Clasificación de tipos de documento

Ignacio Ramírez

13 de julio de 2020

1. Descripción del problema

Se trata de clasificar (agrupar) los documentos según diferentes *tipos*, pero sin tener esos tipos identificados de antemano, es decir, es un problema de agrupación no supervisada (o clustering).

El lograr una buena agrupación facilitaría muchísimo otros aspectos del procesamiento previo a LUISA así como del posterior de los resultados.

La figura 1 muestra algunos ejemplos de distintos posibles tipos de documentos. Algunos son cartas, otros son formularios, otros son recortes de diario, etc.

2. Consideraciones técnicas

Este es un problema abierto y como tal tiene mucho trabajo de investigación, ensayo y error.

La idea es que el método sea razonablemente rápido. Las formas en que hemos atacado este problema hasta el momento (sin una solución elegida en particular) es identificar rasgos generales de los documentos a nivel visual, como ser *textura* o variaciones globales de la intensidad a lo largo del documento. Ideas como la de los perfiles de intensidad de las filas que se aplican para alinear pueden servir aquí. Hay que ser creativos.

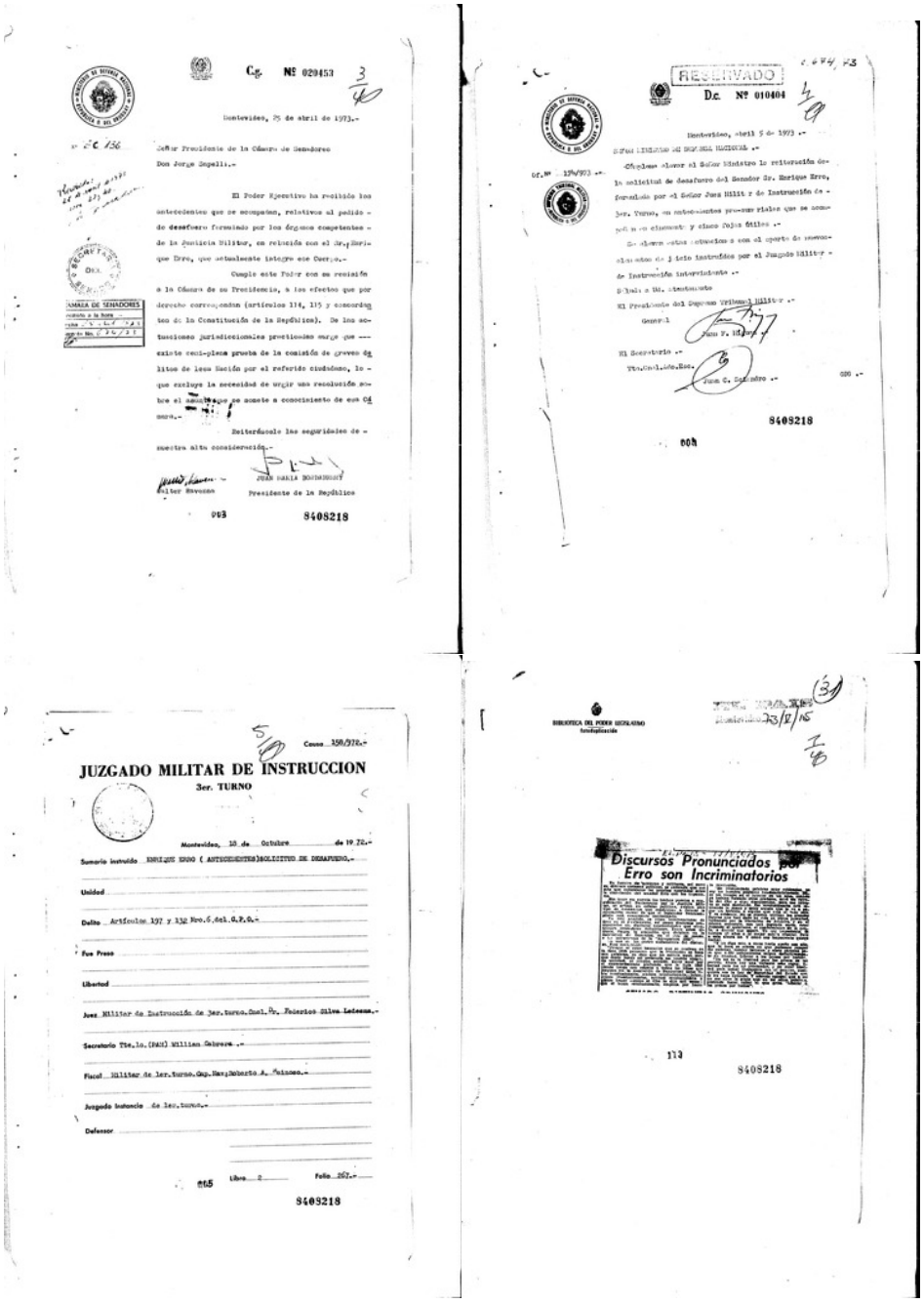

Figura 1: Ejemplos de distintos tipos de documento (hay muchísimos más!). La idea es identificar el tipo de acuerdo a su *aspecto general* más que a su contenido textual, que no está disponible.