

Objetivo y plan

Armar un pipeline donde ingresen invoices crudos en PDF y una lista de invoice numbers que aparecen en esos PDFs. Generar un archivo CSV con el invoice number de cada una junto con la suma de sus líneas profesionales, o lo más cerca posible.

Como muchos de los pasos intermedios son pesados en cómputo, intento separar cada etapa y guardar en disco los resultados. De esta manera, cada etapa levanta los datos de la anterior y no es necesario volver a ejecutarla.

Además utilizo memorización a nivel función para evitar otros cálculos intermedios (por ejemplo el fuzzy matching).

Cabe aclarar que el orden en las que las expongo no es el orden en el cual fueron desarrolladas, fue un proceso iterativo.

Pipeline

Extracción

El primer paso es tomar los archivos PDFs y extraer los metadatos, las imágenes y el texto. Esto genera un archivo JSON que contiene, para cada muestra:

- La ruta del PDF
- El tamaño del PDF
- La cantidad de hojas
- El texto (nativo del PDF)
- Las rutas a las imágenes. Se guardan como PNG todas las imágenes que tenga embebido el PDF.

Por ejemplo:

```
{  
  "filename": "10726698.pdf",  
  "size": 1879573,  
  "pages": 2,  
  "text": "",  
  "images": [  
    "<recortado>/ParaiSUR/data-extracted/10726698.pdf/0.png",  
    "<recortado>/ParaiSUR/data-extracted/10726698.pdf/1.png"  
  ]  
}
```

Esto demoró aproximadamente 6 horas de cómputo.

Detalles adicionales:

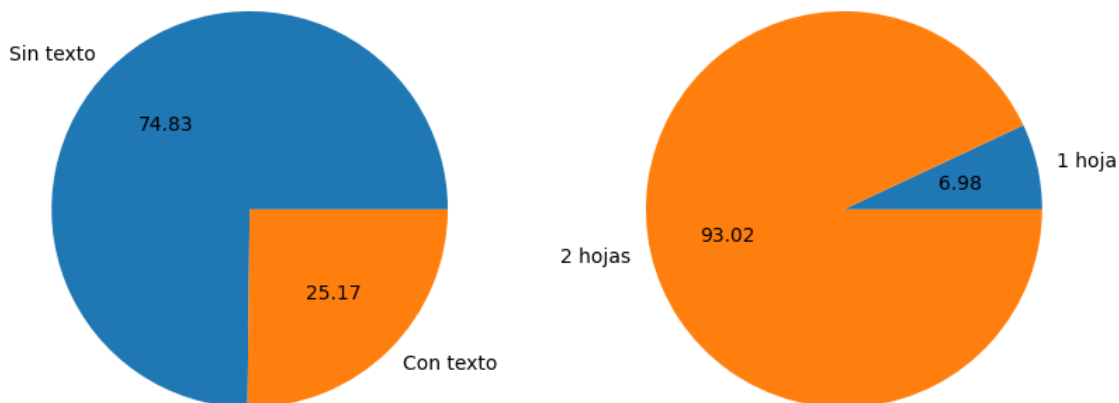
- Algunas facturas tienen una imagen QR, se excluyen sabiendo que tienen un tamaño fijo.
- La librería *pypdf* (o los PDFs en sí) devuelven las imágenes duplicadas, así que se deduplican utilizando un hash convencional.

EDA preliminar

Antes de empezar, conviene realizar un análisis del JSON que obtuvimos en el paso anterior. Nos podríamos preguntar:

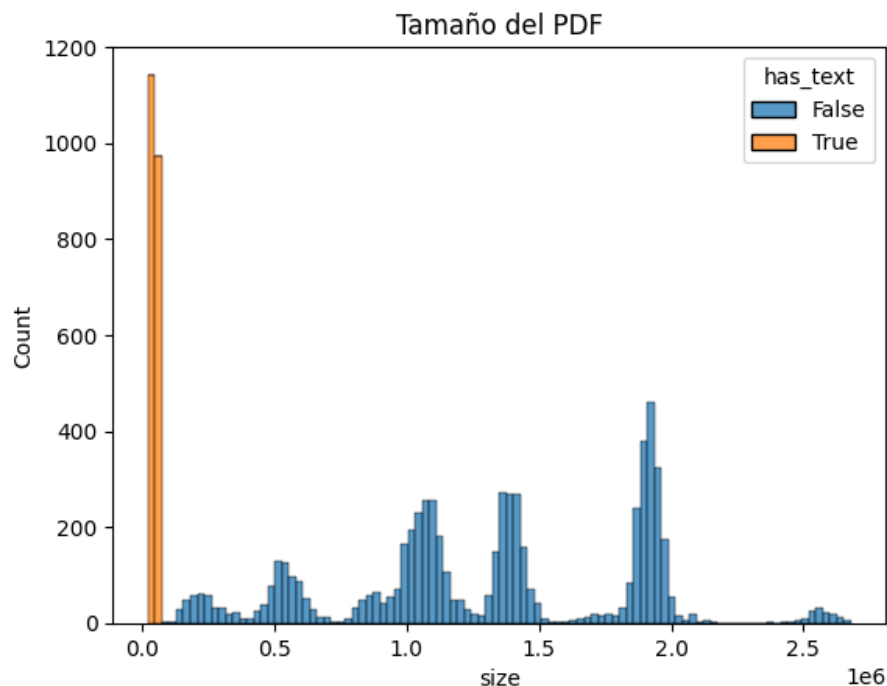
- ¿Cuántas hojas tienen los documentos? ¿Hay alguna factura extraña?
- ¿Cuántas facturas cuentan con texto? ¿Con imágenes? ¿Hay alguna vacía?
- ¿El tamaño nos puede decir algo?

