New algorithm searches historic documents to discover noteworthy people

October 14 2021, by Kevin Manne

Credit: Pixabay/CC0 Public Domain

Old newspapers provide a window into our past, and a new algorithm co-developed by a University at Buffalo School of Management researcher is helping turn those historic documents into useful, searchable data.

Published in Decision Support Systems, the algorithm can find and rank people's names in order of importance from the results produced by optical character recognition (OCR), the computerized method of converting scanned documents into text that is often messy.

"It's a known fact that when OCR software is run, very often the text gets garbled," says Haimonti Dutta, Ph.D., assistant professor of management science and systems in the UB School of Management. "With old newspapers, books and magazines, problems can arise from poor ink quality, crumpled or torn paper, or even unusual page layouts the software isn't expecting."

To develop the algorithm, the researchers partnered with the New York Public Library (NYPL) and analyzed more than 14,000 articles from New York City newspaper The Sun published during November and December of 1894. The NYPL has scanned more than 200,000 newspaper pages as part of Chronicling America, an initiative of the National Endowment for Humanities and the Library of Congress that is working to develop an online, searchable database of historical newspapers from 1777 to 1963.

Their algorithm ranks people's names by importance based on a number of attributes, including the context of the name, title before the name, article length and how frequently the name was mentioned in an article.

The algorithm learns these attributes only from the text—it does not rely on external sources of information such as Wikipedia or other knowledgebases. But since the OCR text is garbled, it can't determine how effective these attributes are for ranking people's names. So the researchers used statistical measures to model the many data attributes, which helped provide the desired ranking of names.

The researchers used two sets of the historic articles to test their

algorithm: One set was the raw text produced from the OCR software, the other set had been cleaned up manually by New York City schoolchildren, who are using the articles to write biographies of local, notable people of the time.

When compared to the cleaned-up versions of the stories, the ranking algorithm is able to sort people's names with a high degree of precision even from the noisy OCR text.

Dutta says their process has wide reaching implications for discovering important people throughout history.

"We recently used this technique on African American literature from the Civil War to learn more about the important people during the era of slavery," says Dutta. "Going forward, we'll be expanding the technique to examine relationships between people and build out the social networks of the past."

More information: Haimonti Dutta et al, PNRank: Unsupervised ranking of person name entities from noisy OCR text, *Decision Support Systems* (2021). DOI: 10.1016/j.dss.2021.113662

Provided by University at Buffalo