Extracción automática de contenidos en documentos electrónicos

Daniel Prieto Bargados
Trabajo fin de Grado – Ingeniería Informática
Cátedra TotalEnergies de Analítica de Datos e Inteligencia Artificial

Introducción

Totalenergies cuenta con un software de procesamiento de facturas llamado eSave a través del cual consigue extraer de forma automática diferentes datos, que son empleados para proporcionar al cliente una oferta con reducciones de gasto tanto en luz como en gas.

Este software analiza documentos electrónicos o fotografías de los mismos, extrayendo el texto y otras informaciones sin intervención de un operador.

Objetivos del estudio

El objetivo final de la empresa es el desarrollo de su propio software de procesamiento de facturas, obteniendo datos de las facturas de los clientes y proporcionándoles a estos una oferta que les favorezca más que otras compañías.

Métodos

Se ha comenzado a desarrollar un código que genera imágenes de las facturas y obtiene datos como el CIF de la compañía o el NIF del cliente, así como la obtención de parámetros que van asociados al QR que podemos encontrarnos en una factura de luz.

```
#Obtengo datos de la empresa
cif_empresa.append(obtencion_datos_factura(imgs))
#Obtengo los datos del cliente
datos_clientes(cif_empresa,imgs[0])
#Obtengo la URL del qr
url_codificada = lector_qr(x)
#Obtengo los parametros de la URL
dic = obtencionDatos(url_codificada)
#Genero el directorio en el que se van a guardar los parametros obtenidos del qr
prueba_path_json = directorio_json_datos(cif_empresa)
#Exporto los datos obtenidos del QR a un archivo de extensión .json
datosJson(dic,nombres[nom],prueba_path_json)
nom = nom + 1
#Elimino los archivos temporales
eliminarArchivosTemp(temp_files_path)
```

Figuras y Resultados

```
def obtencionDatos(URLS):
    code_string = []
    parsed_url = []
    captured_value = []
    dic = {}

    for i in range(len(URLS)):
        code_string.append(URLS[i][0].decode('UTF-8'))
        parsed_url.append(urlparse(code_string[i]))
        captured_value.append(parse_qs(parsed_url[i].query))

    for i in range(len(captured_value)):
        for clave, valor in zip(captured_value[i].keys(), captured_value[i].values()):
            dic[clave] = valor[0]
    return dic
```

Los resultados obtenidos durante la ejecución se dividen en dos partes:

- resultados obtenidos a partir de extraer los parámetros del código QR (que se almacenan en un archivo json)
- parámetros que tiene que incluir este QR, definidos por la CNMC.

Del código QR se extrae la siguiente información:

• Datos del consumidor::
o Código postal.
o Potencia contratada.
o Consumo por periodos horarios de cada peaje de transporte y distribución en el último año.

• Posibles penalizaciones: Con este dato conseguimos saber si un cliente tiene
penalización debido a una cancelación anticipada del contrato.

• Parámetros relacionados con el periodo de facturación:
o Importe total.
o Importe de servicios adicionales.
o Importe de excedentes de consumo.

• Otros datos como, por ejemplo:
o CUPS.
o Potencia máxima demandada.

Conclusiones

El prototipo desarrollado tiene unos tiempos de ejecución demasiado elevados como para que sea un producto que pueda ser publicado de cara al cliente.

Estos tiempos de ejecución son tan elevados debido a que hay que proporcionar un DPI elevado con el objetivo de mejorar la precisión del OCR a la hora de obtener el texto de las facturas, aunque esto solo afecta a la parte de obtener datos analizando la factura; los tiempos de ejecución obtenidos para la obtención de los datos a partir de QR se encuentran en torno a los 10 segundos por factura, dependiendo de cuántas imágenes se generen.

Trabajo futuro

Se han presentado algunas posibles soluciones para mejorar el tiempo de respuesta, como aplicar el OCR solo a unas partes en concreto, utilizar programación multihilo lanzando varios hilos de ejecución para que cada hilo procese una factura o utilizar una IA.

En el futuro se emplearán técnicas inteligentes de análisis de documentos, de forma que se pueda deducir de forma rápida la estructura de la factura y dirigir el OCR a partes concretas de la imagen.

TotalEnergies

Universidad de Oviedo