Para responder estas preguntas, realizo un par de gráficos:



Con esto obtenemos bastante información, sabemos que en general tienen una o dos hojas y el 25% de los datos viene en forma de texto, mientras que el resto sólo tiene imágenes.

Y finalmente, el tamaño nos dice algo?



A primera vista, las facturas con texto (como era de esperarse) son las más pequeñas. Además podemos notar que aparecen grupos claramente separados. Esta información podría resultar útil, se puede hipotetizar: ¿Cada proveedor genera facturas de distintos tamaños? Podría ocurrir si utilizan escáners o cámaras de mayor o peor calidad, por ejemplo.

OCR

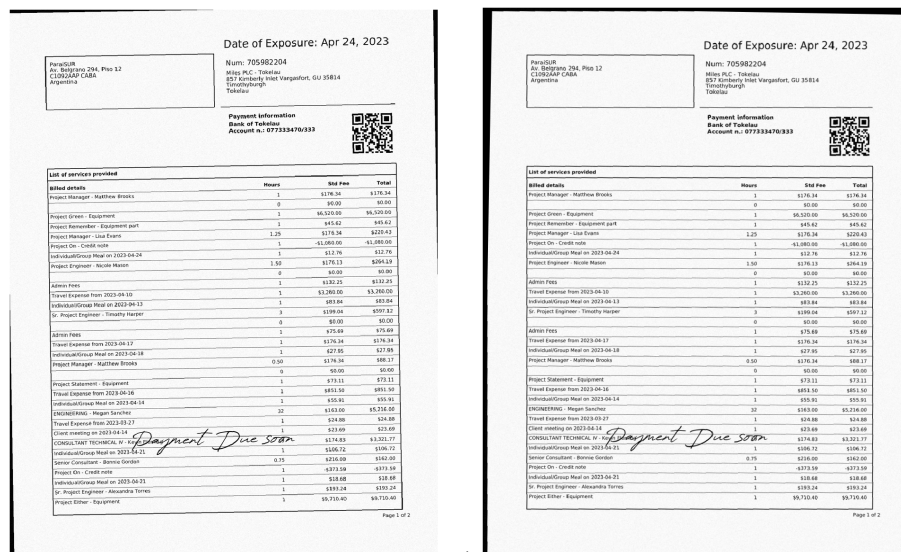
Necesitamos el contenido en texto para poder extraer la información. El 75% de las muestras no tienen texto nativo, así que tengo que emplear alguna técnica de OCR. Existen muchas librerías para realizar OCR: tesseract, easyocr, paddleocr, y más. Antes de pasar cada imagen por las librerías de OCR, es útil hacerle una especie de preprocesamiento. Por ejemplo: pasarla escala de grises, thresholding¹ y deskew (enderezar).

Como no sabía qué engine me iba a servir para este problema, corrí el OCR sobre todas las facturas con distintos engines y preprocesamientos, a la espera de obtener buenos resultados con alguno. Algunos ejemplos:

- paddleocr, sin deskew
- paddleocr, con deskew
- tesseract sin preprocesamiento
- tesseract con umbrales 150, 170, 190 y 210
- easyocr sin preprocesamiento

Este proceso demoró más de 30 horas de cómputo (perdí la cuenta). Como en todas las etapas, el proceso se podía parar y reiniciar en cualquier momento sin perder el progreso. Además se podía comenzar a utilizar lo ya procesado para avanzar con las otras etapas mientras esto seguía corriendo.

El procesamiento de thresholding filtra ruido en la imagen y es bastante estándar. El proceso de deskew (enderzar) se ve así:



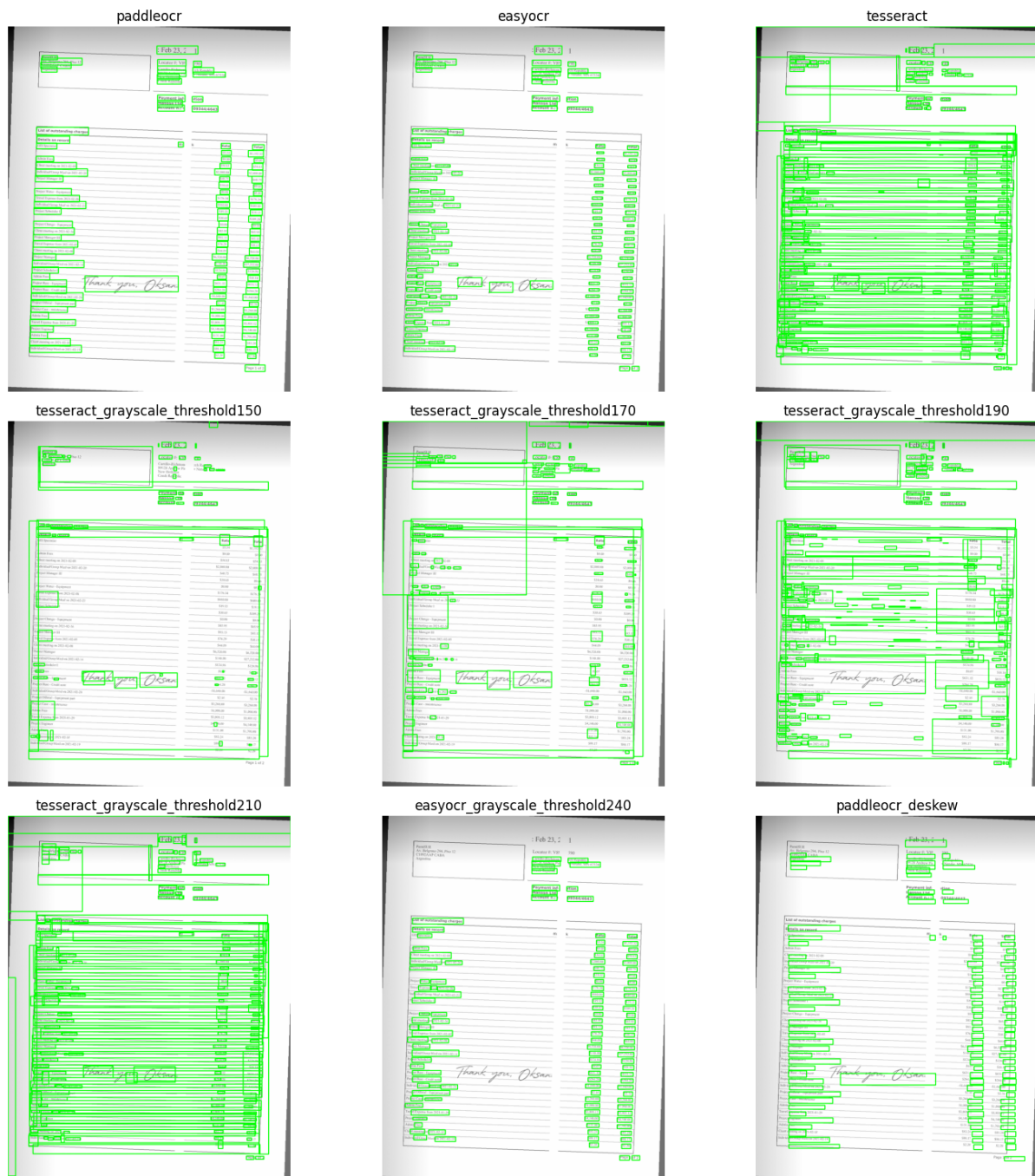
Both images show a scanned invoice from ParaiSUR. The left image is the original scan, and the right image is the result of deskewing (straightening). The text in the right image is clearly aligned and readable.

