

ENCAPSULADOR DE CONTENIDOS EN ARCHIVOS DE TEXTO CON PYTHON

AUTOR: RAÚL SERRANO CAMPILLO

TUTOR: VERA POSPELOVA




OBJETIVOS

Objetivo principal:

Crear una estructura en Python que encapsule el árbol de contenidos de un archivo PDF.

Objetivos secundarios:

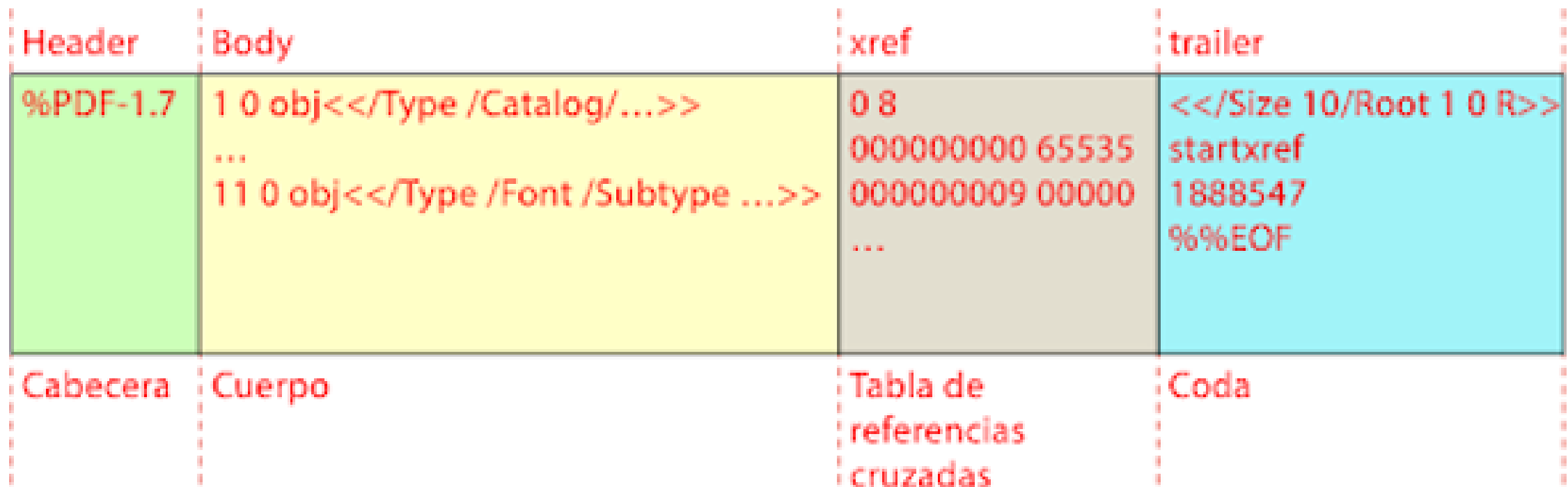
- Estructurar los objetos en los que se divide el árbol de contenidos.
 - Ubicar los elementos de mayor importancia.
 - Almacenar las etiquetas finales de la estructura (Instrucciones).
 - Generar un nuevo archivo TXT modificado.
- 

MARCO TEÓRICO



ESTRUCTURA GENERAL

- **Cabecera:** Indica la extensión y la versión del archivo.
- **Cuerpo:** Almacena los contenidos reales y está dividida en objetos.
- **Tabla de referencia cruzada:** Permite localizar los objetos rápidamente.
- **Tráiler:** Contiene información de la estructura e indica donde se encuentra la tabla de referencia



- De PDF a TXT



MARCO TEÓRICO

ESTRUCTURA DEL CUERPO: OBJETOS

```
1 0 obj
<<
/Lang (en-US)
/MarkInfo <<
/Marked true
/Suspects false
>>
/Metadata 2 0 R
/Outlines 3 0 R
/Pages 4 0 R
/StructTreeRoot 5 0 R
/Type /Catalog
/ViewerPreferences <<
/DisplayDocTitle true
>>
>>
endobj
```

