Issues in Data Capture from Historical Government Records

2023-10-07 16:59:56

Authors

Enda O Shea

Abstract

Segmenting, detecting, and capturing historical handwritten data from historic government records presents a myriad of challenges. First, the age and degradation of these documents often result in faded ink, smudges, tears, and stains, which complicates the extraction process. Machine learning algorithms and traditional OCR tools often struggle to recognize document structures, line separators, and character features from degraded sources. Additionally, historical records often employ intricate handwriting styles that vary widely across multiple authors, and may contain ligatures, abbreviations, and symbols unfamiliar to modern systems, presenting obstacles in accurate interpretation. Addressing these challenges necessitates the development of specialized tools and algorithms tailored to the unique properties of historical data, as well as comprehensive training sets that account for such diversity. The preservation and digitization of these records is crucial, as they represent invaluable historical assets, but doing so demands innovative solutions to surmount the inherent difficulties.

Keywords

--