



⚠ More than 1 year has passed since last update.

 @YosukeArai13 updated at 2020-09-09

# Batch processing forms with Azure Form Recognizer

🔖 Python, Azure, AI, Cognitive

There is a service called Azure Form recognizer.

<https://azure.microsoft.com/ja-jp/services/cognitive-services/form-recognizer/>

It is an excellent one that reads the form nicely and extracts the target data. Since there is also an API, I wrote a Python script that can process multiple forms at once

<https://github.com/yosukearaiMS13/formrecognizerbatch/blob/master/fy.py>

The contents of the script and how to use it are explained below.

## The contents of the script

The script is made by extending the sample in the document

<https://docs.microsoft.com/ja-jp/azure/cognitive-services/form-recognizer/quickstarts/python-labeled-data?tabs=v2-0>

The script consists of 4 sections

<https://github.com/yosukearaiMS13/formrecognizerbatch/blob/master/fy.py>

fr.py

```
# Configurations: 各種設定パラメータ

# Post 分析対象pdf section
## Form recognizerに対し、分析対象データを一旦全部postします

# Get analyze results section
## 先ほどpostしたデータの分析結果（抽出されたデータ含む）を取得します。

# 抽出結果のcsv出力 section
## 抽出結果を出力します。余計な空白の除去と、信頼性が低い抽出値の置き換え
## （しきい値以下の場合抽出値は採用せず、代わりに信頼度を[]囲みで出力）
## を行っています
```

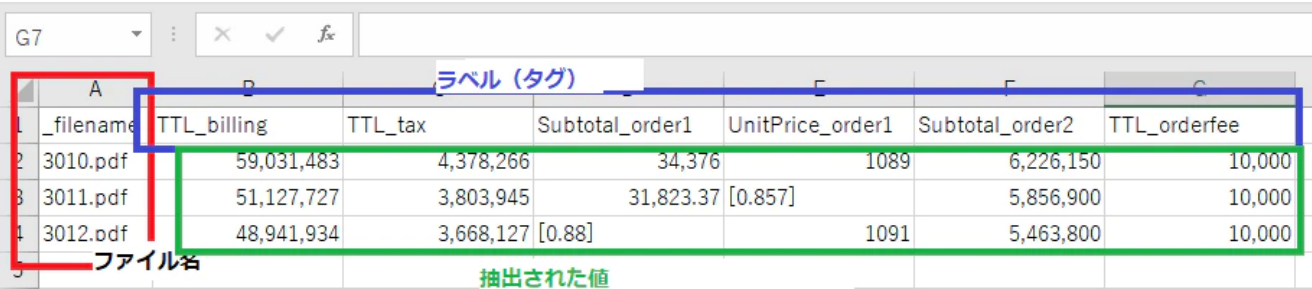
Get analyze results and csv output section of the extraction results parse the json returned by the Form recognizer. The format of json is here

[https://github.com/Azure-Samples/cognitive-services-REST-api-samples/blob/master/curl/form-recognizer/Invoice\\_1.pdf.ocr.json](https://github.com/Azure-Samples/cognitive-services-REST-api-samples/blob/master/curl/form-recognizer/Invoice_1.pdf.ocr.json)

The format of the output csv is as follows.

--First column: Form file name to be analyzed

--Second and subsequent columns: All labels (tags) set in the analysis model and the corresponding extracted values



	A	B	C	D	E	F	G
1	_filename	TTL_billing	TTL_tax	Subtotal_order1	UnitPrice_order1	Subtotal_order2	TTL_orderfee
2	3010.pdf	59,031,483	4,378,266	34,376	1089	6,226,150	10,000
3	3011.pdf	51,127,727	3,803,945	31,823.37	[0.857]	5,856,900	10,000
4	3012.pdf	48,941,934	3,668,127	[0.88]	1091	5,463,800	10,000
5	ファイル名	抽出された値					

The API used in each section is as follows-

Post analysis target pdf: [Analyze Form](#)

