Sight & Science: Vision 2020

Detecting Complex Text Layouts of Newspaper Data

Principal Investigator: Melissa Dell, PhD. Faculty of Arts Sciences

This project will use recent advances in deep learning to develop innovative methods for inferring complex document layouts. These will be used to build a full-text database of historical newspaper content for American communities across the 19th and 20th centuries. Existing historical newspaper databases typically do not provide information on text structures. This is because the off-the-shelf OCR (Optical Character Recognition) tools used to digitize historical newspaper scans are incapable of detecting the layouts of complex texts. They can recognize words in image scans of historical newspapers, often with high levels of noise, but cannot detect how these words are combined into sentences, paragraphs, and articles. This makes these documents inaccessible to the visually impaired. An analogous failure of off-the-shelf tools to detect layouts applies not only to newspapers but to many other structured and semi-structured documents, including websites. The layout detection methods that we develop will combine external memory learning and multi-task learning to recognize complex layouts while reducing the amount of costly data labeling that is required. The broad project aims are twofold: 1) to develop an open source document layout analysis pipeline that will contribute to research advances in layout analysis more generally, aiding the development of facilitative technologies for the visually impaired, and 2) to make a structured, searchable database of newspaper content available to the visually impaired.