

Challenge - Text Mining con Facturas

Detalles de la entrega

Este trabajo corresponde al 60% de la evaluación de la asignatura. Para que el trabajo se considere completo, deberá contener al menos las siguientes partes:

- Un script de Python ejecutable sin errores.
- Un informe breve explicando el trabajo realizado en Python. Este informe se puede presentar en el formato que os parezca más adecuado, no necesariamente como un documento de Word; por ejemplo se podría entregar un Jupyter-Notebook en el que se explique lo realizado.

Para la evaluación, se considerarán los siguientes criterios:

- Que el código se ejecute sin errores.
- Que el código esté limpio y comentado.
- Que el código sea capaz de cumplir **al menos** los objetivos mínimos que se piden en el trabajo (identificar los datos, clasificarlos, exportarlos a Excel y generar una o más gráficas resumen).
- Que el informe sea claro.

Se valorarán positivamente en la nota otras funcionalidades que queráis darle al script, como por ejemplo que hagáis una aplicación a partir del script. No obstante, recordad que al final tiene que haber un script de Python que cumpla con los criterios que hemos pedido y que podamos ejecutar.

El trabajo se subirá a la actividad de ADI con los grupos previamente formados.


El plazo máximo para la entrega será el día 15 de mayo.

Contexto

Estáis trabajando como programadores de Python en una empresa de software que hace reconocimiento automático de texto en imágenes, también llamado OCR (Optical Character Recognition). La empresa se llama ClasiFact, y su producto principal consiste en una aplicación para analizar y clasificar automáticamente fotografías de facturas (**Figura 1**).

ATENCIÓN NUEVAS CONDICIONES DE PAGO ENVIAR FACTURAS ANTES DEL 30 Y PAGAN EN SABADELL POR TRANSFERENCIA A 60 DIAS FECHA FACTURA.

MULTIMAC,C.B
Apdo Correos 20104
06080 - Badajoz
N.I.F: E-06227490



FACTURA 2017 12

SU PEDIDO 36572
SU PEDIDO

FECHA: 16-ene-17
CLIENTE: Vegenat, S.A
DIRECCION: Ctra. Badajoz - Montijo, Km 24.9
06184 - Pueblo Nuevo del Guadiana
N.I.F: A-06141345
acacho@vegenat.es

CANTIDAD	CONCEPTO	UNIDAD	TOTAL
	Nuestro Albarán N°: 20170091		
78 Uds	Botellones de agua garrafas	6,00 €	468,00 €
2 Uds	Palet	10,00 €	20,00 €
4 Uds	Tableros	10,00 €	40,00 €
78 Uds	Botellones de agua producto (10%iva)	4,06 €	316,68 €
11 uds	Fuentes de agua en deposito sin cargo: Lumesn0008386, ebac15011982 Romen0803000001, ebac158000570, ebac136005753, ebac158000546 ebac148007726, inox 131200186, ebac170006792, ebac149009367 nobo368, 0712001554Romen Valor en caso de perdida, robo o siniestro por mal uso	200€7ud	
0 Uds	Vasos de 220 cc para fuente	0,02 €	
0 Uds	Palet de garrafas de 8 litros no retornables 96uds (10%iva) lot080715	1,62 €	- €
78 UDS	BOTELLONES RETIRADOS	- 6,00 €	-468,00 €
2 UDS	PALET	- 10,00 €	- 20,00 €
4468 UDS	TABLEROS	- 10,00 €	- 40,00 €
	SUMA:		316,68 €
	I.V.A: 10%		31,67 €
	I.V.A: 21%		- €
	TOTAL:		348,35 €

Según nuevas condiciones 60 dias fecha/factura abono en:
Sabadell Atlantico. ES24 0081 7850 3200 0112 9413
Un saludo vencimiento 16/3/2016

Ci. Juan Labrado nº3 2º A. CP 06008 Badajoz

Figure 1. Ejemplo de factura. DISCLAIMER: Las facturas han sido obtenidas de internet. Se ocultan o modifican los datos confidenciales.

La empresa está desarrollando una aplicación con la que el usuario puede sacar fotos a las facturas que desea, y la aplicación reconoce, clasifica y analiza los datos que aparecen en las facturas, de forma automática.

Un equipo diferente al vuestro ha desarrollado el sistema de OCR que realiza la primera parte del trabajo: reconocer todo el texto que aparece en las facturas (**Figura 2**).