--Get analyze results: [Get Analyze Form Result](#)

--CSV output section: [Get Custom Model](#) --Get

all the labels defined in the API and get all the labels. I am using it as the value of the csv header

# How to use the script

## 1. Environment

Win10 Enterprise, Python 3.8.5, IDE is optional

## 2. Data extraction preparation

(\* From the prerequisite work to data extraction preparation 1, [this Qiita article](#) will also be helpful)

- Prerequisites: Do the following first
  - [Creating a Form Recognizer resource](#)
  - [Create Azure blob \(create storage account-> create container\)](#)
  - [Azure blob settings \(creating a Shared Access Signature\)](#) (useful using the Storage Explorer menu in the Azure portal)

- # Shared Access Signature
- アクセス ポリシー: (なし) ▼

開始時刻: 2020/09/08 16:36 🗓

有効期限: 2020/10/31 16:36 🗓

タイム ゾーン:

☒ ローカル

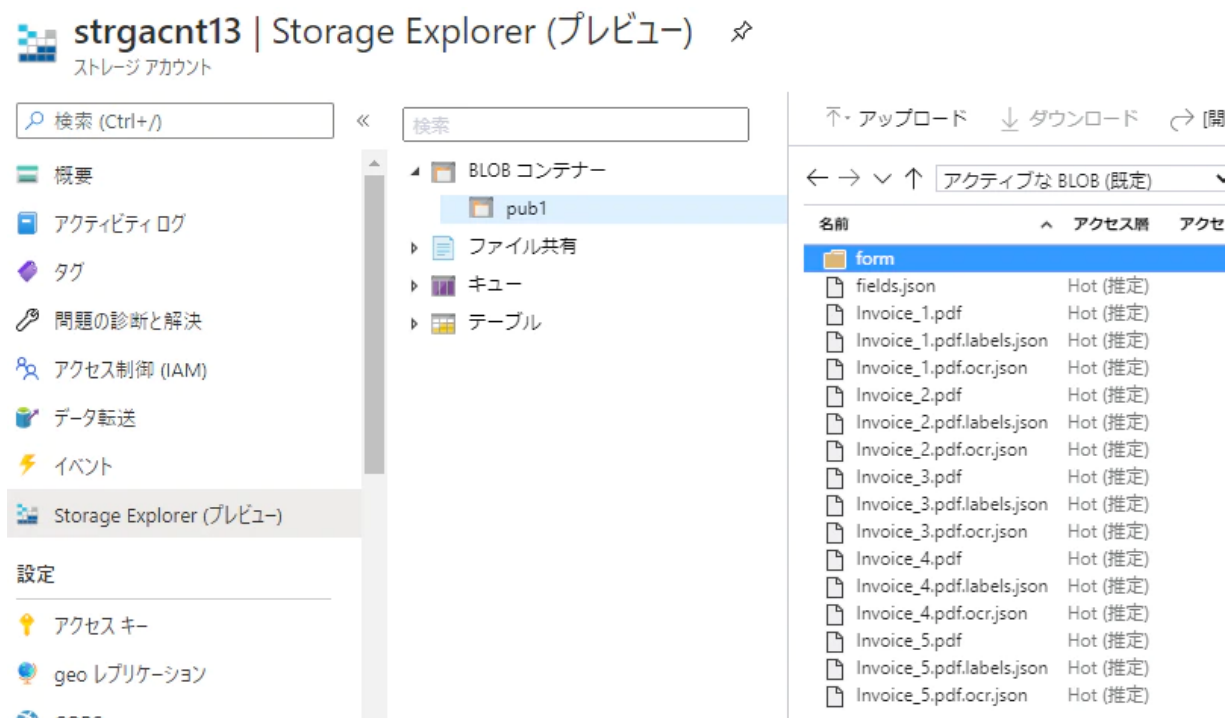
☐ UTC

アクセス許可:

  - ☒ 読み取り
  - ☒ 追加
  - ☒ 作成
  - ☒ 書き込み

- [illegible]

- Data extraction preparation 1 (implemented only for the first time)
  - Store training data for model creation in Azure blob: Place at least 5 files (invoice\_1 ~ 5.pdf in this case) in the following form (xx.json is a file created later, so ignore it here)



- 
- Labeling tool settings:
  - Click here for labeling tools <https://fott.azurewebsites.net/>
  - Labeling procedure: Follow the steps in the document below to connect to the labeling tool-> create a project.
    - <https://docs.microsoft.com/ja-jp/azure/cognitive-services/form-recognizer/quickstarts/label-tool?tabs=v2-0#connect-to-the-sample-labeling-tool>
- Settings for python script # 1

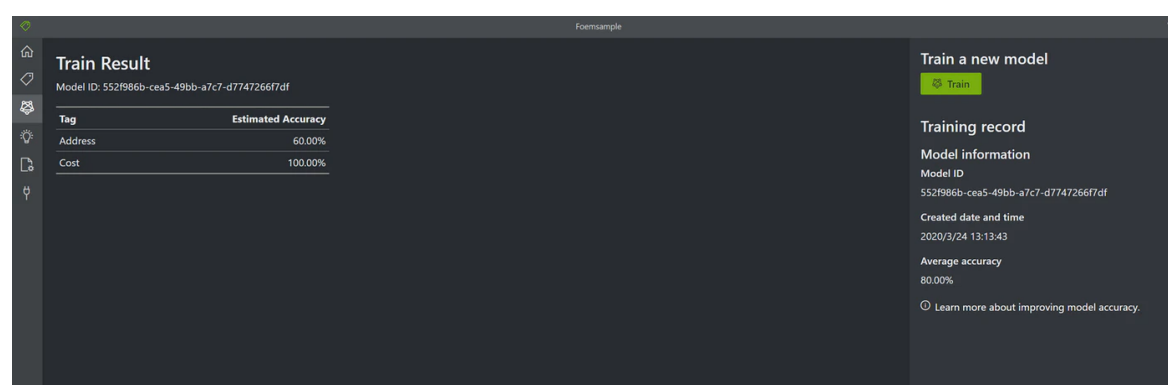
fr.py

```
## Configurations
endpoint = r"https://xxxxx.cognitiveservices.azure.com/"
apim_key = "xxxxx"
model_id = "xxxxx"
sourceDir = r"C:\xxxxx\"
confidence_setting = 0.9 # 0~1. 信頼性がこの値以下の場合採用しない
```

--endpoint: Endpoint of Form Recognizer --apim\_key: Key of Form Recognizer 1 or 2 --sourceDir: Describe the location of the form file to be analyzed with the full path --confidence\_setting: Set the value from 0 to 1 (\* As a script specification, If the reliability is less than or equal to this value, the extracted value is not adopted, and instead the reliability evaluation value is output in []).



- Data extraction preparation 2 (implemented every time a label is added or modified)
  - Label the read training data (form) with the label (tagging) tool ( <https://fott.azurewebsites.net/> ). Train when you're done and generate a model
    - Steps in the documentation below: Label the form-> Train your custom model, go on
      - <https://docs.microsoft.com/ja-jp/azure/cognitive-services/form-recognizer/quickstarts/label-tool?tabs=v2-0#label-your-forms>
    - You can also refer to the procedure in [this Qiita article](#) .
    - A Model ID will be generated after the train (below). This value will be used later