] Identificador

] Diccionario

] Final

- Estructura de árbol
- Tipos de valores:
 - Básicos: números, strings, claves, bool, null
 - Complejos: listas, diccionarios y streams

MARCO TEÓRICO

TIPOS DE OBJETOS

```
1 0 obj
<<
/Lang (en-US)
/MarkInfo <<
/Marked true
/Suspects false
>>
/Metadata 2 0 R
/Outlines 3 0 R
/Pages 4 0 R
/StructTreeRoot 5 0 R
/Type /Catalog
/ViewerPreferences <<
/DisplayDocTitle true
>>
>>
endobj
```

- Catalog
- Pages
- Page
- Content
- Resources
- Font
- XObject

MARCO TEÓRICO

ESTRUCTURA DE ÁRBOL

```
1 0 obj
<<
/Lang (en-US)
/MarkInfo <<
/Marked true
/Suspects false
>>
/Metadata 2 0 R
/Outlines 3 0 R
/Pages 4 0 R
/StructTreeRoot 5 0 R
/Type /Catalog
/ViewerPreferences <<
/DisplayDocTitle true
>>
>>
endobj
```

```
5 0 obj
<<
/K 11 0 R
/ParentTree 12 0 R
/ParentTreeNextKey 20
/Type /StructTreeRoot
>>
endobj
```

```
11 0 obj
<<
/K [33 0 R 34 0 R 35 0 R 36 0 R 37 0 R 38 0 R 39 0 R 40 0 R 41 0 R 42 0 R
43 0 R 44 0 R 45 0 R 46 0 R 47 0 R 48 0 R 49 0 R 50 0 R 51 0 R 52 0 R]
/P 5 0 R
/S /Document
>>
endobj
```

MARCO TEÓRICO



OBJETOS DE CONTENIDO

```
1170 0 obj
<<
/Length 28871
>>
stream
BT
/P <</MCID 1 >>BDC
/T1_0 17.215 Tf
249.201 665.282 Td
(The)Tj
/TT0 1 Tf
27.252 0 Td
( )Tj
/T1_1 17.215 Tf
5.229 0 Td
[(axessibilit)26 (y)]TJ
/TT0 1 Tf
71.168 0 Td
( )Tj
/T1_0 17.215 Tf
5.541 0 Td
[(pac)26 (k)52 (age)]TJ
EMC
/P <</MCID 6 >>BDC
/T1_2 11.955 Tf
-198.72 -28.892 Td
(Dragan)Tj
EMC
```

Objeto de contenido

Instrucciones

```
/P <</MCID 25 >>BDC
/T1_2 11.955 Tf
3.909 0 Td
(Capietto)Tj
EMC
```