Invoice Details:

- ParaiSUR - R. Rodríguez 294, Piso 12 - C/2000AP CABA Argentina
- Núm: 705982204
- Edigo P.C. - Toluca - 557 Kennedy Hotel Vargavelos, GU 35814 Toluca
- Date of Exposure: Apr 24, 2023
- Payment Information: Bank of Toluca, Account no: 97733470/933

List of services provided:

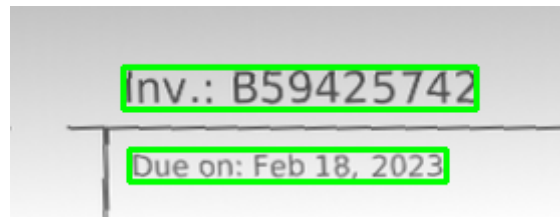
Item	Hours	Unit Price	Total
Project Manager - Matthew Brooks	1	\$176.34	\$176.34
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83.84
Project Engineer - Timothy Harper	3	\$199.04	\$597.12
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-17	1	\$75.88	\$75.88
IndividualGroup Meet on 2023-04-18	1	\$176.34	\$176.34
Project Manager - Matthew Brooks	1.50	\$176.34	\$264.51
Project Manager - Equipment	0	\$0.00	\$0.00
Project Manager - Equipment part	1	\$45.82	\$45.82
Project Manager - Lisa Evans	1.25	\$176.34	\$220.42
Project Dir - Credit note	1	\$1,086.00	\$1,086.00
IndividualGroup Meet on 2023-04-24	1	\$12.76	\$12.76
Project Engineer - Nicole Hester	1.50	\$176.34	\$264.51
Admin Fees	0	\$0.00	\$0.00
Travel Expense from 2023-04-10	1	\$12.25	\$12.25
IndividualGroup Meet on 2023-04-13	1	\$1,365.00	\$1,365.00
Project Engineer - Timothy Harper	1	\$83.84	\$83



Como se puede ver, los resultados varían drásticamente entre distintos engines y preprocesamientos.

Para cada uno, se guarda cada caja detectada junto con sus coordenadas (dibujadas en verde arriba).

Por ejemplo:



```
{
  "paddleocr": [{
    "boxes": [
      {"bounds": [[446, 60], [636, 60], [636, 80], [446, 80]],
      "text": "Inv.: B59425742", "confidence": 0.97},
      ...
    ],
    "text": ...
  }, ... ]
}
```

Invoice numbers

El siguiente paso es asociar cada muestra con su invoice number.

Algo extremadamente útil es que proveen una lista con todos los invoice numbers que debería haber en las muestras. Esto permite hacer fuzzy matching con cadenas arbitrarias y rellenar algunos espacios faltantes (debido a errores en el OCR)..

Se intenta extraer el invoice number, en orden:

1. Si el sample fue **etiquetado manualmente**, usamos ese invoice number. Algunas muestras las etiqueté a mano, ya que el fuzzing a veces falla y dos facturas terminan con el mismo invoice number (la real y la cercana). De esta manera arreglás dos facturas por cada una etiquetada. Etiqueté aproximadamente 50 muestras.
2. Si el **nombre del archivo** es un invoice number exacto, lo utilizamos.
3. Se intenta **extraer del texto** nativo del PDF. Se buscan todos los candidatos con la regex `([A-Z0-9@]{6,})` y se busca en la lista de invoice numbers un match perfecto.
4. Se intenta **extraer del texto obtenido por OCR**: sólo se consideran las cajas en la parte superior de la primera hoja (para evitar los Amount Overdue y Account Numbers). Se buscan candidatos igual que en el paso anterior y se matchean usando fuzzy matching.

