

Herramientas de Extracción

Dr. Cristian Enrique Muñoz Villalobos

Introducción

Existe una gran variedad de bibliotecas en python para la extracción de información. A continuación vamos a revisar las herramientas mas conocidas.

python-docx

Biblioteca para crear y actualizar archivos Microsoft Word (.docx)

Site: <https://python-docx.readthedocs.io/en/latest/>

```
>>> import docx  
  
>>> doc = docx.Document('demo.docx')  
  
>>> len(doc.paragraphs)  
7  
  
>>> doc.paragraphs[0].text  
'Document Title'  
  
>>> doc.paragraphs[1].text  
'A plain paragraph with some bold and some italic'
```

pdfminer

Herramienta para extracción de información en documentos PDF. Tiene excelentes funcionalidades para realizar parser y análisis.

```
from pdfminer.pdfinterp import PDFResourceManager, process_pdf
from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from cStringIO import StringIO

def convert_pdf(path):
    rsrcmgr = PDFResourceManager()
    retstr = StringIO()
    codec = 'utf-8'
    laparams = LAParams()

    with TextConverter(rsrcmgr, retstr, codec=codec, laparams=laparams) as device:
        with open(path, 'rb') as fp:
            process_pdf(rsrcmgr, device, fp)

    str = retstr.getvalue()
    retstr.close()
    return str

text = convert_pdf('demo.pdf')
```

pandas

Es una biblioteca de código abierto diseñada principalmente para trabajar con datos relacionales o anotados de forma fácil e intuitiva. Tienes muchos recursos para manipular datos textuales, numéricos y series temporales.

Documentación: https://pandas.pydata.org/docs/getting_started/index.html

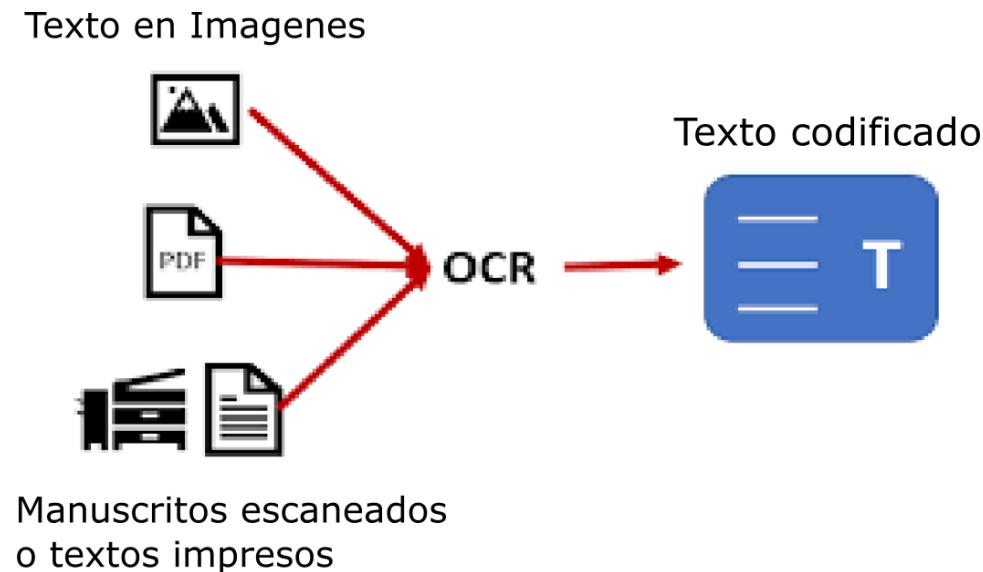
```
import pandas as pd  
  
df = pd.read_excel('comentarios.xlsx')  
  
df.head()
```

Herramientas de Extracción basadas en IA

Dr. Cristian Enrique Muñoz Villalobos

Reconocimiento Óptico del Carácteres

Los reconocedores ópticos de caracteres o *OCR-Optical Character Recognition* son una tecnología que transforma el texto renderizado en texto codificado entendible por el computador. Para esto se aplica técnicas de Visión Computacional (segmentación y detección de objetos). Los OCR modernos son basados en técnicas de Deep Learning usando redes convolucionales y redes recurrentes.





Extracción de información em documentos.



