, basically, the PDF reader is a Java Spring framework-based application.

The main objective of this application is to detect the text being written in a PDF file.

So, for the first step which is reading the text from any PDF file, our application basically uses the Amazon Textract OCR APIs.

The text reading part from any PDF is done in such a way that our application segregates the block and words being mentioned in the PDF file.

Eventually, when the reading and segregating part is done from the PDF file next step comes into action and that is to reframe the output being obtained into HTML format.

Once everything is done from reading PDF to converting the content into HTML format, the final step comes and that is to store the document in the S3 bucket, which is done using the credentials and environment variables class.

SRC

Documentcontroller.java

Class Description:

In this class, we integrate the API into our front and back end.

Also, in this class, we have called textractservice.java and S3service.java.

Imports:

In this particular class we did the following imports and that are java.io*, orgspringframework.beans, orgspringframework.web.

Functions:

a) Get dock: The main task of this function is to extract the document from the S3 bucket.
b) Single File Upload: The task of this function is to upload the document in the S3 bucket.

c) Analyze Dock: The task of this function is to analyze the text of the given PDF file and then convert it into the HTML format.

S3 service.java:

Class Description:

The main task of this class is to build a connection between the application and the S3 bucket for fetching and uploading documents to S3 bucket.

Imports:

In this particular class, we did the following imports and java.util*, java.io*, javax.xml* and software.amazon.awssdk.

Functions:

a) Get client: This function is assigned the task of creating an S3 client object using the credentials and region of the S3 bucket.
b) Get object bytes: The task of this particular function is to fetch the document from the S3 bucket in form of bytes.
c) List all objects: This function is needed to return or we can say show all the available files in the S3 bucket.
d) Put object: So basically, this function is needed to put the PDF format document in the S3 bucket.
e) To XML: As the name says this function is responsible for converting the document into XML format so that it can be passed back to view.

Textract Service.java:

Class Description:

This class basically helps in connecting to Amazon textract using the textract client objects.

This class performs various operations to convert PDF into HTML format.

Imports:

In this particular class we did the following imports and are java.util*, java.io*, javax.xml* and software.amazon.awssdk, javax.xml.

Function:

a) Analyze Document: This function is helps in picking up the document from S3 bucket and converts it into byte format and eventually return the text.
b) Convert to string: As the name suggest this function simply converts the output of any format into string.
c) Get text: This function is useful for displaying the converted text.

So now to sum up the whole project, as per the task assigned this project was able to read the PDF file and convert it into the HTML format and eventually display the output text on screen.