

Code Instructions for Digitization and Data Frames for Card Index Records

Before following the below steps: create an AWS account, upload the RFC application files to the S3 bucket, and authenticate AWS's Boto3.

1. In `AWS_Code.ipynb`, input the document, AWS, and location names. The `docuName` refers to the `Cards_34.pdf` file that is in the S3 bucket (`s3Name`).

```
#initial document details
s3Name = #s3BucketName
docuName = #document name
saved_folder = #Folder to put textract outputs
box_file_name = #Box Name
csv_file_name = #Name of CSV output file
```

2. Compile the code blocks and run the `convert_card_image()` function. The function will print “finished” once Textract has finished running. The text file for each card will then be uploaded to the S3 bucket and downloaded locally (in the same directory as the code file).
3. Before running AWS Comprehend, the two custom models need to be trained. Training is as follows:
 - a. Launch AWS Comprehend and go to custom entity recognition.
 - b. Click create new model (BankLocation Model)
 - i. For custom entity type, enter Bank and Location.
 - ii. In data specifications, select “CSV file” and “Using annotations and training docs” for the training type.
 - iii. Input the training CSV and text files from the Training>BankLocation subfolder.
 - iv. Select your IAM role and click create.
 - c. Click create new model (Decision Model)
 - i. For custom entity type, enter Decision.
 - ii. In data specifications, select “CSV file” and “Using entity type and training docs” for the training type.
 - iii. Input the training CSV and text files from the Training>Decision subfolder.
 - iv. Select your IAM role and click create.
4. Once training is completed, input the requested information that appears at the beginning of the Comprehend block of code in `AWS_Code.ipynb`.
 - a. Enter the ARN values for each custom model under “decision_ARN” and “bankloc_ARN”.
 - b. Enter the dARN values. The `general_dARN`, `decision_dARN`, and `bankloc_dARN` refer to the data access role for the models. Generally, all of these values are the same. To obtain these values, go to AWS IAM > Roles > Your preferred Role.
 - c. Enter the input folder for the Comprehend model. This should be where the Textract text files are located.Also, enter a file path for the output of the Comprehend files.

```
#general comprehend
general_dARN = "arn:aws:iam::12312132:role/test-role"
general_s3_input = "s3://input-path"
general_s3_output = "s3://output-path"

#decision comprehend
decision_dARN = "arn:aws:iam::12312132:role/test-role"
decision_ARN = "arn:aws:comprehend:us-west-2:12312312:testmodel"
decision_s3_input = "s3://input-path"
decision_s3_output = "s3://output-path"

#bankloc comprehend
bankloc_dARN = "arn:aws:iam::12312132:role/test-role"
bankloc_ARN = "arn:aws:comprehend:us-west-2:12312312:testmodel"
bankloc_s3_input = "s3://input-path"
bankloc_s3_output = "s3://output-path"
```

5. Compile all of the Comprehend code blocks and run the `start_all_comprehend()` function.
6. Go to the S3 bucket and download the Comprehend output files. Rename each of the files to the following: “General.txt” for General Comprehend output; “BankLoc.txt” for BankLocation Custom Comprehend output; “Decision.txt” for Decision Custom Comprehend output. Ensure that these files are in the same directory as the code file.
7. Compile the remaining code blocks and run `get_csv_test()`. Once executed, the final CSV output file will appear in the same directory as the code file.