

Combinación de texto e información visual

Texto plano

Texto plano extraído con pypdf

Separar texto en palabras
(txt_words)

Información visual

Texto extraído con OCR
y reconstruido

Separar texto en palabras
(ocr_words)

Obtener lista de diferencias
con difflib

Sin diferencia

Capítulo
Capítulo

Capítulo

Palabra sin correspondencia al inicio

- 3
DICTAMEN

DICTAMEN

Palabra separada

+ PRESENTE
- P
- R
- E
- S
- E
- N
- T
- E

PRESENTE

Palabras cortadas

- Constituciona
+ Constituciona-
? +
- -

Constituciona-

Palabra sin correspondencia

- 1.
La

1. La

Palabra diferente

- II. - 5.
? ^ + s.
+ IT.
? ^

II.

5.

Remociones sin correspondencia

- Decreto
- por
- el

Agregar a lista de
pendientes

Adiciones sin correspondencia

+ Decreto
+ por
+ el

Agregar a lista de
pendientes

Alinear lista de pendientes donde más palabras coincidan

- el
- Decreto + Decreto
- por + por
- el + el



Decreto
por
el