ATENCIÓN NUEVOS CONDICIONES DE PAGO ENVIAR FACTURAS ANTES DEL 30 Y PAGAN EN SABADELL POR TRANSFERENCIA A 60 DIAS FECHA FACTURA.

MULTIMAC,C.B
Apdo Correos 20104
06080 - Badajoz
N.I.F: E-06227490

FECHA: 16-ene-17
CLIENTE: Vegenat, S.A
DIRECCIÓN: Ctra. Badajoz - Montijo, Km 24.9
06184 - Pueblo Nuevo del Guadiana

CANTIDAD	CONCEPTO	UNIDAD	TOTAL
78 Uds	Botellones de agua garrafas	6,00 €	468,00 €
2 Uds	Palet	10,00 €	20,00 €
4 Uds	Tableros	10,00 €	40,00 €
78 Uds	Botellones de agua producto (10%iva)	4,06 €	316,68 €
11 uds	Fuentes de agua en deposito sin cargo. Lumen0008386, ebac15011982 Romen0803000001, ebac158000570, ebac136005753, ebac158000546 ebac148007726, inox 131200186, ebac170006792, ebac148009367 nobo368, 0712001554Romen Valor en caso de perdida, robo o siniestro por mal uso	200€7ud	
0 Uds	Vasos de 220 cc para fuente	0,02 €	
0 Uds	Palet de garrafas de 8 litros no retornables 96uds (10%iva) lot080715	1,62 €	- €
78 UDS	BOTELLONES RETIRADOS	- 6,00 €	-468,00 €
2 UDS	PALET	- 10,00 €	- 20,00 €
4468 UDS	TABLEROS	- 10,00 €	- 40,00 €
SUMA:		316,68 €	
I.V.A: 10%		31,67 €	
I.V.A: 21%		- €	
TOTAL:		348,35 €	

Según nuevas condiciones 60 días fecha/factura abono en:
Sabadell Atlantico. ES24 0081 7850 3200 0112 9413
Un saludo vencimiento 16/3/2016

Code + Text

```
print(next(value_iterator))
```

IVALINEAS2 (1)

ATENCIÓN NUEVOS CONDICIONES DE PAGO ENVIAR FACTURAS ANTES DEL 30 Y PAGAN EN SABADELL POR TRANSFERENCIA A 60 DIAS FECHA FACTURA

MULTIMAC,C.B

FACTURA
2017 12

Apdo Correos 20104
06080-Badajoz
N.I.F: E-06227490

SU PEDIDO
36572

SU PEDIDO
FECHA:
16-ene-17

CLIENTE:
Vegenat, S.A.

DIRECCIÓN:
Ctra. Badajoz - Montijo, Km 24.9

N.I.F:
A-06141345

06184 - Pueblo Nuevo del Guadiana
acacho@vegenat.es

CANTIDAD CONCEPTO

UNIDAD

TOTAL

Nuestro Albaran N°:
20170091

78 Uds

SUMA:
316,68 €

I.V.A: 10%
31,67 €

I.V.A: 21%
-

€

TOTAL:
348,35 €

Figure 2. Salida OCR (derecha) junto con factura original (izquierda).

Problema

El algoritmo que ha desarrollado el equipo de OCR devuelve como output un diccionario de Python en el que cada uno de sus elementos contiene todo el texto que ha reconocido en cada una de las facturas.

El archivo podéis descárgalo de ADI.



Figure 3. Archivo que devuelve el OCR de las facturas. Para nosotros será el input de nuestro programa.

El equipo del que formas parte es el encargado de:

1. Identificar y clasificar los distintos elementos de las facturas (NIF, importe total, IVA, fecha, etc). Ver Figura 4.
2. Exportar un archivo Excel con la información.
3. Realizar gráficas que muestren un resumen de las facturas (la cantidad de gráficas y los resúmenes de valores a representar quedan a vuestro criterio).
4. Cualquier extra que se quiera añadir a la aplicación también será valorado.



Factura

Nº factura 4043136967
 Nº cliente 1275465485
 NIF 01921327F
 Fecha 05.09.2019
 Página 1/1