La versión final logra extraer correctamente 8210 de 8411 (97.6%) invoice numbers. Sin repetidos (arreglados a mano) y 201 muestras con imágenes que no se reconocen bien (con texto el 100%). Podría etiquetar las otras 201 a mano pero no es reutilizable y la ganancia es 1 en vez de 2 como las duplicadas (además hay mejores cosas que hacer 😊).

Details (texto)

El texto nativo de los PDFs es una sola cadena separada por saltos de línea, donde cada columna es una línea. Quiero extraer cada *línea de detalle*, es decir [descripción, unidades, \$/unidad, total], de la lista de líneas sin separación alguna. Lo primero es separar la cabecera de la tabla del contenido, esto se puede detectar usando keywords simples ("Description", "\$/H", "Total", y muchos más). Una vez dentro de la sección de items, podemos armar la *línea de detalle* juntando todas las líneas de manera golosa hasta encontrar una que tiene letras (de esta manera corta en description y no las otras).

Las líneas extraídas se ven así:

```
['Alexandra Torres - Technical Supervisor', '2 ', '$193.24', '$386.48']
```

Y se retorna un arreglo con la siguiente estructura:

```
{
  "desc": "Alexandra Torres - Technical Supervisor",
  "price": 193.24,
  "hours": 2,
  "total": 386.48,
}
```

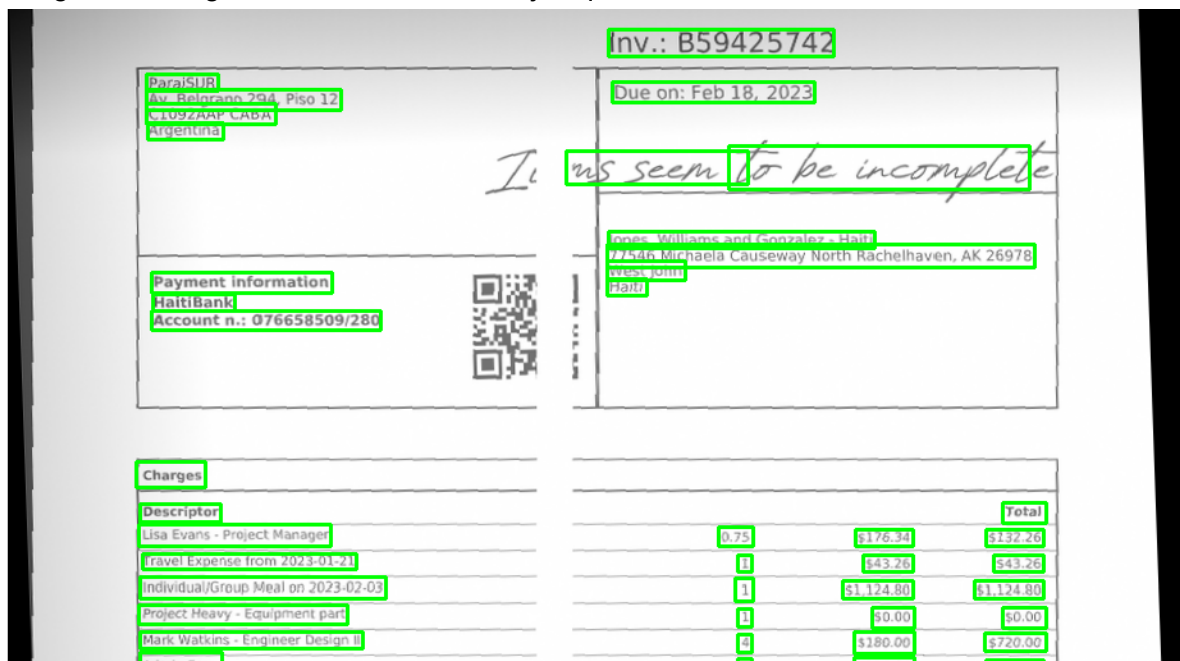
Hay una aserción que verifica que $price * hours = total$.

Details (OCR)

Para realizar esta extracción elegí usar paddleocr con deskew. paddleocr da muy buenos resultados con poca configuración. Y deskew es necesario para en los próximos pasos separar bien las líneas. Mas información al final.

Queremos filtrar todo lo que no sea una línea de detalle.

La siguiente imagen muestra todas las cajas que detecta el OCR.



Inv.: B59425742

Due on: Feb 18, 2023

ParaiSUR
Av. Belgrano 294, Piso 12
C1003ZAAJ, CABA
Argentina

Payment information
HaitiBank
Account n.: 076658509/280

QR code

Items seem to be incomplete

Jones, Williams and Gonzalez - Haiti
77546 Michaela Causeway North Rachelhaven, AK 26978
West Point
Haiti

Charges	Descriptor	Units	Price	Total
Lisa Evans - Project Manager		0.75	\$176.34	\$132.26
Travel Expense from 2023-01-21		1	\$43.26	\$43.26
Individual/Group Meal on 2023-02-03		1	\$1,124.80	\$1,124.80
Project Heavy - Equipment part		1	\$0.00	\$0.00
Mark Watkins - Engineer Design II		4	\$280.00	\$720.00
Admin Fees		1	\$124.80	\$124.80

Ahora agrupamos las cajas en líneas, usando la técnica descrita [acá](#). La idea es agrupar (mirando solo el eje Y) cajas cercanas tal que estén a la mitad de la mediana de las alturas de todas las cajas. En criollo, calculamos la mitad del line-height² y agrupamos en base a eso.

