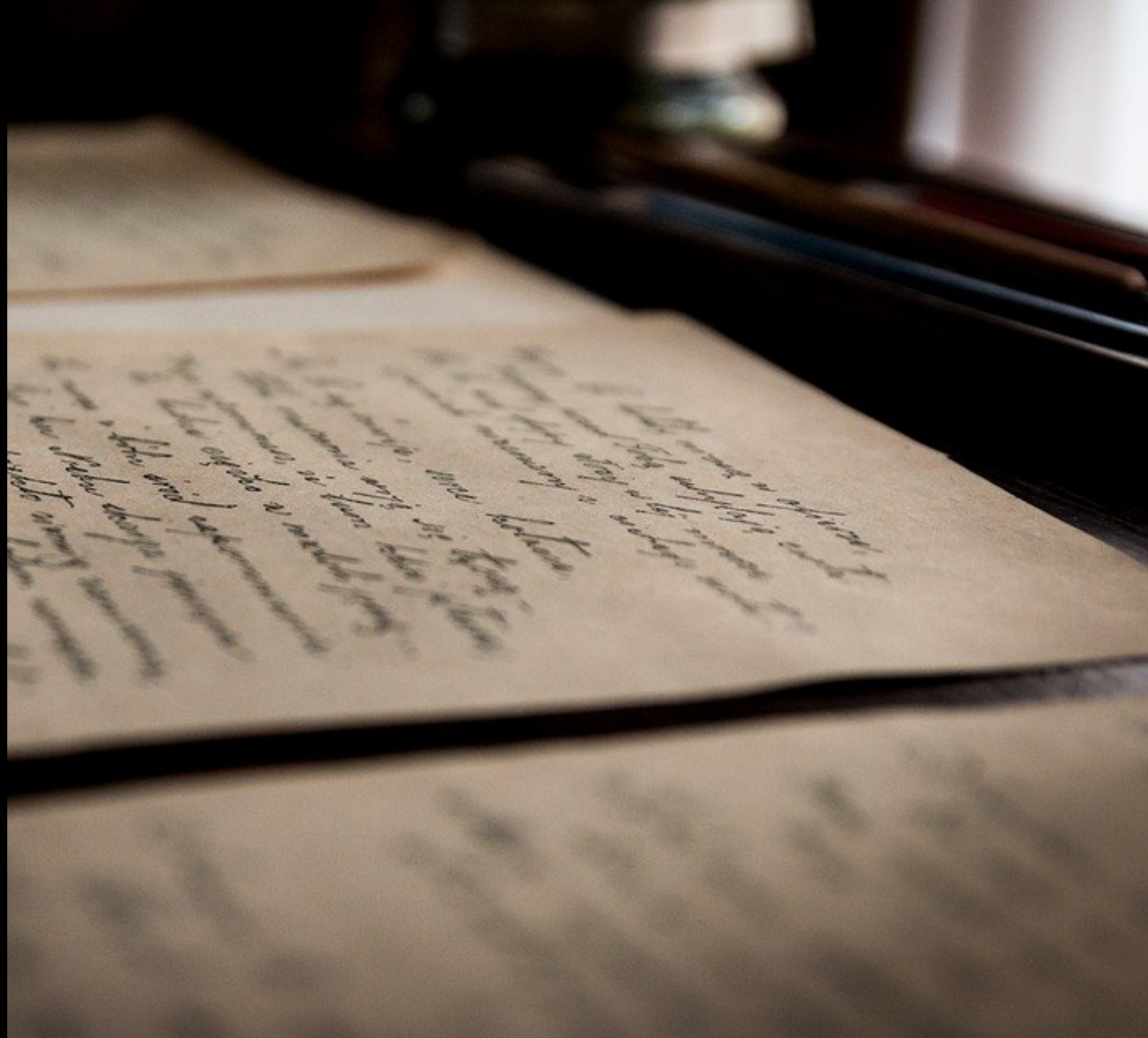
Cursive & Recursive:

Generating Transcriptions of Archival Documents Using Machine Learning



Week 4

Buchanan Fellowship, Spring 2020

Vanderbilt University Library

Tesseract Recap

Try out ABBYY FineReader

- Work through this tutorial:
  https://guides.nyu.edu/c.php?g=823477&p=5878688

- Windows partition on DHC Computer #1
  (There is a Mac version, but the Windows version is more developed)

- Check DHC hours on DHC homepage:
  https://www.vanderbilt.edu/digitalhumanities/

Dirty OCR

- Raw, uncorrected OCR text is dirty, and it can only become clean until it is corrected.

- HathiTrust OCR is dirty and uncorrected.