Ejemplo 1

```
/P <</MCID 28 >>BDC
/T1_2 11.955 Tf
T*
(,)Tj
EMC
```

Ejemplo 2

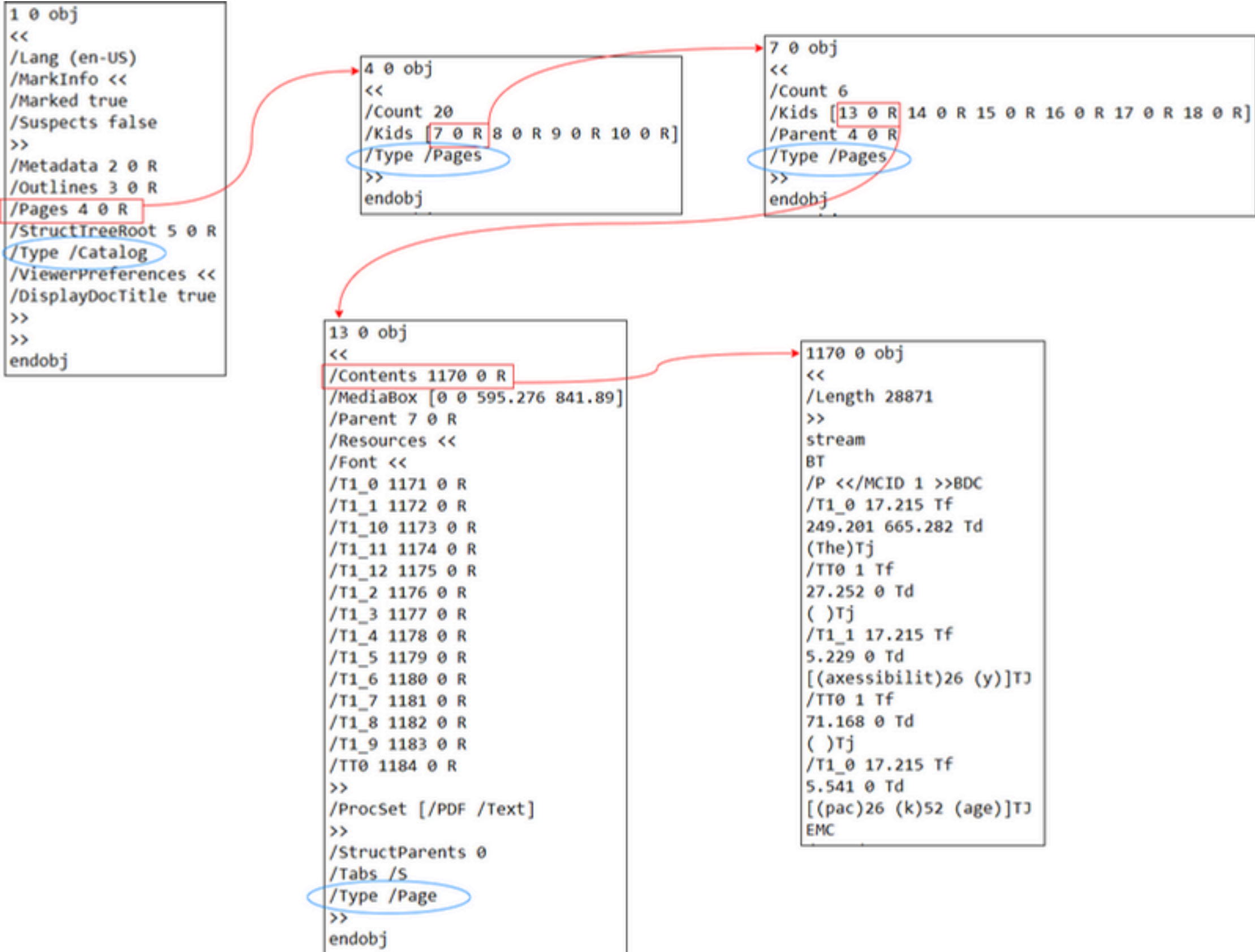
```
/P <</MCID 8 >>BDC
/T1_2 11.955 Tf
3.897 0 Td
[(Ahmeto)27.001 (vic)]TJ
EMC
```