Inv.: B59425742

ParaiSUR
Av. Belgrano 294, Piso 12
C1092AAP CABA
Argentina

Due on: Feb 18, 2023

Items seem to be incomplete

Jones, Williams and Gonzalez - Haiti
77546 Michaela Causeway North Rachelhaven, AK 26978
West John
Haiti

Payment information
HaitiBank
Account n.: 076658509/280

Charges

Descriptor			Total
Lisa Evans - Project Manager	0.75	\$176.34	\$132.26
Travel Expense from 2023-01-21	1	\$43.26	\$43.26
Individual/Group Meal on 2023-02-03	1	\$1,124.80	\$1,124.80
Project Heavy - Equipment part	1	\$0.00	\$0.00
Mark Watkins - Engineer Design II	4	\$180.00	\$720.00
Admin Fees	1	\$117.63	\$117.63

Ahora es más fácil filtrar cajas que no nos interesan, nos quedamos con las líneas más largas que cierto umbral ($0.9 * \text{mediana del ancho de todas las líneas}$):

Inv.: B59425742

ParaiSUR
Av. Belgrano 294, Piso 12
C1092AAP CABA
Argentina

Due on: Feb 18, 2023

Items seem to be incomplete

Jones, Williams and Gonzalez - Haiti
77546 Michaela Causeway North Rachelhaven, AK 26978
West John
Haiti

Payment information
HaitiBank
Account n.: 076658509/280

Charges

Descriptor			Total
Lisa Evans - Project Manager	0.75	\$176.34	\$132.26
Travel Expense from 2023-01-21	1	\$43.26	\$43.26
Individual/Group Meal on 2023-02-03	1	\$1,124.80	\$1,124.80
Project Heavy - Equipment part	1	\$0.00	\$0.00
Mark Watkins - Engineer Design II	4	\$180.00	\$720.00
Admin Fees	1	\$117.63	\$117.63

² https://fonts.google.com/knowledge/glossary/line_height_leading

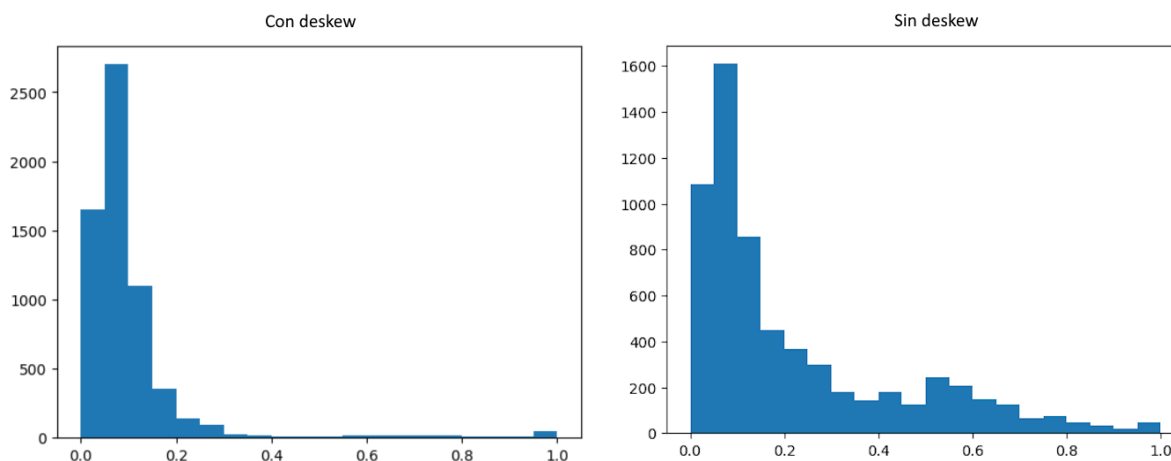
Finalmente intentamos construir la línea de detalle final. De todas las cajas que forman una línea, tomamos la que está más a la derecha y la intentamos parsear a float con un regex complejo. Si fue posible, la descripción es el resto de las cajas, ordenadas de izquierda a derecha. Esto incluye el precio unitario y la cantidad, pero no es una molestia para los pasos siguientes y es más conveniente dejarlo.

Entonces obtenemos un arreglo con una estructura parecida que para las líneas de texto:

```
{ 'desc': 'Lisa Evans - Project Manager 0.75 $176.34',  
  'total': 132.26 },  
{ 'desc': 'Travel Expense from 2023-01-21 I $43.26',  
  'total': 43.26 }
```

¿Por qué elegí paddleocr con deskew?

Comparando el resultado de varios motores de OCR, llegué a la conclusión de que paddleocr me daba los mejores reconocimientos. Aparte de eso, aplicando deskew como preprocesamiento mejoraba aún más los resultados:



Los gráficos anteriores muestran la diferencia entre utilizar deskew y no utilizarlo. El eje X es el porcentaje de líneas que no pudieron ser detectadas. El eje Y, la cantidad. Notar la escala del eje Y (perdí el código para crear el gráfico).

Utilizando deskew, en la mayoría de las muestras el porcentaje de líneas no detectado es bajo, mientras que sin deskew, la distribución se desparrama y es mucho peor.

En total se encuentran 56261 líneas de detalle en invoices con texto (2117) y 157010 en invoices con OCR (6294). Esto significa que hay 26.57 líneas por invoice en facturas de texto y 24.94 líneas por invoice en facturas de OCR. Los números son cercanos por lo que parece que estoy haciendo bien las cosas. Es un poco menor como era de esperarse, por no estar detectando bien algunas líneas (gráfico anterior).

Line classification

El siguiente paso es clasificar cada línea en profesional o no profesional para determinar si hay que incluirla en el total o no. Hay muchas líneas obvias que son claramente profesionales y otras que no. Tomó varios intentos de prueba y error ("probing") para ver si ciertas líneas eran o no profesionales. Ver comentarios adicionales para mi opinión en este tema.