Reconocimiento de captchas

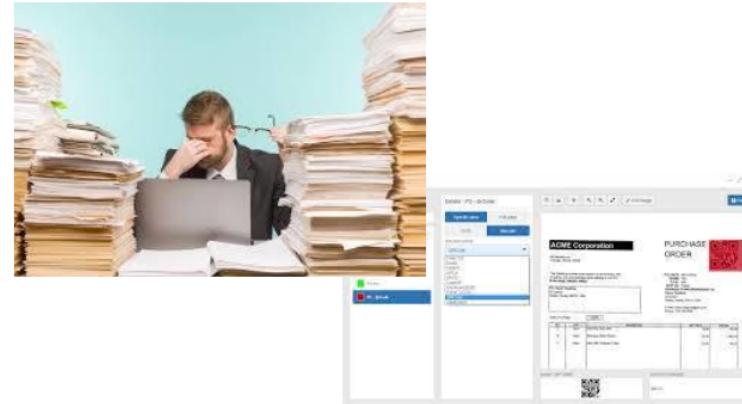


Reconocimiento de texto em ipad.



Detección de número de placas de carro.

Digitalización y Extracción de texto en documentos



Tarea 1 :

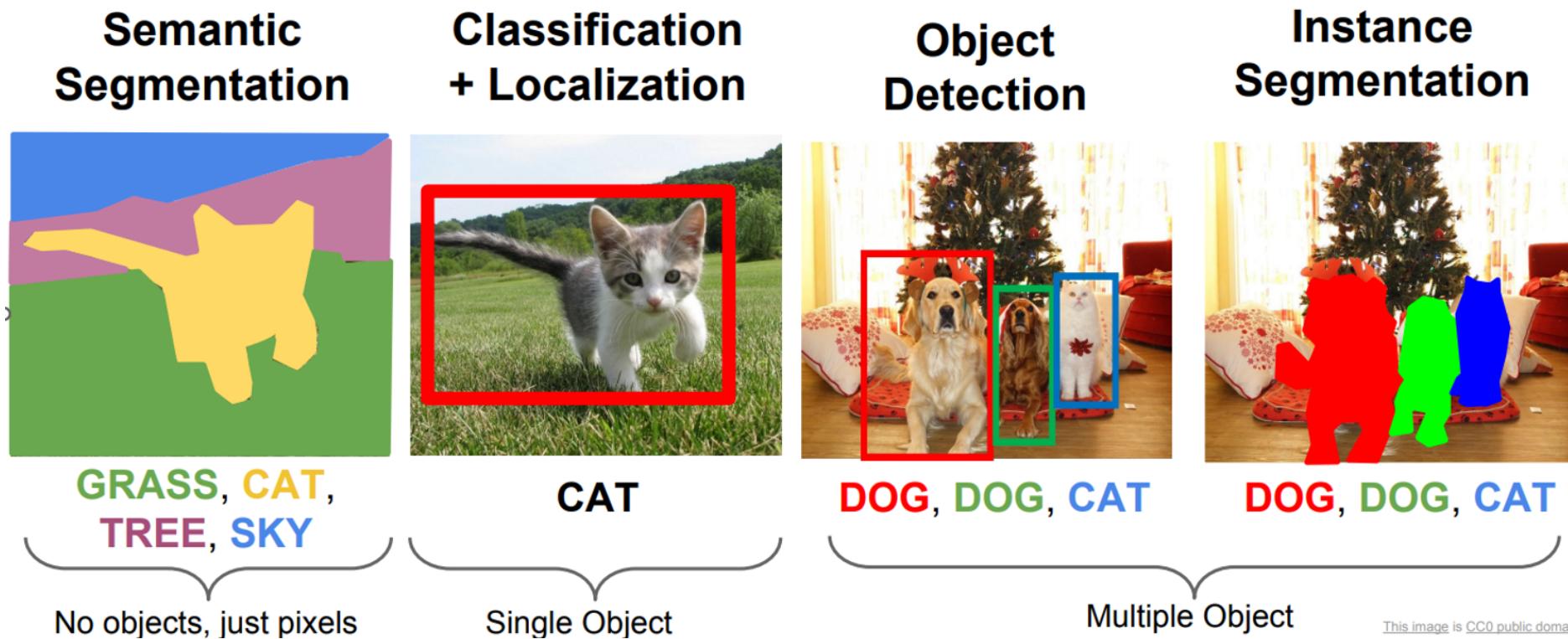
Extracción de texto en DNI usando Tesseract OCR: <https://tesseract-ocr.github.io/>.

