Propuesta de Tesis

Maestría en Ciencias de Datos

- **Área**: Procesamiento de imágenes
- **Subárea**: Procesamiento y análisis de imágenes de documentos de texto
- **Tema**: Análisis de estructura de documentos del Archivo Berruti
- **Tutor**: Ignacio Ramírez (nacho@fing.edu.uy)
- **Formación específica recomendada**:
    - **Competencias**:
        - Python
        - Aprendizaje Automático
    - **Cursos**:
        - Tratamiento de Imágenes por Computadora
        - Aprendizaje Automático
        - DLVIS

Resumen de la propuesta:

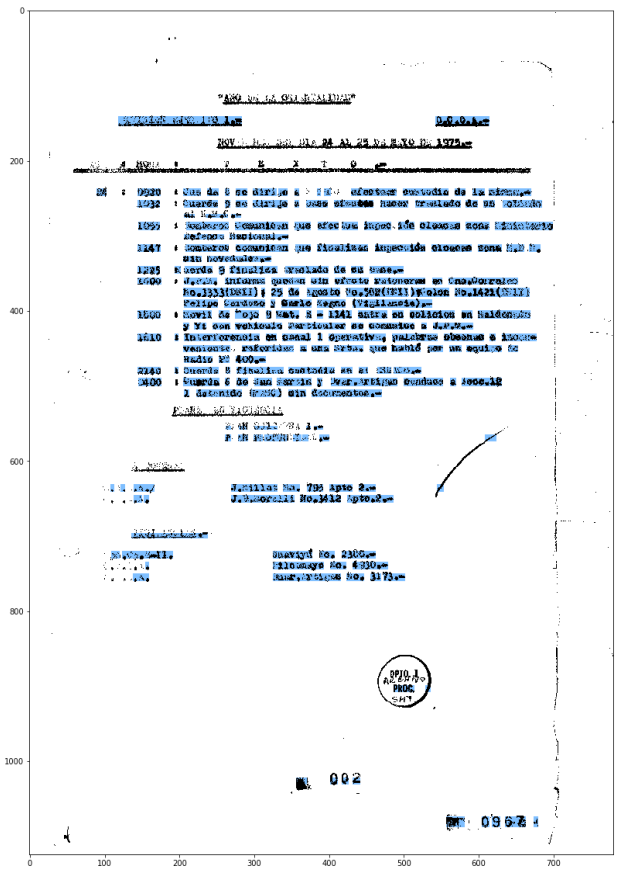
El estudiante deberá familiarizarse y ponerse al día con los métodos existentes de Análisis de Estructura de Documentos (Document Layout Analysis). Deberá hacer una revisión exhaustiva de la bibliografía, e implementar por lo menos tres métodos de referencia en el área. La implementación deberá ser realizada 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).

La etapa previa a la transcripción, ya sea usando LUISA o un OCR, es aislar, segmentar y dar un orden lógico (separar) al texto de una página. De esa manera, luego puede reconstruirse el documento a partir de las transcripciones de las distintas palabras. El siguiente ejemplo es un caso (imperfecto) de identificación de palabras en un documento.



Las técnicas que hacen al procesamiento e identificación de componentes (cuerpo, columnas, filas, recuadros, palabras, etc.) en un documento escaneado se engloban en el área conocida como Document Layout Analysis (DLA). La bibliografía incluye por lo menos dos libros completos sobre el tema los cuales pueden (el más reciente de 2019) y se recomienda que sean la base del estudio en las etapas iniciales de la maestría del estudiante, y de trabajos posteriores.

Algunas Referencias

Shafait, F. (2008). *Geometric Layout Analysis of Scanned Documents*. 163. http://kluedo.ub.uni-kl.de/files/2008/shafait-dissertation.pdf

Breuel, T. M. (2003). High Performance Document Layout Analysis. *Proceedings 2003 Symposium on Document Image Understanding Technology*, *03*(May 2003), 209–218.

http://books.google.com/books?hl=en&lr=&id=Rw7f-vuaX7IC&oi=fnd&pg=PA209&dq=High+performance+document+layout+analysis&ots=LtngWLoU_a&sig=nP4wmQveoTytxQIy2aUdBgKj0mQ

Doermann, D., & Tombre, K. (2014). Handbook of Document Image Processing and Recognition. In *Handbook of Document Image Processing and Recognition*. https://doi.org/10.1007/978-0-85729-859-1