Primero se separa la línea en palabras y frases de 2 y 3 palabras. Por ejemplo, si la entrada es "Tina Obrien -Principal 35.80 \$395.00", se obtienen:

- **words:** ["tina", "obrien", "principal"] (o... phrases1)
- **phrases2:** ["tina obrien", "obrien principal"]
- **phrases3:** ["tina obrien principal"]
- **phrases:** phrases2 + phrases3

Mi solución se basa en fuzzy matching de keywords/phrases y nombres/apellidos:

- **Nombres/Apellidos:** si una línea contiene un nombre seguido de un apellido, es un profesional. Tomé una base de datos de nombres y apellidos de [acá](#). Se itera **words** de a pares y se calcula el mejor puntaje de matching para el nombre y apellido, si la suma está por encima de un umbral, se considera profesional.
- **Keywords positivas/negativas:** se tiene una lista fija de palabras clave que indican que la línea corresponde o no a un profesional, por ejemplo "engineer", "analyst" o "specialist" son positivas y "equipment", "workshop" y "meeting" son negativas. Se hace fuzzy matching en todas las palabras en **words** y se marca como profesional o no profesional si el match está por encima de un umbral.
- **Phrases positivas/negativas:** funcionan exactamente igual que las keywords, sólo que con otro set de frases clave. Por ejemplo "extra hours", y "special on site" son positivas mientras que "travel expense" y "group meal" son negativas. Se hace matching en **phrases**. Las separé de las keywords para tener más control sobre el set y los umbrales.

En caso de empates, la negativa tiene precedencia.

Para optimizar las palabras/frases clave y optimizar los umbrales evitando regresiones armé archivos de referencia (tests/professional_lines.txt, tests/non_professional_lines.txt), con líneas que clasifiqué yo a mano. Cada cambio realizado se verificaba mirando los falsos positivos y falsos negativos, intentando minimizar la cantidad de ambos.

En resumen el proceso era: abrir invoices al azar y copiar líneas que no haya visto al archivo correspondiente, ajustar el algoritmo para que no tenga falsos +/- y repetir.

A continuación, el recuento de líneas.

Tipo de factura	Cantidad de facturas	Líneas totales	Líneas profesionales	Líneas <u>no</u> profesionales
Sólo texto	2117	56261	15552	40709
Sólo OCR	6294	157010	43548	113462
Total	8411	213271	59100 (27.7%)	154171

Submission

El paso final es generar el CSV, utilizando el resultado de los pasos anteriores. Se suman los totales de las líneas que son clasificadas como profesionales. Se aprovecha el dataset de referencia hecho a mano (es gratis).

Hay dos maneras de que una factura sea considerada ilegible:

- Si el invoice number no es reconocido correctamente
- Si el porcentaje de líneas no reconocidas es mayor a **0.3** (lo que vimos en el deskew)

En estos casos, debemos decidir qué valor asignar a cada una de estas **442** facturas.

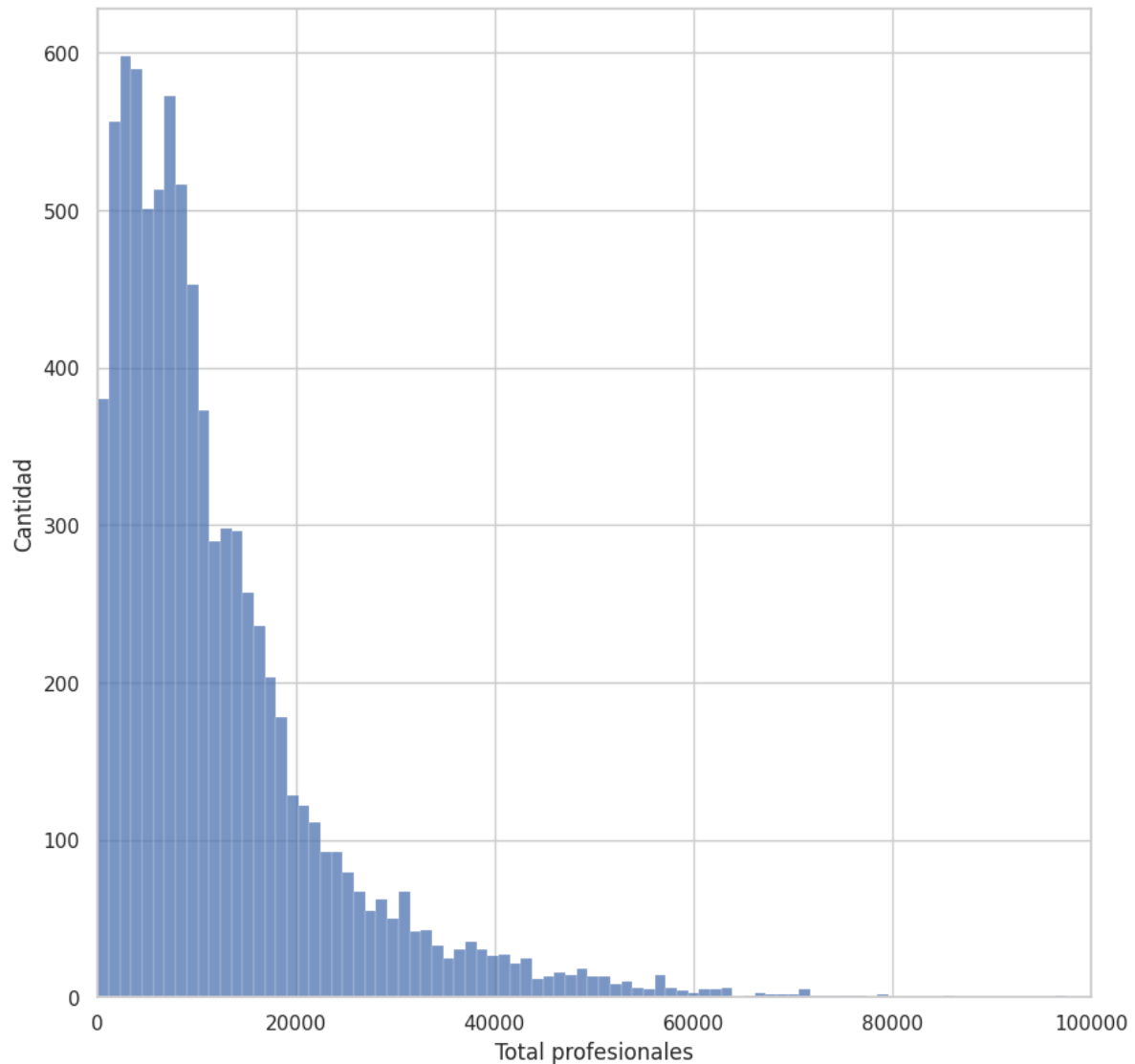
Dado que la métrica utilizada es MAE, sabemos que el mejor estimador para éste es la **mediana**: https://en.wikipedia.org/wiki/Mean_absolute_error#Proof_of_optimality.

