Propuesta de solución Reto Hackathon

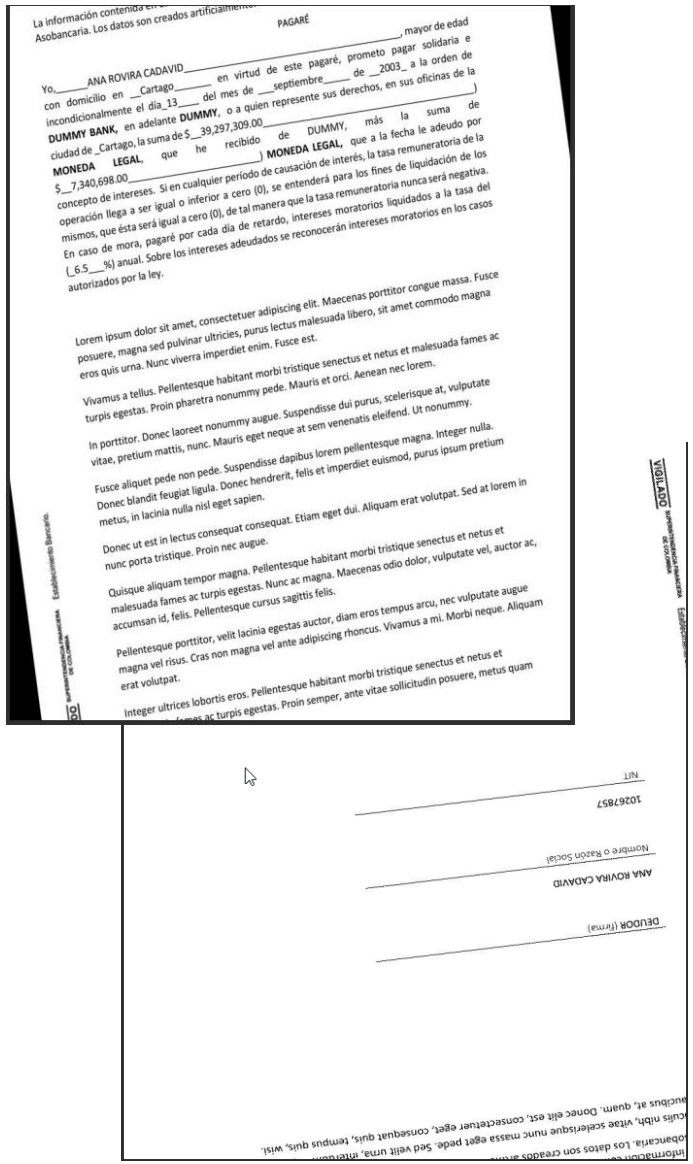
19 de septiembre de 2024

Aso Ban Caria | Acerca la Banca a los Colombianos

Consideraciones iniciales:

- Es un ejercicio académico, los datos son creados artificialmente
- Nos salimos del mundo determinístico y nos adentramos en el mundo de las estimaciones (aproximación)
- No existe una única solución para resolver el problema
- La versión presentada no necesariamente es la forma óptima de resolver el problema - Pueden existir otras formas y otras herramientas para mejorar la puntuación final
- Premisa inicial: Herramientas open-source para la solución del reto.

Objetivo del reto: Tabular información no estructurada



PAGARÉ

Yo,_____ ANA ROVIRA CADAVID_____, mayor de edad con domicilio en __Cartago_____ en virtud de este pagaré, prometo pagar solidaria e incondicionalmente el día_13___ del mes de ___septiembre_____ de __2003_ a la orden de **DUMMY BANK,** en adelante **DUMMY**, o a quien represente sus derechos, en sus oficinas de la ciudad de _Cartago, la suma de $__39,297,309.00_____ **MONEDA LEGAL**, que he recibido de DUMMY, más la suma de $__7,340,698.00_____ **MONEDA LEGAL,** que a la fecha le adeudo por concepto de intereses. Si en cualquier período de causación de interés, la tasa remuneratoria de la operación llega a ser igual o inferior a cero (0), se entenderá para los fines de liquidación de los mismos, que ésta será igual a cero (0), de tal manera que la tasa remuneratoria nunca será negativa. En caso de mora, pagaré por cada día de retardo, intereses moratorios liquidados a la tasa del (_6.5___%) anual. Sobre los intereses adeudados se reconocerán intereses moratorios en los casos autorizados por la ley.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna. Nunc viverra imperdiet enim. Fusce est.

Vivamus a tellus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci. Aenean nec lorem.

In porttitor. Donec laoreet nonummy augue. Suspendisse dui purus, scelerisque at, vulputate vitae, pretium mattis, nunc. Mauris eget neque at sem venenatis eleifend. Ut nonummy.

