Document Analyzer

Aim

To create an application which can extract data from any documents and store them.

Technologies

Amazon Web Services, Docker, Python/C#, Machine Learning/OCR.

Features

- Amazon Textract uses Optical Character Recognition (OCR) technology to automatically detect printed text, handwriting, and numbers in a scan or rendering of a document, such as a legal document or a scan of a book.
- Form Extraction: Amazon Textract enables you to detect key-value pairs in document images automatically so that you can retain the inherent context of the document without any manual intervention.
- Table Extraction: Amazon Textract preserves the composition of data stored in tables during extraction. This is helpful for documents that are largely composed of structured data, such as financial reports or medical records that have column names in the top row of the table followed by rows of individual entries.
- Handwriting Recognition: Many documents such as medical intake forms or employment applications contain both handwritten and printed text. Amazon Textract can extract printed text and handwriting from documents written in English with high confidence scores, whether it is free-form text or text embedded in tables and forms. Documents can also contain a mix of typed text or handwritten text.
- Document Text Detection: It can help to detect all the words and hence sentences from the document. This can be use to further process or get the information from the form.

Use Cases

- Financial Services (Mortgage apps, invoices, receipts)
- Healthcare and Life Sciences (Healthcare and Insurance forms)
- Public Sector (loans, federal tax forms)

