# Form Parsing Using Document AI

cloudskillsboost.google/games/2854/labs/17207

## GSP827


## Overview

As consumers, we are used to filling out forms to apply for insurance, make insurance claims, specify healthcare preferences, apply for employment, tax withholdings, etc. Businesses on the other side of these transactions get a form that they need to parse, extract specific pieces of data from, and populate a database with.

In this lab you will use Google Cloud's Document AI solution to parse forms within a Jupyter Notebook so that you can automatically extract information from digitally scanned paper forms.

This lab you learn how to:

- Create a Jupyter Notebook instance on Cloud Vertex AI
- Create a Service Account so that you can automate form processing.
- Upload a PDF document to Cloud Storage.
- Invoke Document AI.
- Parse the response using low-level functions based on the visual layout of the form.
- Parse the response using high-level functions based on the semantic structure of the form.

### What you will build

You will parse a use it to parse a campaign disclosure form that all US political campaigns are required to file. From this form, you will pull out the cash that the campaign has on hand.

## Setup

### Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab**, shows how long Google Cloud resources will be made available to you.

This hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

To complete this lab, you need:

Access to a standard internet browser (Chrome browser recommended).

**Note:** Use an Incognito or private browser window to run this lab. This prevents any conflicts between your personal account and the Student account, which may cause extra charges incurred to your personal account.
Time to complete the lab---remember, once you start, you cannot pause a lab.

**Note:** If you already have your own personal Google Cloud account or project, do not use it for this lab to avoid extra charges to your account.
Be sure to read all of the instructions here and in the Jupyter Notebook.

## Deploy code to Vertex AI Notebooks

Open this link in a new tab to deploy the code to Vertex AI Notebooks. Follow the tutorial on the right to complete the required prerequsites.

On the Vertex AI page, change the **Instance name** to "formdemo".

In the **Notebook properties** section, click the **edit** (pencil) icon, then scroll down to **Machine configuration** and change the Machine type to **n1-standard-2**.

Click the **CREATE** button.

Wait for the Jupyter Notebook instance to get created and the notebook to be deployed. This will take about 5 minutes.

The notebook should open, but if not, click the **OPEN** link to open the notebook. Then click **Confirm** to download the notebook source.

### Examine document

In the JupyterLab tab, click on **Edit** > **Clear All Outputs** to clear all the cells in the notebook.

Run the first 5 cells in the notebook.

- Click into a cell - the blue bar to the left shows you which cell you're working in.
- Press the triangular **run selected cells...** button in the top ribbon or press **Shift** + **Enter** for each cell.

These cells download a file named `scott_walker.pdf` and then use the IPython display to display the PDF file.

Click on the PDF file and scroll through it to answer the following two questions:

What is the name of the committee that this form pertains to?

Which page number has the cash that this committee has on hand?

## Upload the file to Cloud Storage

Go back to the Cloud Console tab. In the Cloud Console, from the **Navigation menu**, go to **Cloud Storage** > **Browser**.

Click on the **CREATE BUCKET** button.

Change the BUCKET name to be the name of the lab Project ID (e.g. qwiklabs-gcp-03-00a52b77620f)

Click **Create**.

Go back to the Notebook tab. In the next section called Upload to Cloud Storage, change the BUCKET variable to the name of your bucket.

Run all 3 of the cells in this section to copy the PDF file to the bucket.

## Enable Document AI

Open this link in a new tab to go to the Cloud Document AI page in the Library.

Click the **Enable** button.

In the Jupyter notebook, run the cell in the Enable Document AI section to find out the active account email address. This is who the notebook user is running as. **Copy** this email address.

You can either click this link or click the link in the Jupyter notebook to create a service account that will perform the automated form processing.

In the **Service account details**, give the service account the name "formdemo".

Click the **CREATE AND CONTINUE** button.

Click into the **Select a role** field and start typing "document ai" in the search field, then select the role **Document AI API User**.

Click the **Continue** button.

In the box "Service account users role", add the active account's email address (copied from the notebook).

Click the **Done** button.

## Call Document AI

In the first Call Document AI cell, change the PDF to point to the file in your bucket. Copy the bucket location output from `cell #7` and add it in to `cell #10`.

Run the cells in this section.

**Note:** if you see an error, wait a minute and try running the cells again.
From the left menu, click on the the `response.json` file to open it in the file browser of the notebook.

What are the first words of text extracted?

## Parse the response

Run the cells to parse the response in Python. Does the result match?

## Parse form based on visual layout (Option 1)

Read the instructions for this section in the Jupyter Notebook.

Run the cells in this section to parse the form based on the visual layout of the form. Make sure to read the code and the commentary.

Change the code appropriately to obtain the "TOTAL RECEIPTS THIS PERIOD".

What are the total receipts this period?

## Parse form based on form semantics (Option 2)

Read the instructions for this section in the Jupyter Notebook.

Run the cells in this section to parse the form based on the form semantics. Make sure to read the code and the commentary.

Change the code appropriately to obtain the "DEBTS AND OBLIGATIONS OWED BY THE COMMITTEE".

**Hint:** this is field #9 on the same page; or you can loop through all pages and form fields looking for one whose name starts with the above string
What are the debts and obligations owed?

## Congratulations!

You have learned how to use Cloud Document AI to parse forms from within a Jupyter Notebook.

The Jupyter Notebook is optional -- you can run this code from any Python environment including serverless environments like Cloud Functions.

## Next steps

- Read blog post on which this lab is based.
- Document AI Solution

## Google Cloud Training & Certification

...helps you make the most of Google Cloud technologies. Our classes include technical skills and best practices to help you get up to speed quickly and continue your learning journey. We offer fundamental to advanced level training, with on-demand, live, and virtual options to suit your busy schedule. Certifications help you validate and prove your skill and expertise in Google Cloud technologies.

Manual Last Updated: June 22, 2022

Lab Last Tested: June 22, 2022