Fusce aliquet pede non pede. Suspendisse dapibus lorem pellentesque magna. Integer nulla. Donec blandit feugiat ligula. Donec hendrerit, felis et imperdiet euismod, purus ipsum pretium metus, in lacinia nulla nisl eget sapien.

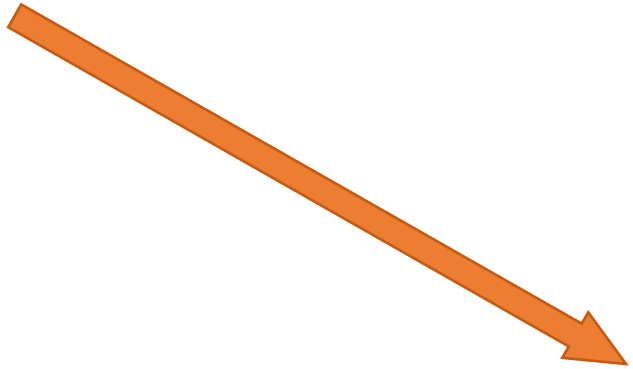
Donec ut est in lectus consequat consequat. Etiam eget dui. Aliquam erat volutpat. Sed at lorem in nunc porta tristique. Proin nec augue.

Quisque aliquam tempor magna. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nunc ac magna. Maecenas odio dolor, vulputate vel, auctor ac, accumsan id, felis. Pellentesque cursus sagittis felis.

Pellentesque porttitor, velit lacinia egestas auctor, diam eros tempus arcu, nec vulputate augue magna vel risus. Cras non magna vel ante adipiscing rhoncus. Vivamus a mi. Morbi neque. Aliquam erat volutpat.

Integer ultrices lobortis eros. Pellentesque habitant morbi tristique senectus et netus et

NIT
10267857
Nombre o Razón Social
ANA ROVIRA CADAVID
DEUDOR (firma)

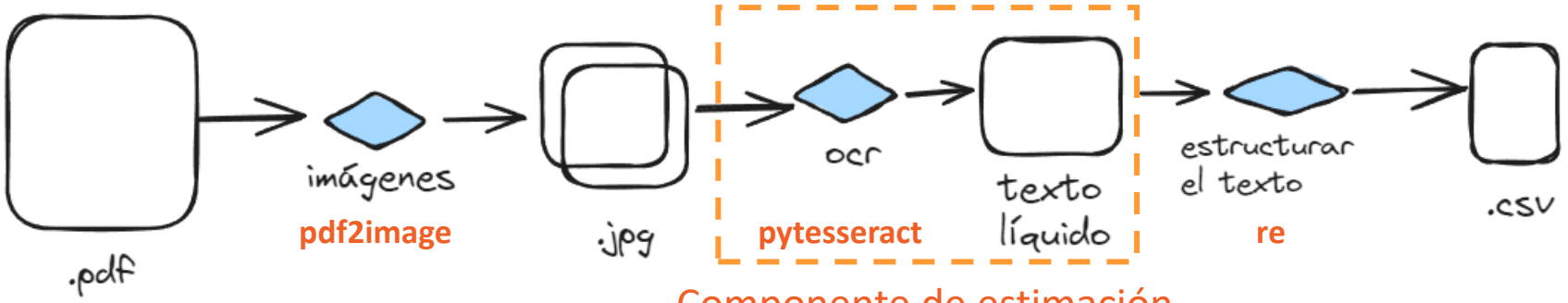


```
ID;nombre;ciudad;anio;mes;dia;valor;intereses;tasa_intereses;id_cliente;tipo_documento;firmado
1;ANA ROVIRA CADAVID;Cartago;2003;septiembre;13;39,297,309.00;7,340,698.00;6.5;10267857;NIT;0
```

ejemploSalida.csv


El Servidor Murió
¿Dónde está el Backup?
En el Servidor

...Lo que vimos en la charla técnica:



Componente de estimación

ÓN – Establecimiento Bancano.

