

Herramientas de Extracción (Parte 2)

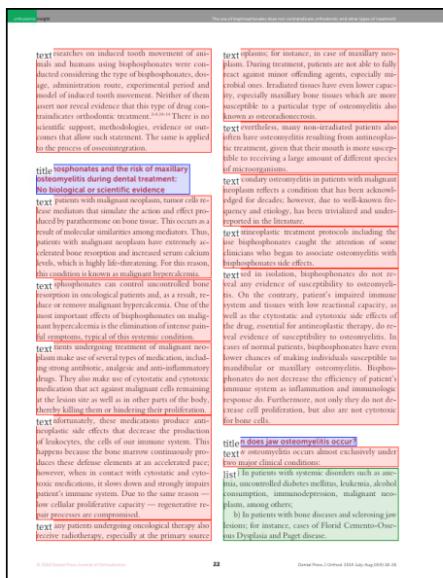
Dr. Cristian Enrique Muñoz Villalobos

Layout Parser

Modelos disponibles: <https://layout-parser.github.io/platform/>

Datasets

Para soportar diferentes estructuras de documentos. Layout Parser utiliza modelos entrenados com diferentes datasets. Actualmente los modelos fueron testados en 5 diferentes datasets.



PubPlayNet



HJDataset



PRImA



Newspaper Navigator

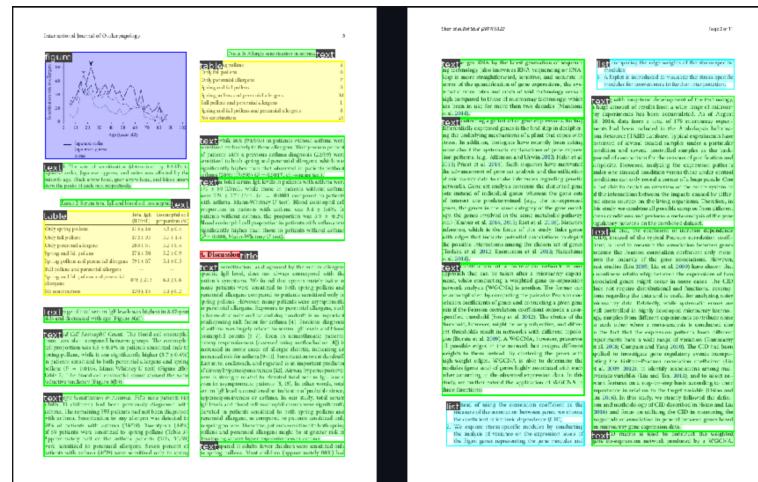
Térítési egység/Feljlesztés cél	4. Geometria	Orakert N. 22 óra I. 10 óra
Sokszögekkel, körrel kapcsolatos ismeretek. Ponttalomzók: nevezetes ponttalomzok ismertséte. Hálózatigazítás vonatával, pántigazítás. Hálózatigazítás, speciális hálózatok vonatkozó ismeretek. Egyszerűbb hálózatok. Hálózatok szülefüggvényei. Ekvivalens egyenletek. Előírások és másodlagos egyenletek. Kötésszerelesen egyenletek általánosítása. Alapvetészek, egyszerű szöveges feladatok körével. Tájékoztatókban előforduló geometriai szavak használata.		
Előzetes tudás		
A tantárgyhoz önműveltséget is feljelölő fejlesztési feladatak.		

TableBank

Publaynet

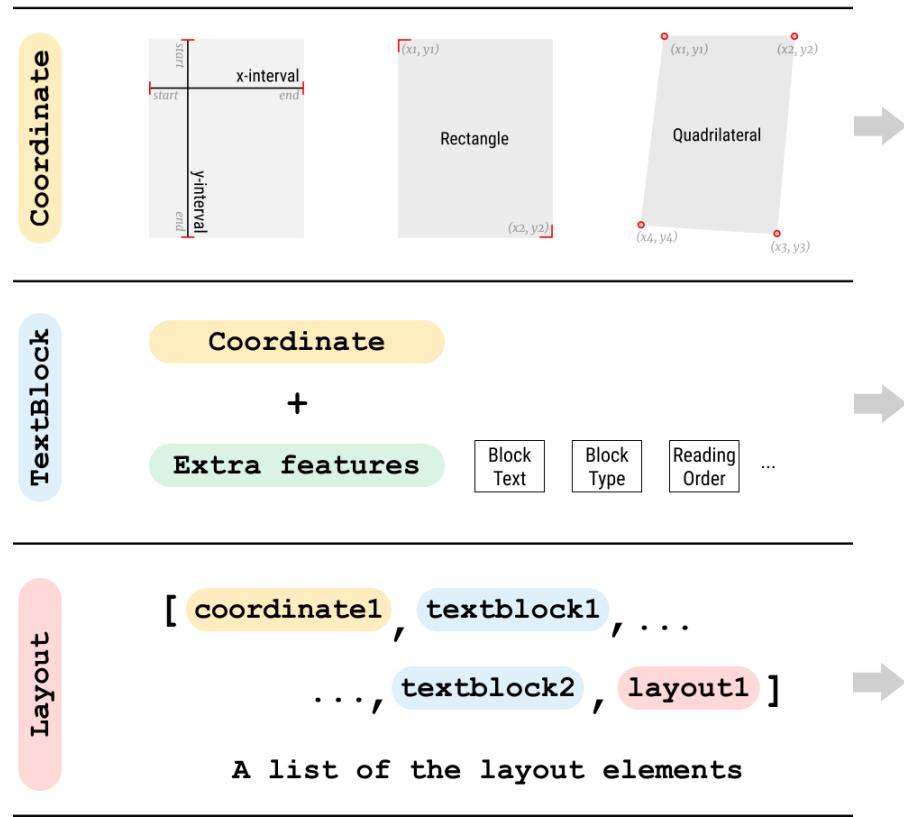
Publaynet es un dataset de imágenes de documentos, donde el template es anotado en bloques (*bounding boxes*) y máscaras poligonales (*polygonal segmentations*).