Observación

El OCR es una herramienta genérica la cual nos puede ayudar en gran variedad de aplicaciones. OCRs, como Tesseract, realizan un proceso de detección de bloques de texto antes de aplicar el reconocedor de caracteres. Este detector puede perder precisión cuando es usado sobre páginas con gran complejidad de organización. Para estos casos, la detección de bloques de texto debe ser realizado por herramientas más especializadas.

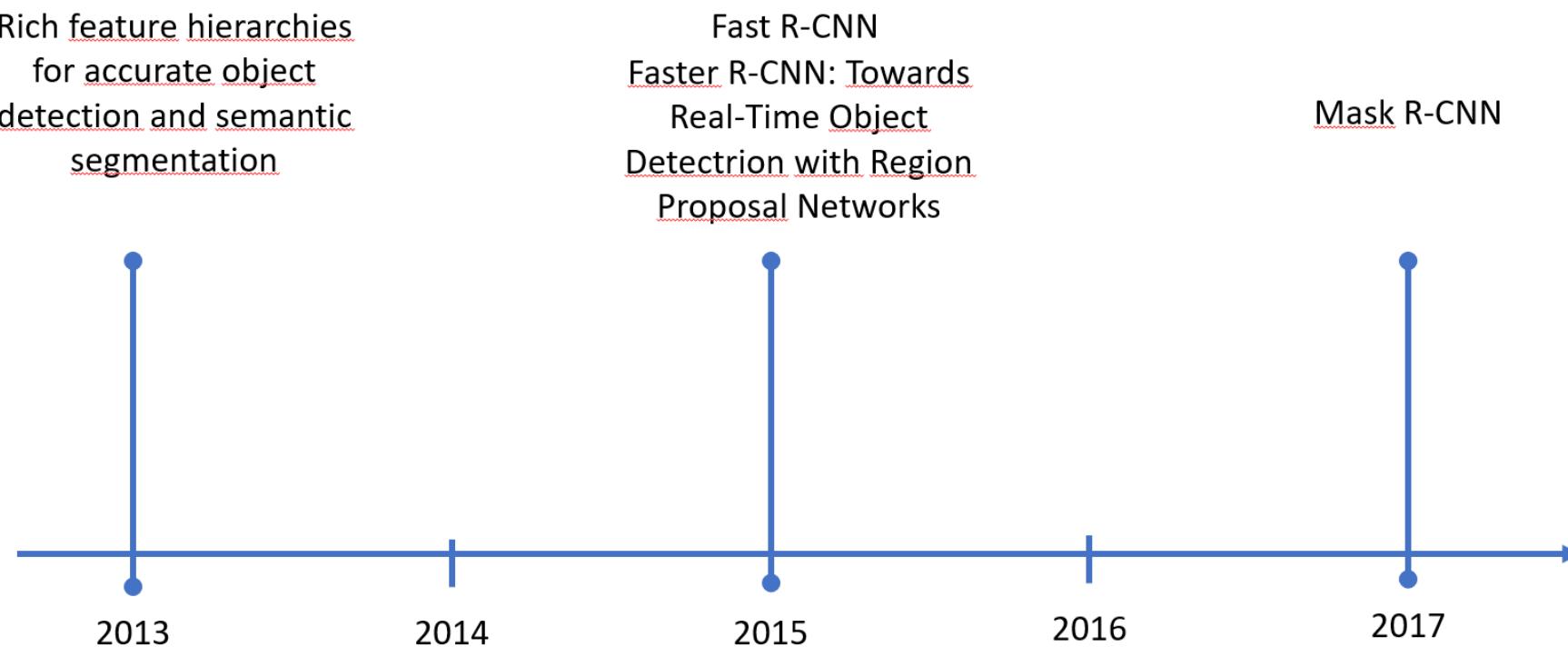
Detección de Objetos

La detección de objetos puede involucrar varios tipos de aprendizaje: Localización, Clasificación y Segmentación. Para dejar esto más claro, veamos ejemplos de cada tipo de tarea:

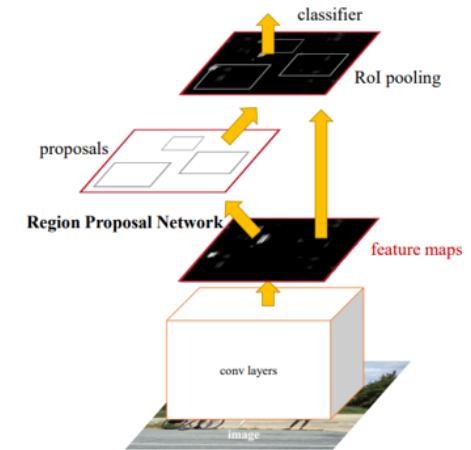
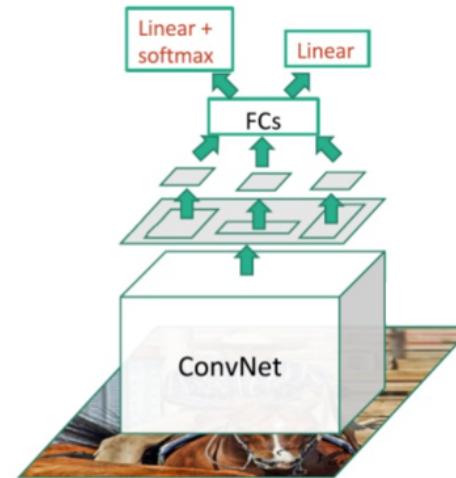
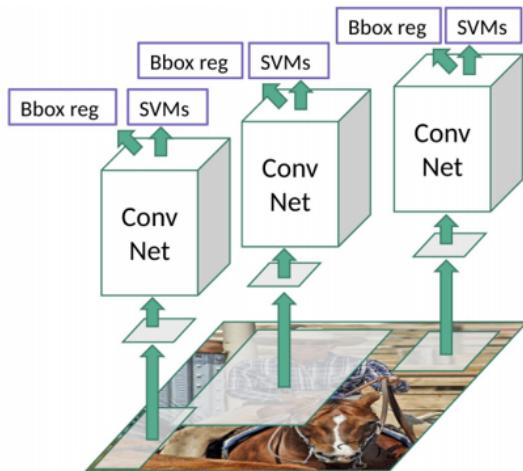


Detección de Objetos

La detección de objetos es una tarea de visión computacional que ha avanzado mucho en los años. Desde el año 2013 donde se utilizaban las tradicionales "Ingenieria de Atributos" hasta 2017 que se consigue automatizar el proceso de aprendizaje de inicio a fin (End to End Learning).