VIGILADO

Asobancarila. Los datos son creados artificialmente. ARS

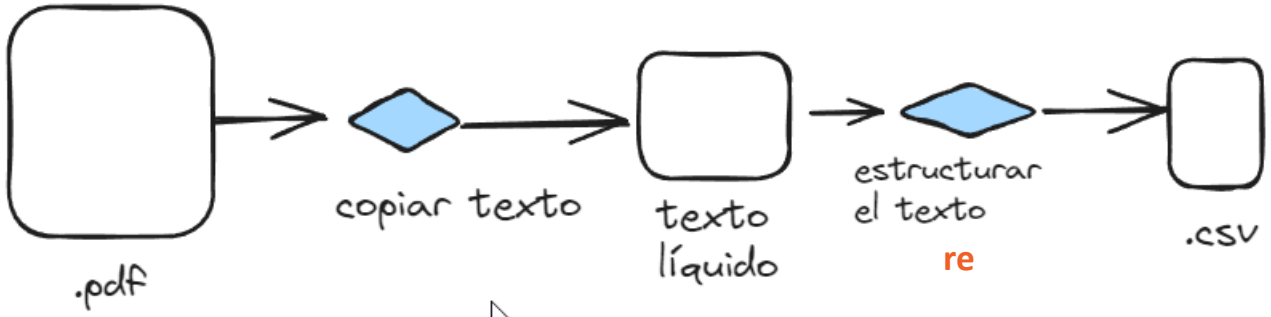
PAGARÉ No. 4
Yo, NATALIA VICTOR RODRIGUEZ
VARGAS , mayor de edad con domicilio en
_ Jamundí en virtud de este pagaré, prometo pagar solidaria e incondicionalmente el
día 2 del mes de __ agosto ____de _ 2010 _ a la orden de DUMMY BANK, en adelante
DUMMY, oa quien represente sus derechos, en sus oficinas de la ciudad de _Jamundí, la suma de
S 99,080,625.00) MONEDA LEGAL, que he
recibido de DUMMY, más la suma de $_ 5,019,933.00) MONEDA

LEGAL, que a la fecha le adeudo por concepto de intereses. Si en cualquier período de causación
de interés, la tasa remuneratoria de la operación llega a ser igual o inferior a cero (0), se entend
para los fines de liquidación de los mismos, que ésta será igual a cero (0), de tal manera que la ta
remuneratoria nunca será negativa. En caso de mora, pagaré por cada día de retardo, intereses
moratorios liquidados a la tasa del (6.6 %) anual. Sobre los intereses adeudados se reconocerán
intereses moratorios en los casos autorizados por la ley.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce
posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna
eros quis urna. Nunc viverra imperdiet enim. Fusce est.

ID	4
nombre	NATALIA VICTOR RODRIGUEZ VARGAS

¿Qué pasaría si el pdf no fuera de imágenes? (Determinístico):



La información contenida en este documento es de uso exclusivo para la hackaton de auditoría Asobancaria. Los datos son creados artificialmente.

DUMMY BANK

PAGARÉ **No. 4**

Yo,_____NATALIA VICTOR RODRIGUEZ VARGAS_____, mayor de edad con domicilio en __Jamundi_____ en virtud de este pagaré, prometo pagar solidaria e incondicionalmente el día_2____ del mes de ___agosto_____ de __2010_ a la orden de **DUMMY BANK,** en adelante **DUMMY,** o a quien represente sus derechos, en sus oficinas de la ciudad de _Jamundí, la suma de $__99,080,625.00_____) **MONEDA LEGAL,** que he recibido de DUMMY, más la suma de $__5,019,933.00_____) **MONEDA LEGAL,** que a la fecha le adeudo por concepto de intereses. Si en cualquier período de causación de interés, la tasa remuneratoria de la operación llega a ser igual o inferior a cero (0), se entenderá para los fines de liquidación de los mismos, que ésta será igual a cero (0), de tal manera que la tasa remuneratoria nunca será negativa. En caso de mora, pagaré por cada día de retardo, intereses moratorios liquidados a la tasa del (_6.6___%) anual. Sobre los intereses adeudados se reconocerán intereses moratorios en los casos autorizados por la ley.

00004.pdf

```
ÓN – Establecimiento Bancano.

VIGILADO

Asobancarila. Los datos son creados artificialmente. ARS

PAGARÉ No. 4
Yo, NATALIA VICTOR RODRIGUEZ
VARGAS , mayor de edad con domicilio en
_ Jamundí en virtud de este pagaré, prometo pagar solidaria e incondicionalmente el
día 2___ del mes de __ agosto_____de _ 2010_ a la orden de DUMMY BANK, en adelante
DUMMY, oa quien represente sus derechos, en sus oficinas de la ciudad de _Jamundí, la suma de
S_ 99,080,625.00 ) MONEDA LEGAL, que he
recibido de DUMMY, más la suma de $_ 5,019,933.00 ) MONEDA

LEGAL, que a la fecha le adeudo por concepto de intereses. Si en cualquier período de causación
de interés, la tasa remuneratoria de la operación llega a ser igual o inferior a cero (0), se entend
para los fines de liquidación de los mismos, que ésta será igual a cero (0), de tal manera que la ta
remuneratoria nunca será negativa. En caso de mora, pagaré por cada día de retardo, intereses
moratorios liquidados a la tasa del (_6.6___%) anual. Sobre los intereses adeudados se reconocerán
intereses moratorios en los casos autorizados por la ley.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce
posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna
eros quis urna. Nunc viverra imperdiet enim. Fusce est.
```

