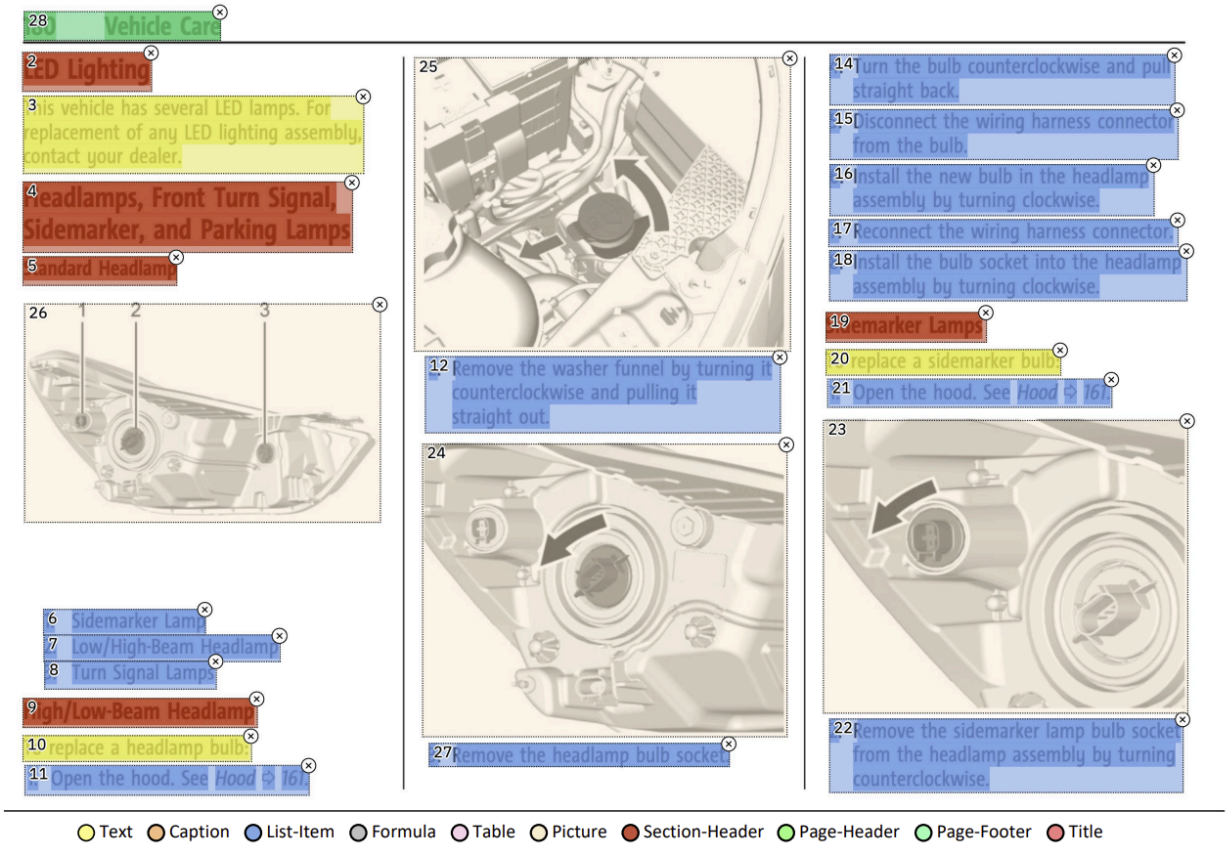
Propuesta de proyecto de Fin de Carrera

Análisis automático de estructura de documentos

Resumen

El presente proyecto trata del estudio, implementación y evaluación sistemática de métodos de análisis de documentos (en inglés, *Document Layout Analysis*). El cometido principal de dichos métodos es, dado un documento, identificar sus partes y la relación lógica entre ellos: títulos, columnas, párrafos, encabezados, pie, gráficos/figuras, imágenes, tablas, etc., como puede verse en el siguiente ejemplo:



Objetivos

Se estudiarán los métodos existentes en la literatura, se implementará aquellos relevantes que no tengan implementación disponible, y se evaluarán sobre dos conjuntos de datos: a) conjuntos de evaluación de uso académico, público, como PubLayNet y b) los documentos del Proyecto Cruzar.uy, que incluyen el Archivo Berruti y otros archivos digitalizados de organismos represivos que operaron durante la dictadura militar en Uruguay (1973-1984).

Antecedentes

A nivel global, el problema de *Document Layout Analysis* (DLA) es tan antiguo como complejo, lejos de estar cerrado a nivel académico e industrial, y paso fundamental en la transcripción e interpretación automática de documentos. Existe muchísima literatura al respecto, e involucra varios sub problemas al punto de que su evaluación suele se restringida a ciertas familias o tipos de documentos.

Desde el punto de vista local, la motivación surge del proyecto Cruzar.uy, que consiste en el análisis de varios millones de documentos que deben ser procesados de manera automática, rápida y eficaz. La dificultad adicional que dicho conjunto de datos plantea con respecto a bases de datos públicas es el estado de degradación de los documentos, a veces muy avanzado, lo cual dificulta enormemente el análisis. Además de ello, existen ciertos tipos de estructuras y documentos únicos o muy particulares en este conjunto de datos como ser: fichas personales, tablas, formularios complejos.

Debido a lo anterior, los objetivos de este proyecto serán apropiadamente delimitados de manera de poder generar una comparación útil y fiable, y en base a ella seleccionar e implementar un método que pueda ser aplicado en los archivos del proyecto Cruzar.uy.

Resultados esperados

La finalización del proyecto implica lograr dos objetivos bien concretos:

1. Una evaluación sistemática de los métodos de DLA más relevantes de la literatura
2. La implementación correcta, eficiente y con interfaz de línea de comandos de un método seleccionado para su aplicación en los archivos de Cruzar.uy

Interesados dirigirse a Ignacio Ramírez, nacho@fing.edu.uy