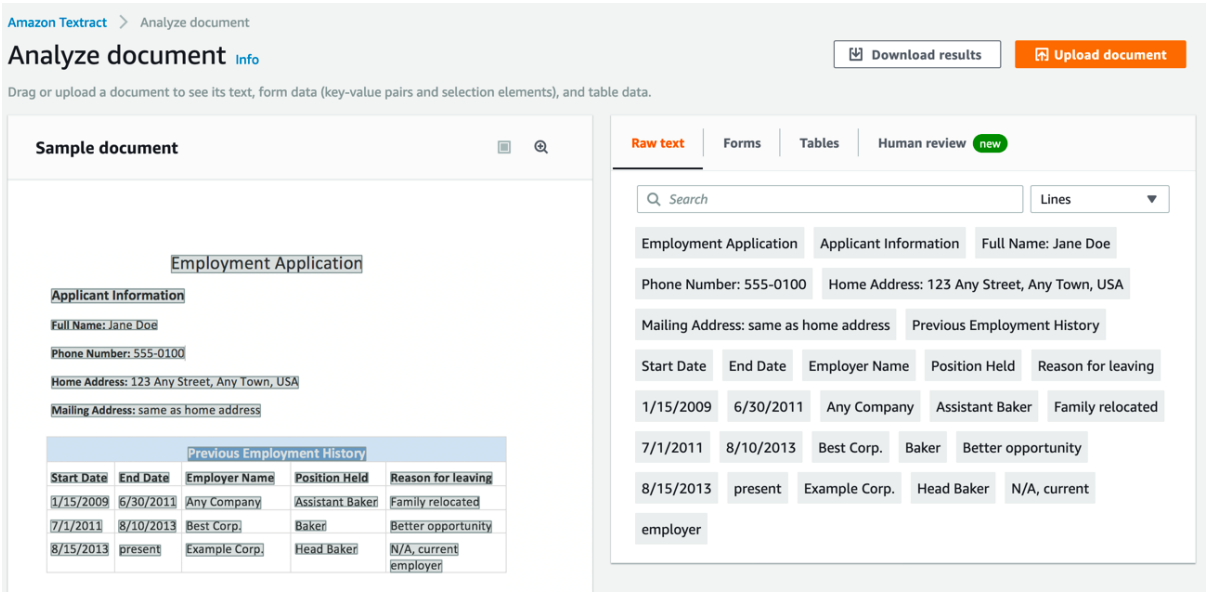
Pros

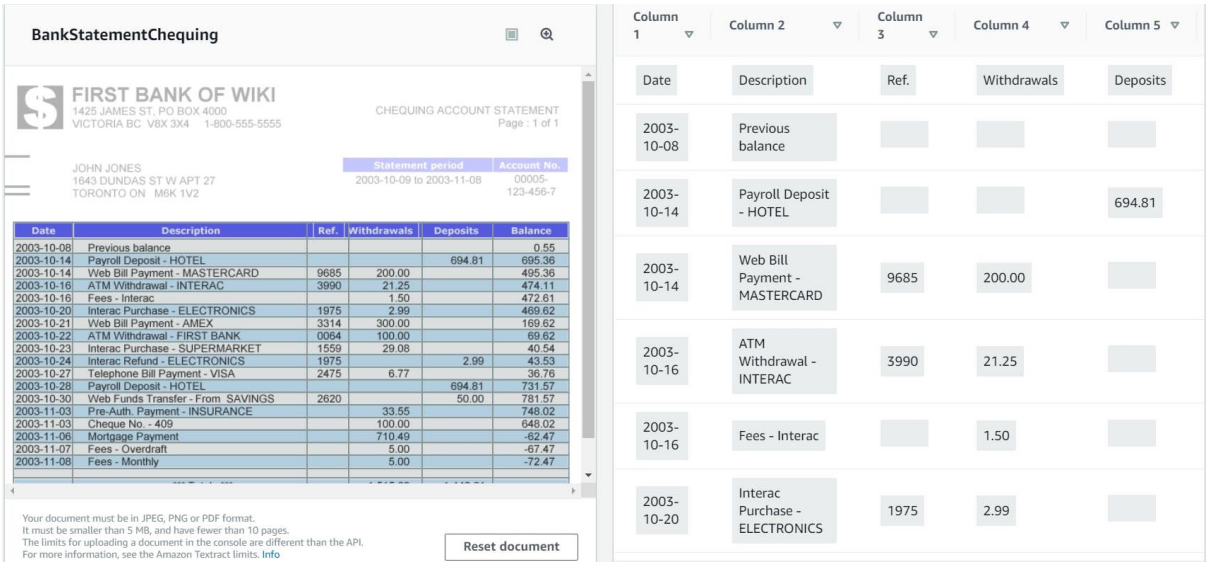
- It’s fast, reliable and efficient.
- It can recognize any sort of handwritten document/forms.
- It provides confidence scores.
- Easy to use.

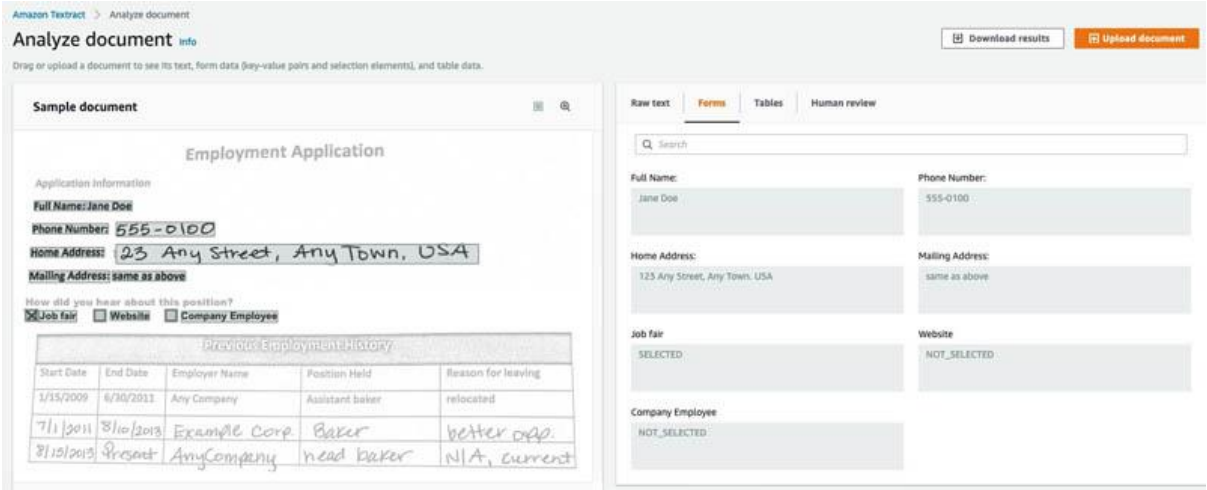
Cons

- Requires internet connectivity.

Samples


Amazon Textract
Analyze document
Analyze document Info
Download results
Upload document
Drag or upload a document to see its text, form data (key-value pairs and selection elements), and table data.
Sample document
Employment Application
Applicant Information
Full Name: Jane Doe
Phone Number: 555-0100
Home Address: 123 Any Street, Any Town, USA
Mailing Address: same as home address
Previous Employment History
Start Date | End Date | Employer Name | Position Held | Reason for leaving
1/15/2009 | 6/30/2011 | Any Company | Assistant Baker | Family relocated
7/1/2011 | 8/10/2013 | Best Corp. | Baker | Better opportunity
8/15/2013 | present | Example Corp. | Head Baker | N/A, current employer
Raw text
Forms
Tables
Human review
new
Search
Lines
Employment Application
Applicant Information
Full Name: Jane Doe
Phone Number: 555-0100
Home Address: 123 Any Street, Any Town, USA
Mailing Address: same as home address
Previous Employment History
Start Date
End Date
Employer Name
Position Held
Reason for leaving
1/15/2009
6/30/2011
Any Company
Assistant Baker
Family relocated
7/1/2011
8/10/2013
Best Corp.
Baker
Better opportunity
8/15/2013
present
Example Corp.
Head Baker
N/A, current
employer