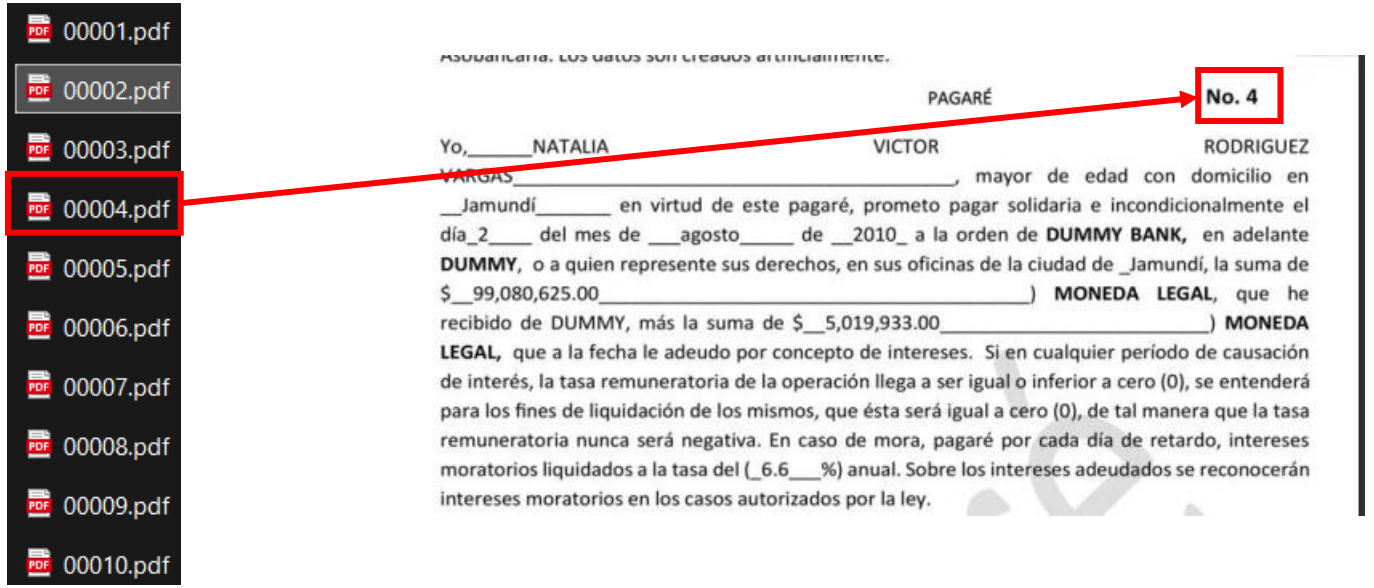
ID	4
nombre	NATALIA VICTOR RODRIGUEZ VARGAS

Estrategia Ágil para el reto:

- Esquema iterativo
  - División del trabajo en partes más pequeñas
- Validación constante.

- 12 campos
  - 9 en la página 1 y
  - 3 en la página 2

- En la iteración 1 iniciemos con el campo ID.



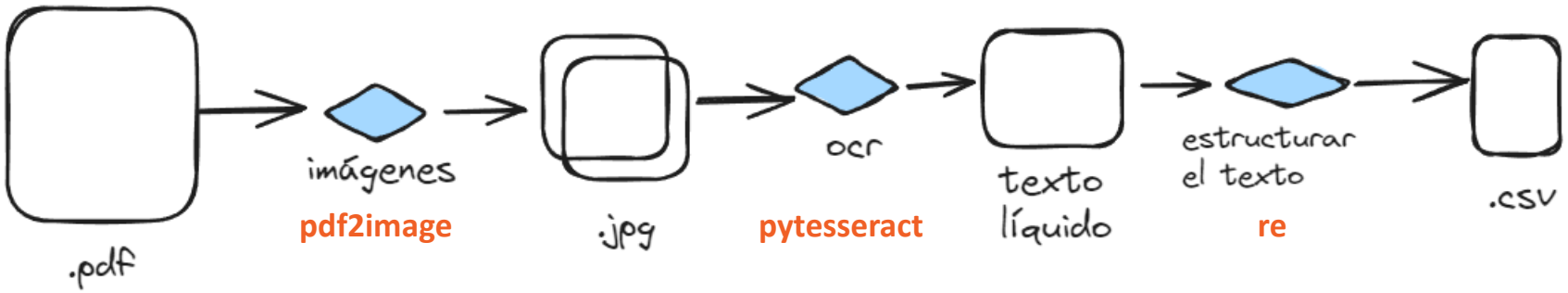
- Implementación en Python
  - Recorrer la carpeta, leer el nombre del archivo
  - Crear un archivo .csv con el nombre del archivo y los demás campos en blanco
- Pasar el archivo resultado para validación

Score de validación: **18.2**

Ranking **3er puesto.**

Iteración 2 campo Nombre.