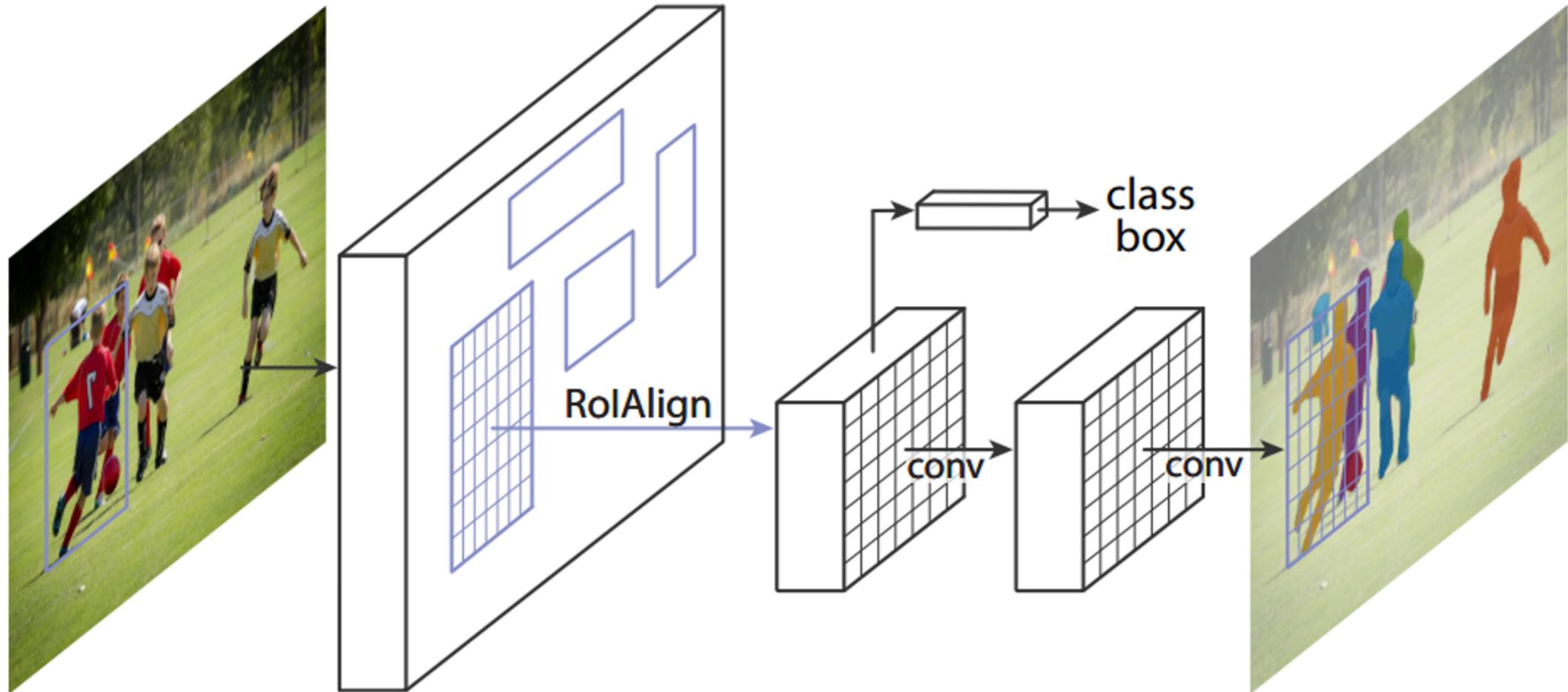


Detección de Objetos



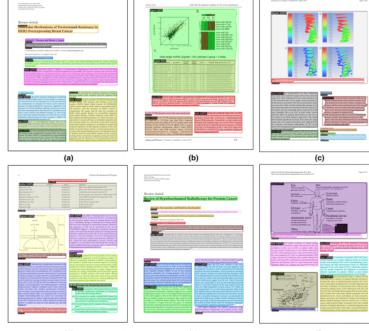
	R-CNN	Fast R-CNN	Faster R-CNN
Test time per image	50 segundos	2 segundos	0.2 segundos
Speed-up	1x	25x	250x
mAP (VOC 2007)	66.0%	66.9%	66.9%

Detección de Objetos (Mask R-CNN)



Construcción Detección de Objetos

La construcción de muchos detectores de objetos democratizados hoy en dia fueron posibles a la gran cantidad de datos anotados disponibles por la comunidad científica.

Objetos Comunes	Documentos	Escaneados
		

LayoutParser

<https://twitter.com/MelissaLDell/status/1380173067624845312?ctx=HHwWgMC4hfaFrqcmAAAA>



Melissa Dell
@MelissaLDell

...

(3/n) We are releasing an open-source deep-learning powered library, Layout Parser, that provides a variety of tools for automatically processing document image data at scale.

Webpage: layout-parser.github.io

Arxiv: arxiv.org/abs/2103.15348

Github:

Layout-Parser/layout-parser



A Unified Toolkit for Deep Learning Based Document Image Analysis

8 Contributors 71 Used by 3 Discussions 3k Stars 301 Forks

github.com GitHub - Layout-Parser/layout-parser: A Unified Toolkit for Deep Learning Base... A Unified Toolkit for Deep Learning Based Document Image Analysis - GitHub - Layout-Parser/layout-parser: A Unified Toolkit for Deep Learning Based ...

11:57 AM · 8 de abr de 2021 · Twitter Web App

342 Retweets 87 Tweets com comentário 1.449 Curtidas

LayoutParser

LayoutParser: A unified toolkit for Deep Learning Based Document Image Analysis

Con la ayuda de modelos de aprendizaje profundo de última generación, Layout Parser permite extraer estructuras de documentos complicadas utilizando solo varias líneas de código. Este método también es más robusto y generalizable ya que no hay reglas sofisticadas involucradas en este proceso.



Tarea 2 :

Detección de componentes de página.