- Settings for python script # 2: model\_id

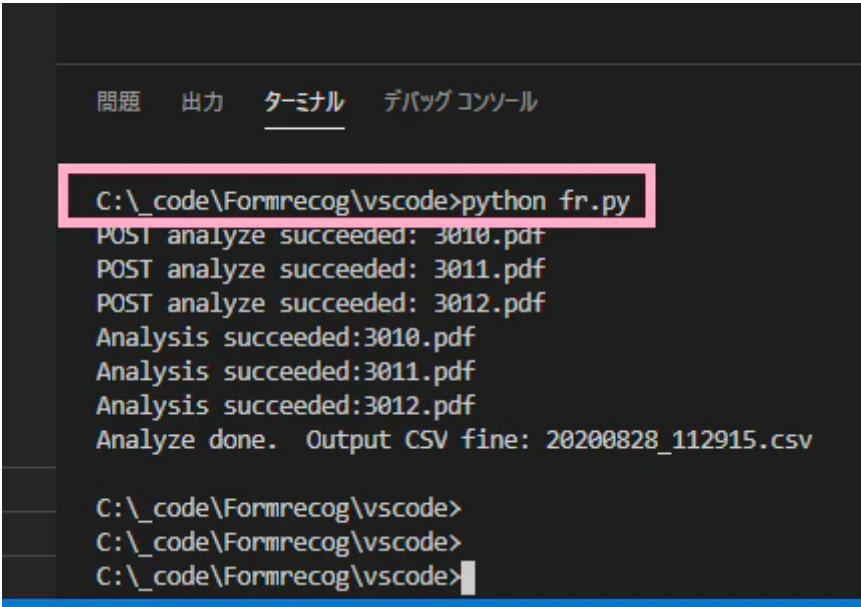
fr.py

```
## Configurations
endpoint = r"https://xxxxx.cognitiveservices.azure.com/"
apim_key = "xxxxx"
model_id = "xxxxx"
sourceDir = r"C:\xxxxx\"
confidence_setting = 0.9 # 0~1. 信頼性がこの値以下の場合採用しない
```

--Model\_id: Set the Model ID obtained above

### 3. Data extraction

- Place the form file to be analyzed in sourceDir
- Run fr.py
- Data extraction result csv is output to the same folder as the script



The screenshot shows a terminal window with a dark background. At the top, there are tabs labeled '問題', '出力', 'ターミナル', and 'デバッグ コンソール'. The 'ターミナル' tab is active. The command prompt shows the path 'C:\\_code\Formrecog\vscode' followed by the command 'python fr.py'. The output of the script is as follows:

```
C:\_code\Formrecog\vscode>python fr.py
POST analyze succeeded: 3010.pdf
POST analyze succeeded: 3011.pdf
POST analyze succeeded: 3012.pdf
Analysis succeeded:3010.pdf
Analysis succeeded:3011.pdf
Analysis succeeded:3012.pdf
Analyze done. Output CSV file: 20200828_112915.csv

C:\_code\Formrecog\vscode>
C:\_code\Formrecog\vscode>
C:\_code\Formrecog\vscode>
```

•

## 4. Constraints, etc.

- It is a file format of training and analysis target forms, but I have only tried PDF
- It is based on the current version v2.0 of Form recognizer. If you use it in other versions, you will need to change the API URL as appropriate and respond to the json format change returned by Form recognizer.



Why not register and get more from Qiita?

1. We will deliver articles that match you

By following users and tags, you can catch up information on technical fields that you are interested in as a whole

2. you can read useful information later efficiently

By "stocking" the articles you like, you can search right away

➡ What you can do with signing up

Sign up

Login



**Yosuke Arai**  
@ YosukeArai13

Follow



Comments Comments

No comments

Sign up for free and join this conversation.

Sign Up

If you already have a Qiita account [Login](#)

Being held events



Qiita 10th Anniversary Event-Technologies you want to study now for 10 years  
2021/09/09~2021/10/01

View Event Details



Post an article about Azure IoT!  
2021/09/10~2021/10/10

View Event Details

➡ All

How developers code is here.



Qiita

- About
- Terms
- Privacy
- Guideline
- Design Guideline
- Release
- API
- Opinion
- Help
- Advertisement

Increments

- About
- Employment information
- Blog
- Qiita Team
- Qiita Jobs
- Qiita Zine