CEAQ/14557/2*2/127CVB
 TOMAS GARCÍA GOMEZ
 TALLERES SAN ROMAN
 HOCES 9
 47011 VALLADOLID-VALLADOLID

Su interlocutor de la red de ventas

ALVARO GONZALEZ RODRIGUEZ
 /14557 T 616831043
 E 014557@ventas.wurth.es

NºArtículo	Pos.	Cantidad	Precio en	CP	Dto %	IVA %	Valor neto
Descripción artículo	Pos. Alb.		EUR				EUR
Nº referencia cliente							
Pedido 2083899179 del 05.09.2019							
Albarán 8318330947 del 05.09.2019							
Dirección de envío							
TALLERES SAN ROMAN							
HOCES 9							
47040 VALLADOLID-VALLADOLID							
0615915018061	1	1	32,20	1	PE	21	32,20
HOJA-SABLE-ZEBRA-BI-METAL-150X1,8-5UDS							
0615815030061	1	2	35,50	1		21	35,50
HOJA SABLE BI-METAL PARA MADERA 5UDS							
20							

Los envases que no llevan Punto Verde se consideran de uso industrial. El poseedor final es el responsable de la entrega del envase usado a los gestores competentes. (Art. 18.1 del Reglamento para el desarrollo y la ejecución de la Ley 11/97 de Envases y Residuos de Envases).

Nº RII RAEE 28 Nº RB-RPA 523

Vencimiento: 05.10.2019

Condición de pago: Giro bancario

CTA XXXXXX1341 BANCO BILBAO VIZCAYA ARGENTARIA

Portes EUR	Valor neto EUR	IVA	Impte. IVA EUR	Importe total EUR
0,00	67,70	21,00 %	14,22	81,92

A	B	E	F	H	J	L	AC	AH	AM	BG
FILENAME	REGISTRO	FECHA REGISTRO	FECHA EXPEDICIÓN	NOMBRE	NIF	NUMERO DE FACTURA	BASE	%IVA	IVA CANT	TOTAL
Wurth.pdf	21032310-01	2021-03-23	2019-09-05	WURTH	01921327F A08472276	4043136967	67,70	21	14,22	81,92
XXXX	21032310-02	2021-03-23	2020-12-04	Empresa 2	73457601M	01/2021	291,74	21	61,26	353
XXXX	21032310-03	2021-03-23	2020-12-05	Empresa 3	84557602M	01/2022	101,65	21	21,35	123
XXXX	21032310-04	2021-03-23	2020-12-06	Empresa 4	76657603M	01/2023	102,48	21	21,52	124
XXXX	21032310-05	2021-03-23	2020-12-07	Empresa 5	81257604M	01/2024	103,31	21	21,69	125

Figure 4. Output del algoritmo FRE.xls.

Identificar y clasificar los distintos elementos de las facturas (NIF, importe total, IVA, fecha, etc).

Se pide detectar mediante el uso de patrones y expresiones regulares el máximo número de campos para el máximo número de facturas. El programa debe exportar un archivo Excel llamado "FRE.xls" con tantas filas como facturas con todos los campos que se detecten (Ver Figura 4).

El código debe estar programado de forma general, es decir, debe funcionar con cualquier archivo dict.json, sin importar el número de facturas que contenga.

Se valora la programación funcional, es decir, que las funciones del programa se definan mediante la sentencia "def" de Python.

Hay que tener en cuenta que los distintos parámetros de las facturas pueden tener distintos patrones, algunos ejemplos:

- Número de factura (no siempre tienen): Cuando tienen, suelen tener la siguiente forma y van precedida de la palabra "factura", "fac.", "fra.", etc.. Ej:

FACTURA	2017 12
---------	---------

- CIF / NIF: el NIF de una empresa puede componerse de 1 LETRA + 8 NÚMEROS ó 1 LETRA + 10 NÚMEROS. Además, puede tener distintas formas como **XXXXXXXX-L**, **XXXXXXXXL**, **XXXXXXXXL**, **L-XXXXXXXX**, **LXXXXXXXX**, **LXXXXXXXX**.
- Fechas: pueden tener también varios formatos. Una factura puede tener varias fechas. Algunos ejemplos (para ver todas las opciones ver fotos de facturas):
 - 14/3/2012
 - 14/3/12
 - 14/03/2012
 - 14-03-2012
 - 14 de marzo de 2012
 - 14-mar-2012
 - 14-mar-12
 - 1 Abril 2020
- Importe total: los importes pueden tener separadores de miles (que pueden ser puntos o comas). Los decimales también pueden ser puntos o comas.
- IVA: En España puede ser del 21%, 10% y 4%.

Notas:

- Si para un campo se encuentran varios elementos (ej: NIF, o Fecha) deben de indicarse todos los elementos encontrados separados por una barra vertical "|".
- El archivo Excel a exportar debe tener algunas columnas vacías para que pueda ser leído por el programa (Ver Figura 4).
- Para la resolución del trabajo puede utilizarse cualquier librería de Python, aunque no se haya visto en clase.