- La fuente del documento es: <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>
- La anotación es automáticamente generada unión el formato PDF com el formato XML de un subconjunto de artículos de PubMed.
- Paper: <https://arxiv.org/abs/1908.07836>



Layout Parser

Estructura de componentes de página:



The same transformation and operation APIs

Operation Name	Description
<code>block.pad(top, bottom, right, left)</code>	Enlarge the current block according to the input
<code>block.scale(fx, fy)</code>	Scale the current block given the ratio in x and y direction
<code>block.shift(dx, dy)</code>	Move the current block with the shift distances in x and y direction
<code>block1.is_in(block2)</code>	Whether block1 is inside of block2
<code>block1.intersect(block2)</code>	Return the intersected region of block1 and block2. Coordinate type to be determined based on the inputs.
<code>block1.union(block2)</code>	Return the union region of block1 and block2. Coordinate type to be determined based on the inputs.
<code>block1.relative_to(block2)</code>	Convert the absolute coordinates of block1 to relative coordinates to block2
<code>block1.condition_on(block2)</code>	Calculate the absolute coordinates of block1 given the canvas block2's absolute coordinates
<code>block.crop_image(image)</code>	Obtain the image segments in the block region

Tarea 1 :

Extraer información (text) de las imágenes extraídas de los PDF de proyectos de ley usando layout parser.

Expresiones Regulares

Expresiones Regulares (ER) son un conjunto de caracteres que especifican cierto patrón procurado en un texto, normalmente en lenguaje formal.

Por ejemplo como podemos buscar ...

- cachorro
- cachorros
- Cachorro
- Cachorros



Expresiones Regulares

Una de las operaciones principales de expresiones regulares es **la disyunción**, que permite escoger entre dos o más caracteres. Por ejemplo, si queremos encontrar la palabra cachorro o Cachorro, podemos representarla con la expression [cC]achorro donde el termino entre corchete representa la disyunción del carácter.

Expresión Regular	Encuentra
[cC]achorro	cachorro,Cachorro
[123456789]	Cualquier dígito

Expresión Regular

Algunas herramientas que nos pueden facilitar la tarea de construcción de expresiones regulares y explorar Regex:

- <https://regex101.com/>
- <https://regexr.com/>
- <https://unicode-table.com/es/>

Operadores: Disyunción

Disyunción ([]): Captura de un rango de valores.

Expresion Regular	Busca	Ejemplo
[A-Z]	una letra mayuscula	Cristian Muñoz
[a-z]	una letra minuscula	.. Si queremos encontrar...
[0-9]	Un único dígito	Parte 1: Introducción

Operadores: Disyunción

Negación (^): También se puede aplicar negación a una disyunción. Esto nos permite identificar cual carácter o rango de caracteres **no deseamos encontrar en el texto**.

Expresion Regular	Busca	Ejemplo
[^A-Z]	Una letra no mayuscula	Cristian Muñoz
[^0-9]	Un caracter (no dígito)	
[x^y]	x^y	En la ecuación z=x^y

El acento circunflexo solo indica la negación solo si es el primer carácter dentro de la disyunción.

Uso del operador Disyunción

Una barra vertical "|" también representa una disyunción. En este caso la disyunción es entre palabras. Por ejemplo, si queremos buscar la palabra "crear" y "cachorro", al mismo tiempo, podemos usar la expression: `crear|cachorro` .

Otros ejemplos:

- Caso básico: `tuyo|mio`
- Multiples situaciones: `x|y|c`
- Mezclando operadores: `[cC]rear|[cC]achorro`

Otros Operadores

Existe un conjunto de caracteres especiales que son muy importantes en expresiones regulares (?., +, *)

Operador	Descripción	Ejemplo (ER)	Encuentra
?	opcional	cachorro?	cachorro y cachorros
.	cualquier caracter	cas.	casa, caso, casi, etc.
+	por lo menos 1 vez	a+	a,aa,aaa,aaa,...
*	cualquier cantidad.	k*	,k,kk,...

Otros Operadores

Existe tambien "metacaracteres" quem nos ayudan a capturar caracteres de una palabra.

\w : Encuentra un caracter a-z,A-Z,0-9,_

\d : Encuentra un dígito 0-9

obs 1:

\ : Actua sobre el caracter que lo procede. Puede ser usado para referenciar metacaracteres como para tambien usar caracteres reservados (\?,\.,\+).

obs 2:

() : Puede usarse para agrupar un elemento, asi los operadores (+,?,...) pueden actuar sobre todo el agrupamiento.

Tarea 2 :

Extraer secciones de interes de los proyectos de ley descargados.

Bibliografia

- Documentación Biblioteca Layout Parser: <https://layout-parser.readthedocs.io/en/latest/>
- Biblioteca Detectron2 (Facebook): <https://detectron2.readthedocs.io/en/latest/>
- PublayNet (artigo): <https://arxiv.org/abs/1908.07836>
- Expressiones Regulares: [Regular Expression \[Pocket Reference\]](#), Tony Stubblebine, 2007.