Page Number Extraction On Scanned Books

Tanaya M. Asnani with Prof. James Allan and John Foley

College of Information and Computer Sciences
University of Massachusetts Amherst

Introduction

Old books are scanned by **Optical Character Recognition** to create their electronic format for preservation and archiving purposes. While scanning, page number recognition is prone to a high variety of errors.

- Pages of a book may be skipped by human error or ripped out
- OCR software might interpret numbers incorrectly
- Page numbers also come in different numbering styles(e.g roman numerals etc.)

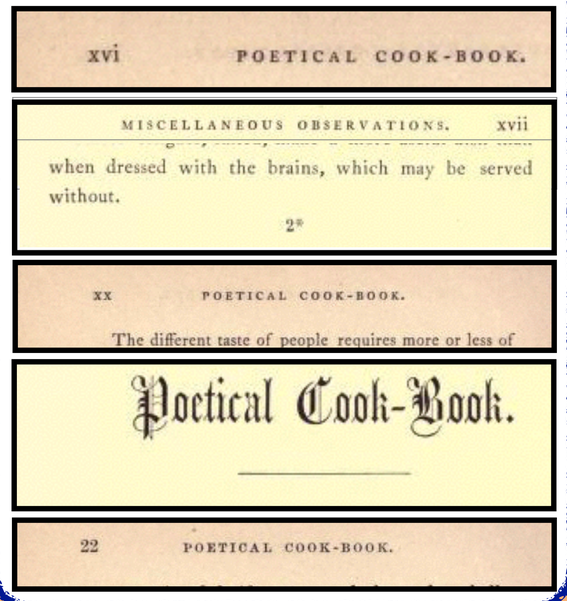
All these factors increase the complexity of this problem.

In 2008, Déjean & Meunier explored this problem on well and consistently formatted books only.

Déjean, Hervé, and Jean-Luc Meunier. "Versatile page numbering analysis." *Electronic Imaging 2008*. International Society for Optics and Photonics, 2008.

Visual examples of a problematic sequence of five consecutive pages from a book:

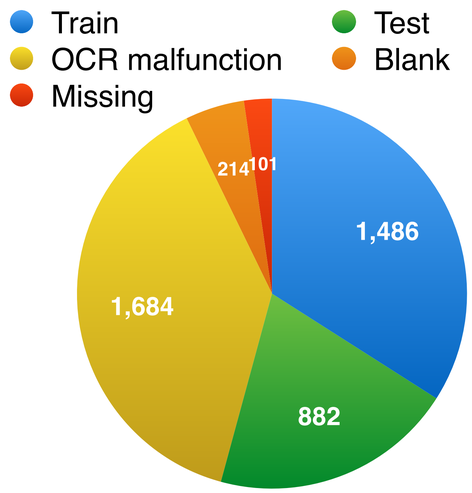
16(xvi), 17(xvii) or 2*, [18? 19?], 20(xx), blank, 22



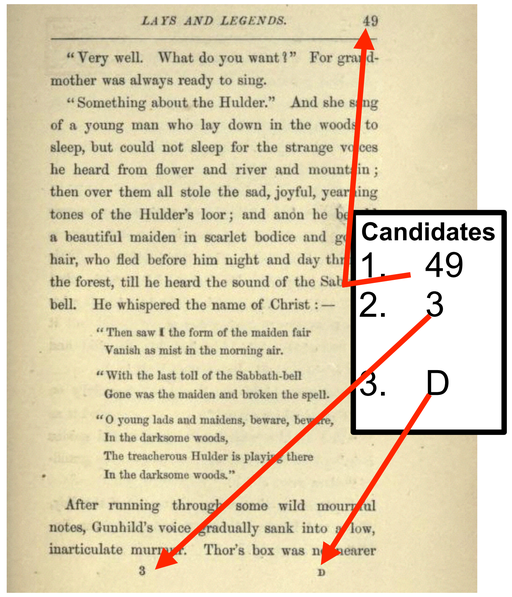
Algorithmic Approach

- Uses ML approach with svmrank
- Create parser to extract page number candidates
- Extract features for each candidate
- Generate train & test splits on 25,000 candidates
- Pass feature & candidate data to learning algorithm
- Feed test data to rank classifier to get best ranked candidate for *each page*
- Obtain accuracy of best ranked candidates

Page distribution in dataset



Evaluation Approach



For each book, Overall Accuracy @ N = proportion of pages with correct answer at rank 1 to N

- To understand misclassification, look at missed correct candidates appearing in ranks 2 to 5
- During development, modify features /create special cases so correct candidates can be moved up to rank 1.

Features used in various models

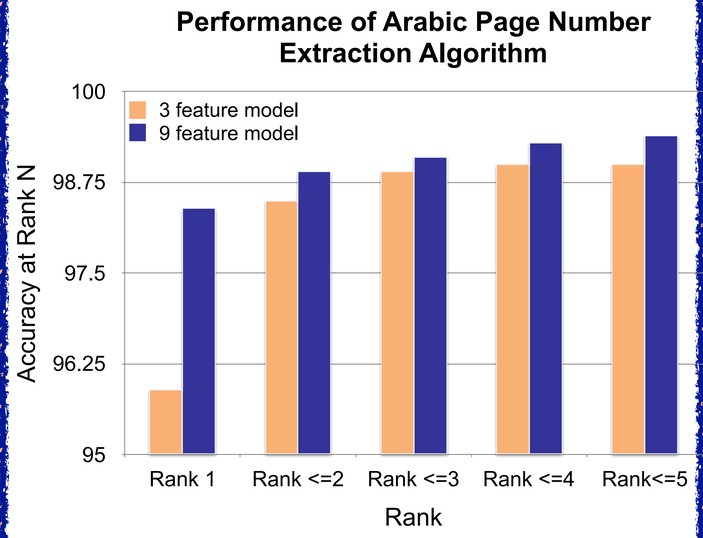
Initial model

- sequence
- % through the line
- % down through the page

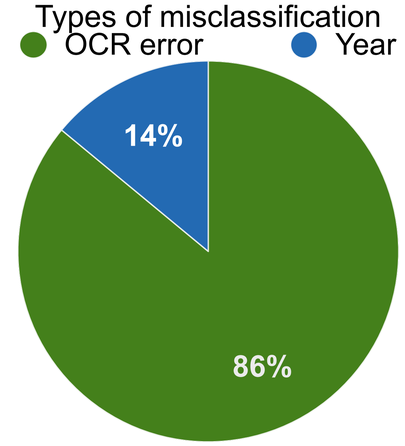
Final model

- is it a number?
- does it look like a number(i0, l, 1o, b,s)?
- % left through the line
- % up through the page
- x coordinate fraction
- y coordinate fraction

Results & Conclusion



- Overall page number extraction accuracy for 3 feature model is 95.9%.
- Overall page number extraction accuracy for 9 feature model is 98.4%.



- Improvement between models is around 3.5 %.
- 6 features added collectively to the model don't have as much as impact as 3 initial features.
- Didn't find a feature that greatly influences algorithm performance. Parser impacts greatly but its performance isn't quantified.