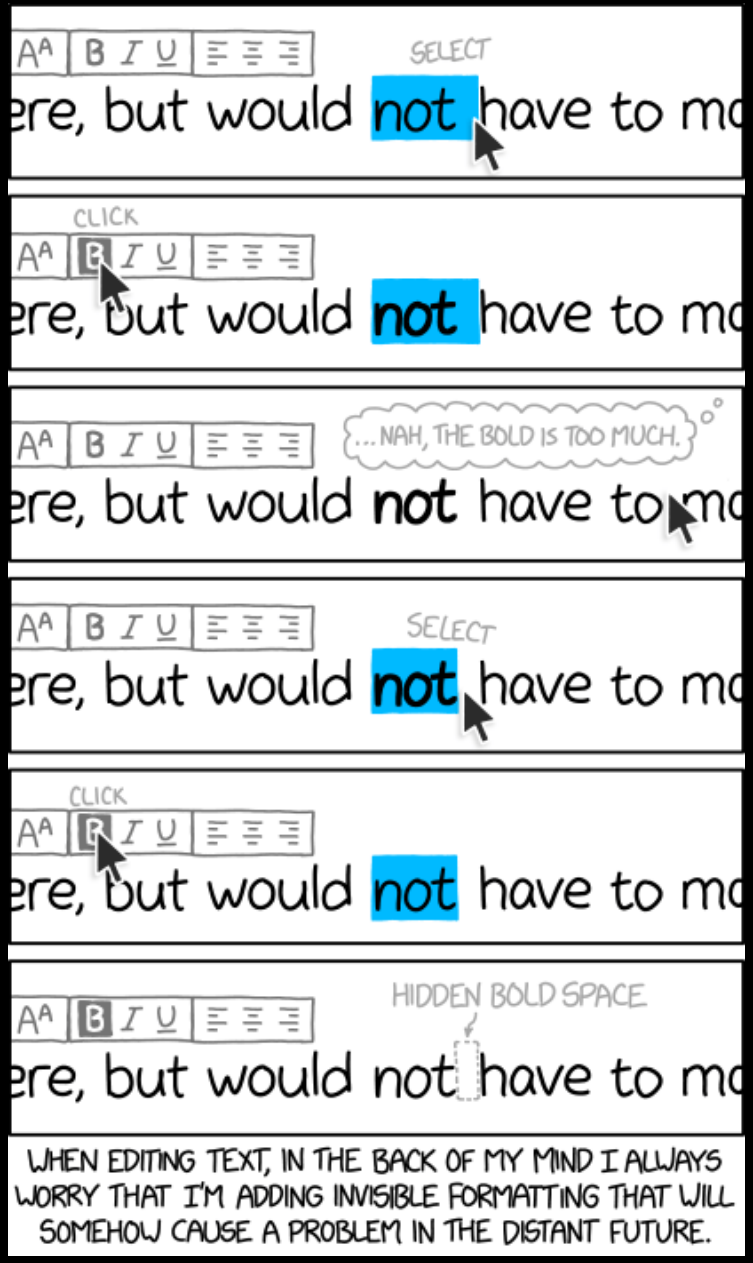
Example:
https://babel.hathitrust.org/cgi/pt?id=nyp.33433074972401&view=1up&seq=388

Cleaning up OCR

- Open a text editor
  - Atom (atom.io)
  - Sublime Text (www.sublimetext.com/)

- Learn regular expressions
  - https://regexr.com/
  - https://regex101.com/
  - See also “Cleaning OCR’d text with Regular Expressions” on Programming Historian website

Why use plaintext



https://xkcd.com/2109/

Grab a text – one way

- Digitized books on Internet Archive
- Search for 1854 edition of Thoreau's Walden from Boston Public Library
- Or go to https://tinyurl.com/WaldenOCR
- Go to Show All download options on right sidebar
- Select the.txt file (waldenorlifeinwo1854thor_djvu.txt) and save as walden.txt

Grab a text – another way

- Open Terminal
- Type:

```
Wget
https://archive.org/download/waldenorlifeinwo185
4thor/waldenorlifeinwo1854thor_djvu.txt
```

Getting started

- Open Walden.txt in your text editor and explore
- Delete front matter
- Turn on regular expression feature
- Enter regular expressions in the search box
  - Edit – Find (or Cmd – F)

Pattern Matching with Regular Expressions (RegEx)

- Useful skill for cleaning texts and spreadsheets
- More powerful than Find & Replace
- Can use in text editor, Open Refine, and most programming languages

Some Basics

Term	Meaning	Sample regex	Matches
+	one or more	he+y	hey, heeeeeey
?	optional	colou?r	color, colour
*	zero or more	toys*	toy, toys, toysss

Some Basics

regex	matches	doesn’t match
/the/	the, isothermally	The
/[Tt]he/	the, isothermally, The	
/\b[Tt]he\b/	the, The	—The

Some Basics

Symbol	Function
\b	Word boundary (zero width)
\d	Any decimal digit (equivalent to [0-9])
\D	Any non-digit character (equivalent to [^0-9])
\s	Any whitespace character (equivalent to [\t\n\r\f\v])
\S	Any non-whitespace character (equivalent to [^ \t\n\r\f\v])
\w	Any alphanumeric character (equivalent to [a-zA-Z0-9_])
\W	Any non-alphanumeric character (equivalent to [^a-zA-Z0-9_])
\t	The tab character
\n	The newline character

Try it out on Walden

- Go to Worksheet
- Remove page headers and page numbers
- Remove hyphenated words at line breaks

Things to think about

- Standardize spelling?
- Expand contractions?
- Remove numbers?
- Change all to lowercase?
- Punctuation
  - @Jane_Smith
  - #BlackLivesMatter

All depends on what you want to do with your OCR’d text

OCR Training Use Case

- Need to match handwriting to text as precisely as possible to train handwriting recognition algorithm

- Therefore:
  - Won’t correct or standardize spelling
  - Won’t expand contractions
  - Won’t change all to lowercase
  - Will be careful about punctuation

Another Use Case

Ted Underwood on Text Mining :

“The algorithms I use can handle a pretty high level of error as long as those errors are distributed in a more-or-less random way. If a word is mis-transcribed randomly in 200 different ways, each of those errors may be rare enough to drop out of the analysis. You don’t necessarily have to catch them all.”

https://tedunderwood.com/2013/12/10/a-half-decent-ocr-normalizer-for-english-texts-after-1700/

Text as Data

- Cleaning data often involves discarding data
- Prepared text may be illegible to the human reader
- Amount of text processing changes your results

Rockwell, G. (2003). *What is Text Analysis, Really?* Literary and Linguistic Computing, 18(2), 209–219. https://doi.org/10.1093/llc/18.2.209

Denny, M. J. and Spirling, A. (2017). *Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It.* https://ssrn.com/abstract=2849145