Data Acquisition Specialist

Transform Documents into Actionable Data in Seconds using Veryfi OCR API

OCR APIs and Mobile SDKs to securely capture, extract, categorize and transform bills, invoices, receipts (SKUs), W2s, into standardized JSON with Level 3 data giving your app and customers superpowers.

https://www.veryfi.com/

Job Description

The ability to correctly identify patterns is becoming a specialized, but essential skill when dealing with large amounts of data. Diverse fields from cyber threat to genomics rely on accurately identifying, acquiring and, at times, correcting a very specific pattern among a sea of potential targets.

Veryfi is seeking an individual who is willing to develop expertise in **regular expressions** for processing texts from around the world. While fluency in multiple languages is not a requirement, the candidate should be able to effortlessly handle and manipulate diverse languages, formats, and symbols. The ideal candidate will possess the ability to create and modify patterns into classes, categories, and rules. A working knowledge of **Python** is necessary.

Qualifications:

While a degree in computer science is not mandatory, we do require candidates to possess a degree or relevant experience in technical or related fields such as bioinformatics data preprocessing, quantitative linguistics, or other areas that involve extracting data from large datasets or logs.

Bonus:

- Proficiency in Python, especially for task automation.
- Familiarity with distance functions and their relationship to error correction.
- Experience with regular expressions in right-to-left languages such as Hebrew and Arabic.
- Exposure to cloud environments, such as AWS or GCP.
- Background in web development.

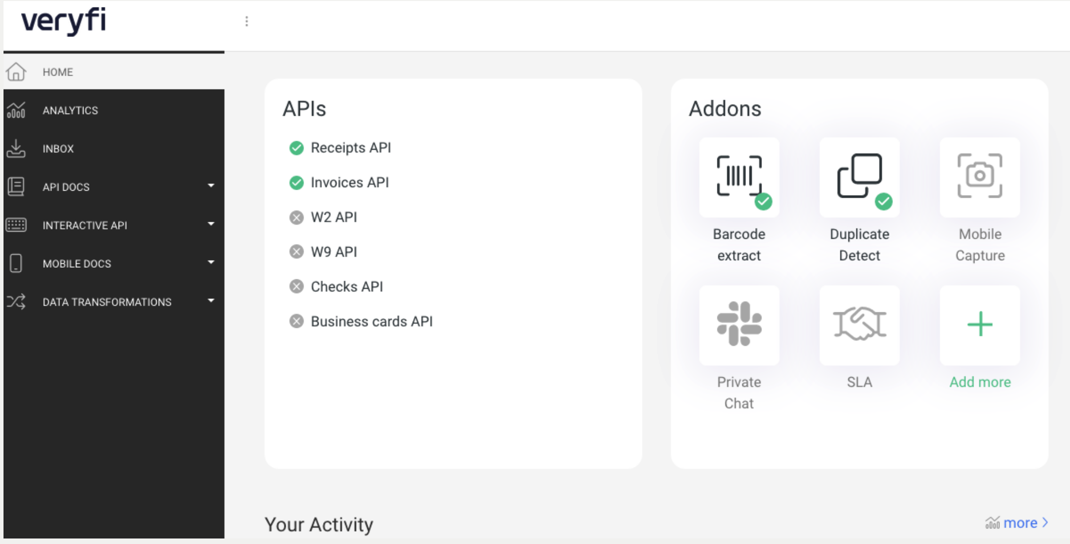
Technical test

We need you to complete an important step in the training of supervised machine learning models: labeling information. We have provided you with a new set of documents (attached), and your task is to accurately extract the following information in a JSON format:

- vendor_name
- bill_to_name
- bill_to_address
- ship_to_name
- ship_to_address
- line_items [Array]
  - quantity
  - Description
  - Price

If you need clarification about the fields requested, please visit: https://faq.veryfi.com/en/articles/5571268-document-data-extraction-fields-explained

1. Create an `OCR API` Veryfi account at https://hub.veryfi.com/signup/

> ⚠️ Please note that creating a Receipts OCR and Expenses App account will not allow you to use our APIs. To ensure access to our APIs, please make sure that the left pane in your hub account is gray. If the pane is green or pink, then you have created the wrong account type.
>
> 
>
>
> APIs
>
> ✓ Receipts API
> ✓ Invoices API
> ⊗ W2 API
> ⊗ W9 API
> ⊗ Checks API
> ⊗ Business cards API

> 💡 Be aware that the free-trial account comes with certain restrictions. Please send us the email address you used to create your account, and we will increase your account's limitations accordingly. This will enable you to test everything you need without any roadblocks.

2. Create a python-based system to extract the required information:
   a. Use the Veryfi's python API to get the OCR output for each document.

      https://github.com/veryfi/veryfi-python

      - The OCR output should be under `ocr_text` within the api response. You can ignore everything else.

   b. Develop an automatic solution that can extract the necessary information from the provided documents in a JSON format. Please keep in mind that the solution should support any document with the same format, while excluding any other documents. You can test the exclusion by processing a document of your own.
3. Include in your solution a file describing in detail your approach, your assumptions, and the coding best practices that you implemented.
   - Code paradigm
   - Unit tests
   - etc…
4. Send the link to the repository of your solution to jmgarzonv@veryfi.com

References

Sample documents were acquired from: https://www.kaggle.com/datasets/holtskinner/invoices-document-ai