BankStatementChequing
FIRST BANK OF WIKI
1425 JAMES ST, PO BOX 4000
VICTORIA BC V8X 3X4 1-800-555-5555
CHEQUING ACCOUNT STATEMENT
Page : 1 of 1
JOHN JONES
1643 DUNDAS ST W APT 27
TORONTO ON M6K 1V2
Statement period
2003-10-09 to 2003-11-08
Account No.
00005-123-456-7
Date | Description | Ref. | Withdrawals | Deposits | Balance
2003-10-08 | Previous balance | | | | 0.55
2003-10-14 | Payroll Deposit - HOTEL | | | 694.81 | 695.36
2003-10-14 | Web Bill Payment - MASTERCARD | 9685 | 200.00 | | 495.36
2003-10-16 | ATM Withdrawal - INTERAC | 3990 | 21.25 | | 474.11
2003-10-16 | Fees - Interac | | 1.50 | | 472.61
2003-10-20 | Interac Purchase - ELECTRONICS | 1975 | 2.99 | | 469.62
2003-10-21 | Web Bill Payment - AMEX | 3314 | 300.00 | | 169.62
2003-10-22 | ATM Withdrawal - FIRST BANK | 0064 | 100.00 | | 69.62
2003-10-23 | Interac Purchase - SUPERMARKET | 1559 | 29.08 | | 40.54
2003-10-24 | Interac Refund - ELECTRONICS | 1975 | | 2.99 | 43.53
2003-10-27 | Telephone Bill Payment - VISA | 2475 | 6.77 | | 36.76
2003-10-28 | Payroll Deposit - HOTEL | | | 694.81 | 731.57
2003-10-30 | Web Funds Transfer - From SAVINGS | 2620 | | 50.00 | 781.57
2003-11-03 | Pre-Auth. Payment - INSURANCE | | 33.55 | | 748.02
2003-11-03 | Cheque No. - 409 | | 100.00 | | 648.02
2003-11-06 | Mortgage Payment | | 710.49 | | -62.47
2003-11-07 | Fees - Overdraft | | 5.00 | | -67.47
2003-11-08 | Fees - Monthly | | 5.00 | | -72.47
Your document must be in JPEG, PNG or PDF format.
It must be smaller than 5 MB, and have fewer than 10 pages.
The limits for uploading a document in the console are different than the API.
For more information, see the Amazon Textract limits. Info
Reset document
Column 1
Column 2
Column 3
Column 4
Column 5
Date
Description
Ref.
Withdrawals
Deposits
2003-10-08
Previous balance
2003-10-14
Payroll Deposit - HOTEL
694.81
2003-10-14
Web Bill Payment - MASTERCARD
9685
200.00
2003-10-16
ATM Withdrawal - INTERAC
3990
21.25
2003-10-16
Fees - Interac
1.50
2003-10-20
Interac Purchase - ELECTRONICS
1975
2.99


Amazon Textract
Analyze document
Analyze document Info
Download results
Upload document
Drag or upload a document to see its text, form data (key-value pairs and selection elements), and table data.
Sample document
Employment Application
Application Information
Full Name: Jane Doe
Phone Number: 555-0100
Home Address: 123 Any Street, Any Town, USA
Mailing Address: same as above
How did you hear about this position?
Job fair
Website
Company Employee
Previous Employment History
Start Date | End Date | Employer Name | Position Held | Reason for leaving
1/15/2009 | 6/30/2011 | Any Company | Assistant baker | relocated
7/1/2011 | 8/10/2013 | Example Corp | Baker | better opp.
8/15/2013 | present | AnyCompany | Head baker | N/A, current
Raw text
Forms
Tables
Human review
Search
Full Name:
Jane Doe
Phone Number:
555-0100
Home Address:
123 Any Street, Any Town, USA
Mailing Address:
same as above
Job fair
SELECTED
Website
NOT_SELECTED
Company Employee
NOT_SELECTED