Ejemplo 3



MARCO TEÓRICO

FLUJO DEL ÁRBOL



DESARROLLO

ENCAPSULACIÓN DEL CONTENIDO

01

PDF → TXT

Expresiones regulares

02

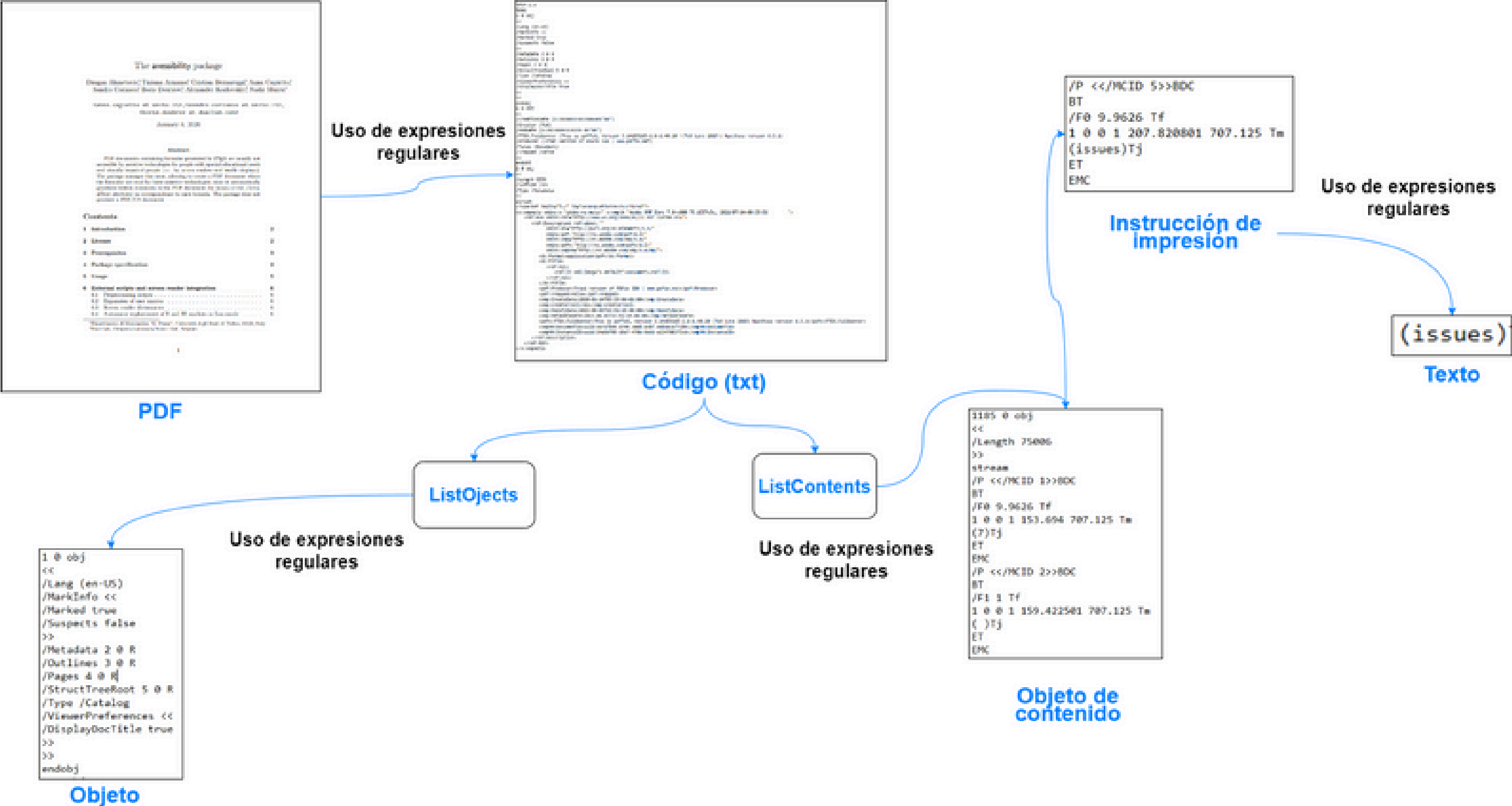
Encapsular objetos (id, referencias, kids, instrucciones)

03

Encapsular instrucciones (id, fuente, tamaño, posición, texto)

DESARROLLO

ENCAPSULACIÓN DEL CONTENIDO



DESARROLLO

FUNCIONAMIENTO

Menú de opciones

1. Imprimir el árbol de estructura del PDF
2. Imprimir una página del PDF
3. Imprimir el PDF entero
4. Imprimir un texto por id
5. Imprimir referencias de un objeto
6. Modificar objeto
7. Descargar PDF
8. Gestionar los hijos de un objeto
9. Salir


Selecciona una opción (1-9):

CONCLUSIONES Y TRABAJOS FUTUROS

Conclusiones:

- Se ha creado una estructura capaz de encapsular los contenidos de un PDF.
- La estructura permite acceder a las referencias de los objetos y navegar por el árbol.
- Se permite modificar la estructura generando un nuevo TXT.

Trabajos futuros:

- Obtener imágenes al igual que se hace con los textos
 - Imprimir los contenidos usando el desplazamiento horizontal y vertical
- 

ENCAPSULADOR DE CONTENIDOS EN ARCHIVOS DE TEXTO CON PYTHON

AUTOR: RAÚL SERRANO CAMPILLO

TUTOR: VERA POSPELOVA



Universidad
de Alcalá