Rellenamos las muestras que no reconocemos con la mediana, que en el submit final el valor fue 9102.57. Como dato adicional, el promedio fue de 12416.72.

Análisis extra

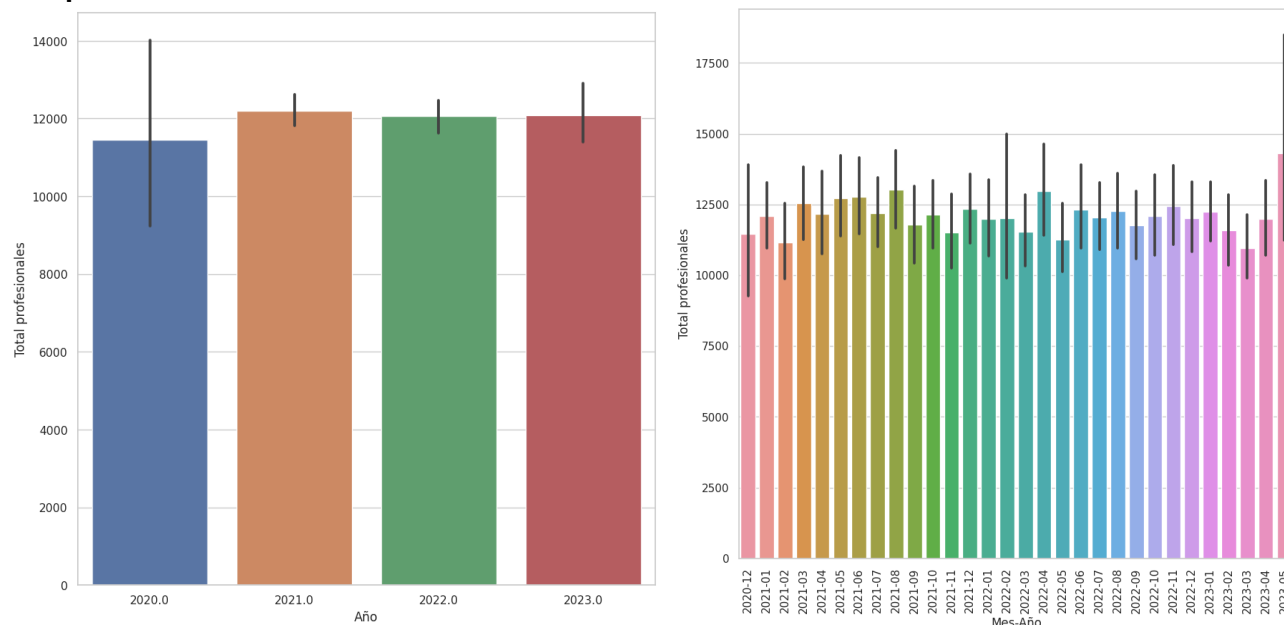
Ahora intento extraer más información teniendo el total de profesionales. El objetivo es encontrar algún patrón que pueda ayudarme a mejorar la predicción de líneas o estimar mejor para las que no tengo información.

Distribución general



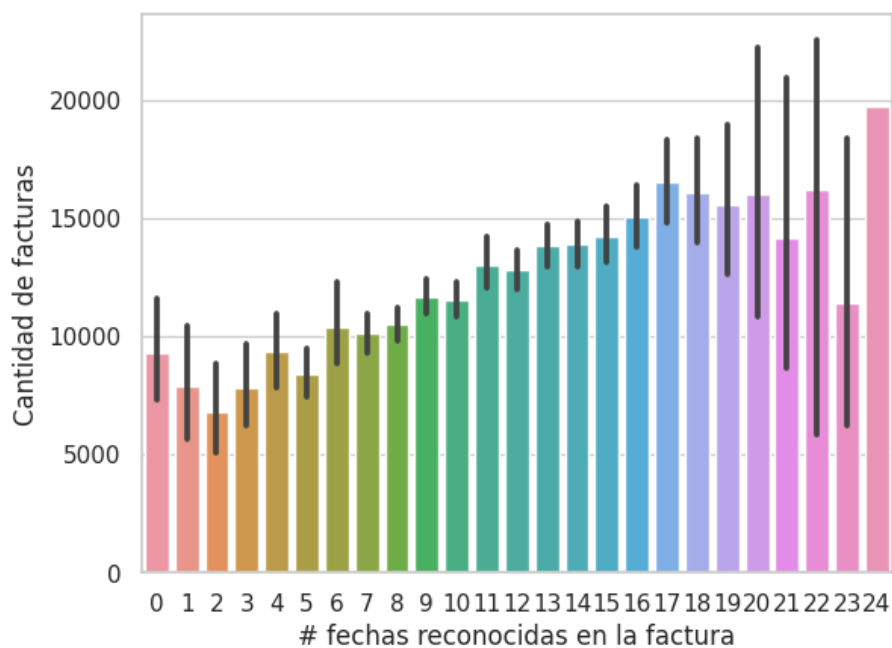
La distribución de los totales es de cola pesada. La mayoría de los totales se encuentran en números razonablemente bajos (con sentido al ser facturas).