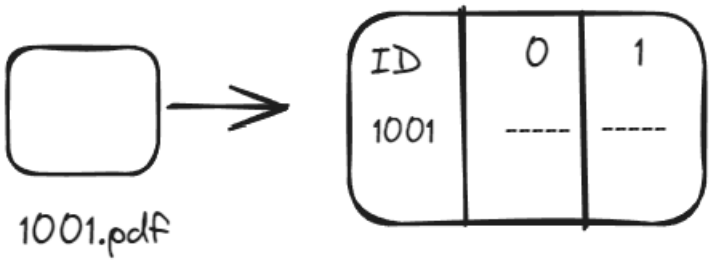
- Utilicemos lo aprendido en la charla técnica y exploremos todos los documentos con un OCR simple.



- Guardaremos las imágenes por si requerimos procesarlas posteriormente.

Iteración 2 campo Nombre.

- Utilicemos lo aprendido en la charla técnica y exploremos todos los documentos con un OCR simple.
- Guardaremos la información de cada página en archivos diferentes - (Cada página tiene transformaciones diferentes)
- Guardaremos el texto líquido de cada página en archivos diferentes (Cada página tiene campos diferentes)
- Cargar el texto a un DataFrame de pandas para procesarlo



- Aplicar expresión regular para extraer la información del nombre.

```
df['0'].str.extract(r'Yo, (.*)[mtMT]ayo[rt]')
```

Score de validación: **22.1**

Ranking **2do puesto.**

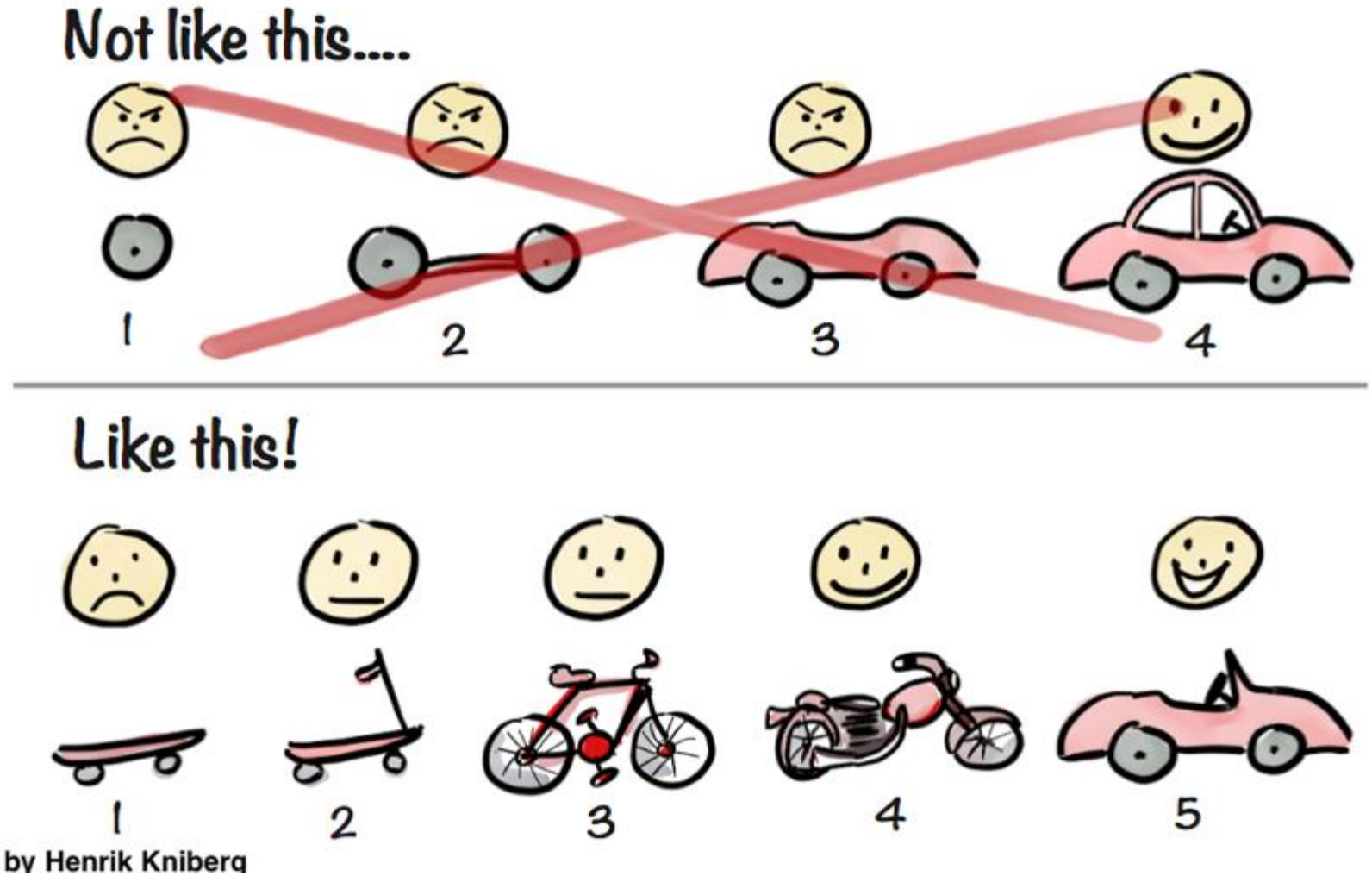
Repetimos el mismo procedimiento para todos los campos

