Propuesta de Tesis

Maestría en Ciencias de Datos

- **Área**: Procesamiento de imágenes
- **Subárea**: Procesamiento y análisis de imágenes de documentos de texto
- **Tema**: Limpieza, alineamiento y rectificación de documentos escaneados
- **Tutor**: Ignacio Ramírez
- **Formación específica recomendada**:
    - **Competencias**:
        - Python
        - Aprendizaje Automático
    - **Cursos**:
        - Tratamiento de Imágenes por Computadora
        - Aprendizaje Automático
        - DLVIS

Resumen de la propuesta:

El estudiante deberá familiarizarse y ponerse al día con los métodos existentes de preprocesamiento de documentos que hacen a las primeras etapas en el Análisis de Estructura de Documentos (Document Layout Analysis). Deberá hacer una revisión exhaustiva de la bibliografía, e implementar por lo menos tres métodos de referencia en el área. La implementación deberá ser realizada en el lenguaje Python, de modo de poder integrarse a las herramientas actualmente existentes en el proyecto LUISA/Cruzar.

Descripción

El Archivo Berruti (AB) consiste en más de 2 millones de imágenes de documentos generados por el aparato represivo antes, durante y después de la dictadura uruguaya (período 1972--1991). Las imágenes no son documentos escaneados comunes, sino escaneos de microfilms (rollos de negativos fotográficos diminutos) que a su vez eran fotografías de documentos físicos. Las imágenes tienen la particularidad de ser binarias, es decir, los pixeles son blanco o negro; no hay tonos de gris.

El proyecto Cruzar es un esfuerzo colectivo de docentes de UdelaR, así como de algunos actores externos, cuyo objetivo es la interpretación y análisis histórico del pasado reciente en Uruguay. El proyecto LUISA/Cruzar es parte de este macroproyecto, y consiste en los aspectos de procesamiento primario de los datos previo a su transcripción mediante herramientas como LUISA (interfaz de transcripción colaborativa) o sistemas automáticos de transcripción de texto (OCR).

Previo a la transcripción se realiza una serie de etapas de preprocesamiento. Las primeras tienen que ver con el acondicionamiento de la imagen. Esto incluye:

1. Limpieza de manchas, marcas de polvo, ruido en general
2. Eliminación de bordes
3. Corrección de iluminación no uniforme (sombras, etc.)
4. Enderezado del texto. Esto puede ser simplemente rotar la página para que el texto quede perfectamente horizontal, o bien corregir deformaciones debidas a la perspectiva (por ejemplo, cuando el texto llega al borde del encuadernado de un libro, las lineas se curvan)

Existe muchísima bibliografía al respecto de estos problemas (nada triviales). La idea es rever esa bibliografía e implementar y comparar por lo menos un par de métodos de dos de esas etapas. En este proyecto estamos interesados especialmente en la etapa de limpieza y del enderezado de texto (1 y 4), ya que el problema 3 no suele estar presente y el 2 es fácil de corregir en nuestro caso.