Temporalidad



Parece que no hay ningún patrón a lo largo del tiempo; los totales profesionales se distribuyen uniformemente a lo largo de los años y meses.

Cantidad de fechas (experimento)



La cantidad de fechas detectables en una factura parece correlacionarse con los totales profesionales. Esto parece que tiene que ver más con la cantidad de líneas que hay una factura y no con las fechas en sí. Si hay más fechas, hay (en general) más líneas, y por lo tanto mayor probabilidad de que una de ellas sea profesional.

Segmentar por proveedor (no me da el tiempo para hacer este análisis pero lo voy a comentar)

Puede llegar a ser posible segmentar los archivos por proveedor, conociendo keywords que usan, el tamaño de los archivos (como mostré al principio), información de pago, el país, etc.

Si se logra segmentar se pueden armar clasificadores más especializados y con mejores resultados. Aunque hacer esto incurrirá un costo más elevado.

Muestra de 50³ facturas

Un día antes del cierre de la competencia, se liberaron 50 muestras con sus totales “bien” calculados. Para esas muestras, obtenía un 3845,04 de MAE en mi submit 48 (el mejor hasta ahora).

Revisando cuidadosamente las muestras con diferencias, no logré encontrar líneas que justifiquen esas sumas. Utilicé el algoritmo subsetsum⁴ para verificar si existe una posible combinación de líneas que resulte en el número que dan como referencia.

El procedimiento consistía en tomar todas las líneas, excluirlas que obviamente no eran líneas profesionales y buscar un subconjunto de las líneas matchee con el número de referencia o se acerque. No logré hacerlo en ninguna de las que probé (que eran las tenían una diferencia).

Viendo que el lore venía como:

> **El equipo de ParaiSUR estuvo trabajando a la par de ustedes y, dado que mañana es el deadline, les compartimos el archivo con las facturas que llegaron a revisar a mano.**

Simularía que las etiquetaron personas a mano y podría contener errores. Otra opción es que hayan copiado del ground truth de Kaggle y filtrado, por ende son los valores reales e hice cualquier cosa.

Ya fuera del lore, mi teoría es que alteraron la mitad de las muestras para que no se las copien directamente a Kaggle. Para ajustar el MAE localmente sirve igual (pero no vas a llegar a 0, ¿para desalentarlo?).

³ 49 en realidad, **869402605** está repetida.

⁴ https://es.wikipedia.org/wiki/Problema_de_la_suma_de_subconjuntos

Mejoras posibles

- Se puede mejorar bastante el OCR, medio que me casé con la primera opción que ví que daba buenos resultados.
- No verifiqué que todas las imágenes de los PDFs midieran lo mismo, eso podría causar problemas con algunos umbrales basados en píxeles que uso.
- Actualmente sólo es posible matchear nombres si se detecta bien un espacio en el medio. Por ejemplo, "Douglas Green" se reconoce bien, pero "DouglasGreen" no es reconocido. Esto es porque hago un fuzzy matching para cada uno (nombre y apellido). No encontré una solución eficiente para esto, ya que hacer fuzzy matching con todas las concatenaciones es prohibitivo (458652000 pares). ¿Quizás aprovechar la mayúscula?
- En el análisis tenemos una hipótesis de que la cantidad de líneas está correlacionada con el total. Se podría ver la altura a la que está el total (teniendo en cuenta la cantidad de hojas también) y estimar la cantidad de items que deberíamos reconocer (o reconocer directamente la estructura de la tabla para saber la cantidad de filas). Esto nos puede servir para filtrar mejor si una factura tiene muchas líneas no legibles o directamente conseguir un mejor estimador para una factura ilegible.
- Intenté usar FastText para clasificar las líneas, ya que se basa en n-gramas y quizás funcionaba mejor que fuzzing pero no logré hacer que funcione mejor.

Comentarios adicionales (y opiniones)

- **Kaggle:** El MAE que muestra Kaggle (22k) en mi opinión muy alto para un dataset donde la media es ~12k y la mediana ~8k. No estoy muy seguro a que se debe eso, quizás hay una factura que no encontré con un total extremadamente alto (o bajo). O quizás le suman un número fijo al MAE (?), ya estoy conspiranoico.
- **Origen del dataset:** el dataset se nota que es muy artificial. Los valores muchas veces no tienen sentido, ni en pesos ni dólares, y el análisis temporal no muestra inflación ni nada. Esto fue un poco desalentador, aunque por otro lado si era real era como hacer trabajo gratis, que tampoco está bueno.
El trabajo que hicieron para que parezcan tomadas como fotografías quedó muy bien, me gustó.
- **Dificultad:** En mi opinión la dificultad de la competencia pasó por saber (como humano) si las líneas eran o no profesionales. Entiendo que quieren capturar el "ser consultores" pero en un caso real vos preguntás y una persona que sabe del negocio tiene que saber exactamente qué es y qué no es. Para mí competencias de este estilo deberían ser de información perfecta, autocontenida. Como fue, todos dependemos de ustedes para clasificar las líneas, no me pareció divertido.