- Nuevo reto: Identificar la expresión regular

Score de validación: **57.3**

Ranking **1er puesto.**


Test
Build
Review
Plan


Not like this....
1
2
3
4
Like this!
1
2
3
4
5
by Henrik Kniberg

Iteración 3

- Identificar las imágenes giradas 180° y rotarlas
- Modificar la configuración del tesseract
- Repetir pasos de la Iteración 2

Opciones de configuración de tesseract:

```
custom_config = r'--oem 3 --psm 6'
txt=pytesseract.image_to_string(img, lang='spa', config =
custom_config)
```

oem: (int) {0, 1, 2, 3}. OEM hace referencia al modo del motor OCR (OCR engine mode en inglés). Tesseract tiene 2 motores, Legacy Tesseract y LSTM, y el parámetro oem permite escoger cada uno de estos motores por separado, ambos al tiempo o automáticamente:

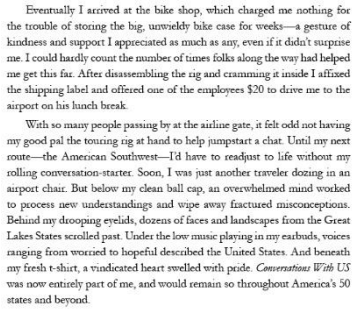
- 0: utilizar únicamente el motor Legacy.
- 1: utilizar únicamente el motor de redes neuronales LSTM.
- 2: utilizar los motores Legacy y LSTM.
- 3: escoger el motor según lo que hay disponible.

psm: (int) {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13}. PSM hace referencia a los modos de segmentación de las páginas (page segmentation modes, en inglés) de la librería Pytesseract. Cada número hace referencia a un modo de segmentación:

Tesseract Page Segmentation Modes (PSMs) Explained: How to Improve Your OCR Accuracy

```
$ tesseract --help-psm
Page segmentation modes:
  0    Orientation and script detection (OSD) only.
  1    Automatic page segmentation with OSD.
  2    Automatic page segmentation, but no OSD, or OCR. (not implemented)
  3    Fully automatic page segmentation, but no OSD. (Default)
  4    Assume a single column of text of variable sizes.
  5    Assume a single uniform block of vertically aligned text.
  6    Assume a single uniform block of text.
  7    Treat the image as a single text line.
  8    Treat the image as a single word.
  9    Treat the image as a single word in a circle.
 10    Treat the image as a single character.
 11    Sparse text. Find as much text as possible in no particular order.
 12    Sparse text with OSD.
 13    Raw line. Treat the image as a single text line,
       bypassing hacks that are Tesseract-specific.
```

--psm 0 “meta-information” OSD detecta ángulos de rotación fijos 0, 90, 180 ,270



Eventually I arrived at the bike shop, which charged me nothing for the trouble of storing the big, unwieldy bike case for weeks—a gesture of kindness and support I appreciated as much as any, even if it didn't surprise me. I could hardly count the number of times folks along the way had helped me get this far. After disassembling the rig and cramming it inside I affixed the shipping label and offered one of the employees $20 to drive me to the airport on his lunch break.

With so many people passing by at the airline gate, it felt odd not having my good pal the touring rig at hand to help jumpstart a chat. Until my next route—the American Southwest—I'd have to readjust to life without my rolling conversation-starter. Soon, I was just another traveler dozing in an airport chair. But below my clean ball cap, an overwhelmed mind worked to process new understandings and wipe away fractured misconceptions. Behind my drooping eyelids, dozens of faces and landscapes from the Great Lakes States scrolled past. Under the low music playing in my earbuds, voices ranging from worried to hopeful described the United States. And beneath my fresh t-shirt, a vindicated heart swelled with pride. *Conversations With US* was now entirely part of me, and would remain so throughout America's 50 states and beyond.

```
Page number: 0
Orientation in degrees: 0
Rotate: 0
Orientation confidence: 4.89
Script: Latin
Script confidence: 4.69
```

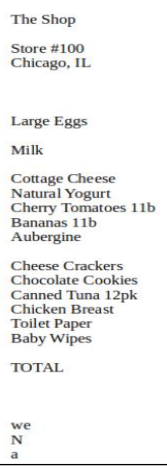
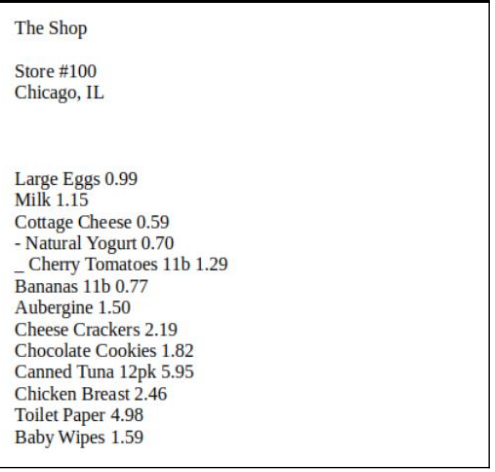
- How the page is oriented, in degrees, where `angle = {0, 90, 180, 270}`.
- The confidence of the script (i.e., graphics signs/writing system), such as Latin, Han, Cyrillic, etc.

--psm 1 La información del OSD se utiliza en de forma automática en el proceso de OCR. Tesseract ejecuta el OSD intermaente pero no se lo retorna al usuario.

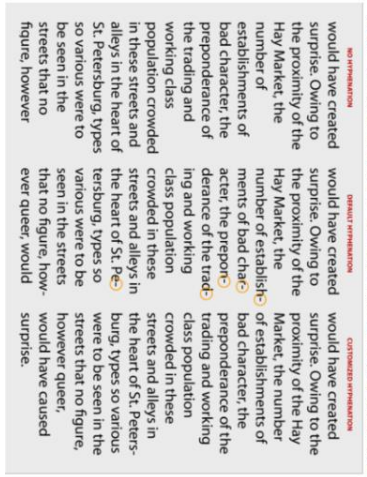
--psm 2 no está implementado

--psm 3 es el comportamiento por defecto. Tesseract intenta segmentar el texto y retorna el texto

--psm 4 Columna de texto de tamaño variable. Hojas de cálculo, tablas, facturas, recetas, etc.

Original	--psm 3	--psm 4
		

--psm 5 Único bloque de Código con texto alineado de forma vertical. Funciona bien solo con imágenes rotadas 90° en sentido de las manecillas del reloj. (psm 5 incluye psm 4)

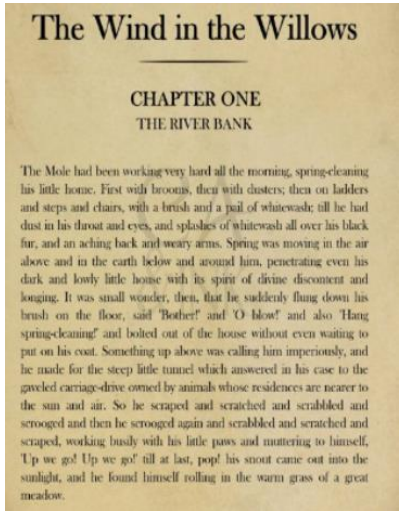


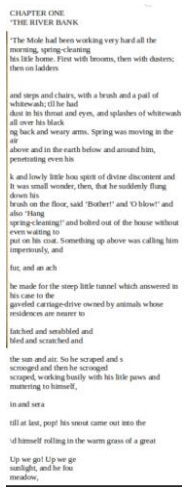
NO HYPHENATION DEFAULT HYPHENATION CUSTOMIZED HYPHENATION
would have created would have created would have created
surprise. Owing to surprise. Owing to surprise. Owing to the
the proximity of the the proximity of the —_ proximity of the Hay
Hay Market, the Hay Market, the Market, the number
number of number of establish-) of establishments of
establishments of | ments of bad char-) bad character, the
bad character, the acter, the prepon- preponderance of the
preponderance of — derance of the trad-)_ trading and working
the trading and ing and working class population
working class class population crowded in these
population crowded crowded in these streets and alleys in
in these streets and streets andalleysin the heart of St. Peters-
alleys inthe heart of the heart of St.Pe-) burg, types so various
St. Petersburg, types tersburg, types so were to be seen in the
so various were to various were tobe _ streets that no figure,
be seen in the seeninthe streets | however queer,
streets that no that no figure, how- would have caused
figure, however ever queer, would surprise.

Store #100
Chicago, IL
Sessessss |
Large Eggs 0.99 |
Milk 1.15
Cottage Cheese 0.59
- Natural Yogurt 0.70
_ Cherry Tomatoes 11b 1.29
Bananas 11b 0.77
Aubergine 1.50
Cheese Crackers 2.19
Chocolate Cookies 1.82
Canned Tuna 12pk 5.95
Chicken Breast 2.46
Toilet Paper 4.98
Baby Wipes 1.59
TOTAL $25.97 |
Sesessas=

--psm 6 Asume un único bloque de texto uniforme: El significado del texto uniforme es un único tipo de letra sin ninguna variación pertinente para un tipo de página de un libro o novela. (el por defecto tiene saltos de línea, espacios y algunos errores que se deben corregir manualmente)

--psm 3 (defecto)





--psm 6

The Wind in the Willows

CHAPTER ONE
'THE RIVER BANK
'The Mole had been working very hard all the morning, spring-cleaning his little home. First with brooms, then with dusters; then on ladders and steps and chairs, with a brush and a pail of whitewash; till he had dust in his throat and eyes, and splashes of whitewash all over his black fur, and an aching back and weary arms. Spring was moving in the air above and in the earth below and around him, penetrating even his dark and lowly little house with its spirit of divine discontent and longing. It was small wonder, then, that he suddenly flung down his brush on the floor, said 'Bother!' and 'O blow! and also 'Hang spring-cleaning!' and bolted out of the house without even waiting to put on his coat, Something up above was calling him imperiously, and he made for the steep little tunnel which answered in his case to the gaveled carriage-drive owned by animals whose residences are nearer to the sun and air. So he scraped and scratched and scrabbled and scrooged and then he scrooged again and scrabbled and scratched and scraped, working busily with his little paws and muttering to himself, 'Up we go! Up we go!" till at last, pop! his snout came out into the sunlight, and he found himself rolling in the warm grass of a great meadow.

--psm 7 Procesa la imagen cómo una sóla línea y PSM 8. procesa la imagen cómo una sola palabra (Horizontal)

PSM 9. Procesa la imagen como una palabra en un circulo (es extraño y se recomienda evitar si es posible)

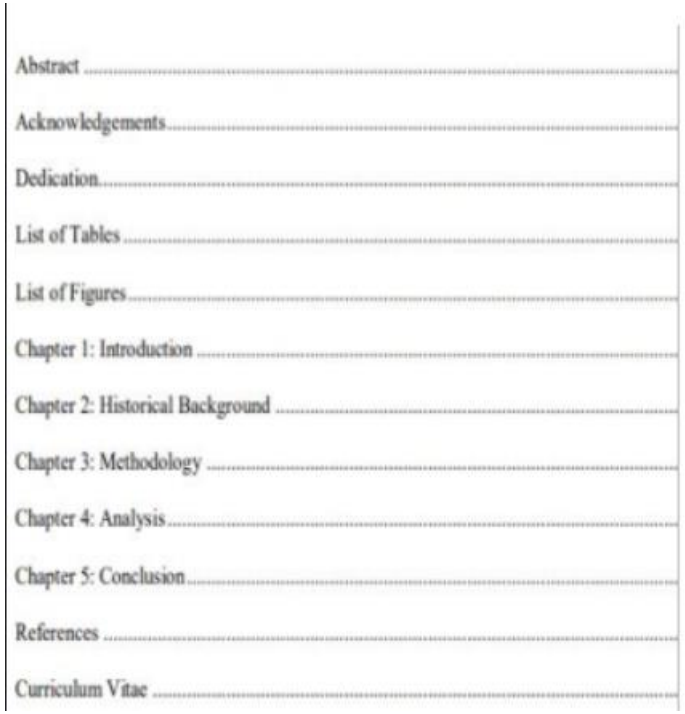
PSM 10. Procesa la imagen cómo un solo caracter.

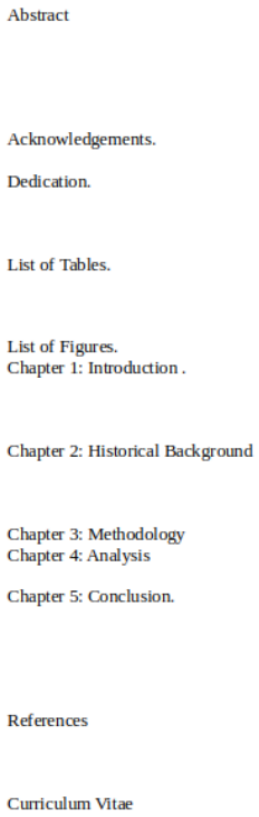
Con -- psm 3 retorna una cadena vacía
Con -- psm 10 retorna el caracter

PSM 11. Texto espaciado: Encuentre todo el texto posible sin un orden en particular.

PSM 11. Texto espaciado: Encuentre todo el texto posible sin un orden en particular. (útil para menús o tablas de contenidos)

--psm 3 (defecto)





--psm 11

```
Abstract

Acknowledgements.

Dedication.

List of Tables.

List of Figures.

Chapter 1: Introduction .

Chapter 2: Historical Background

Chapter 3: Methodology

Chapter 4: Analysis

Chapter 5: Conclusion.

References

Curriculum Vitae
```

PSM 12. Texto espaciado con OSD

PSM 13. Procesa la imagen cómo una sóla linea de texto. Se suele utilizar cuando todo lo demás falla. si el texto está muy recortado, si el texto se ha estilizado de alguna manera, o si se trata de una fuente que Tesseract no reconoce automáticamente.

Iteración n

- Mejorar calidad de las imágenes
- Eliminar ruido
- Recortar área de interés en la imagen
- …

Gracias por su atención

Carrera 9 No. 74-08 piso 9

60 1 3266600

Asobancaria Colombia

@asobancariaco

@Asobancaria

https://www.asobancaria.com