Turning PDFs into Research Data

Unit Three: Part One

Overview

Instructor: John ‘Jack’ Collins (2024)

Turning PDFs into Research Data

Unit Three: Part Two

Text Extraction

Instructor: John ‘Jack’ Collins (2024)

Is Your PDF an Image or a Postscript?

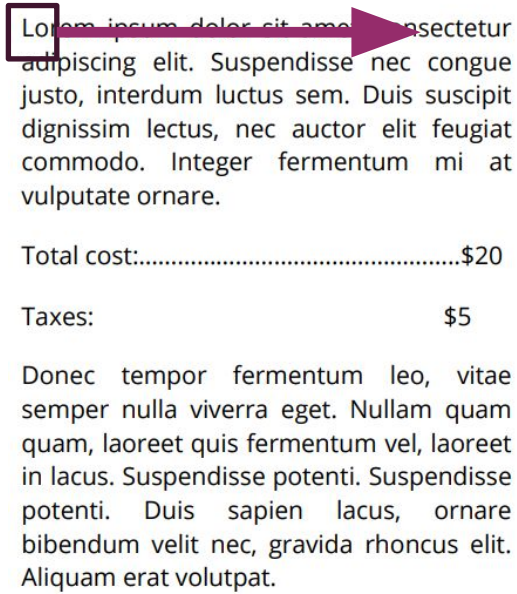
- If the creator of the PDF saved an MS Word document to a PDF, then likely the text can be extracted most easily by parsing the **Postscript** which is the markup language for PDFs.

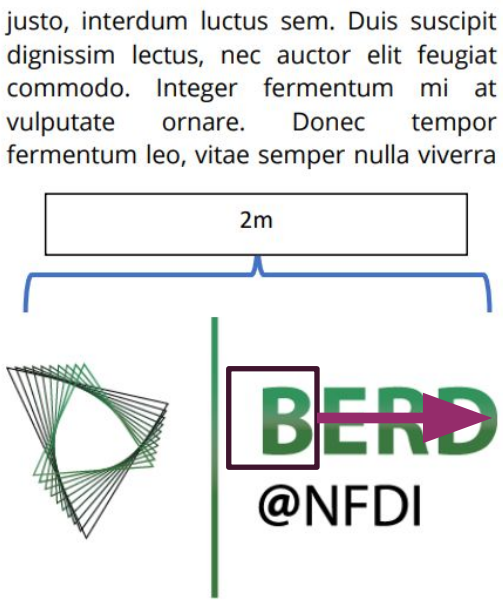
- If the creator of the PDF scanned documents, then the text ill be contained within a image embedded in the PDF and you will need to use **Optical Character Recognition (OCR)** software.

```
%!PS-Adobe-3.0
%%BoundingBox: 0 0 612 792
%%Pages: 1
%%EndComments

%%Page: 1 1
/Times-Roman findfont
24 scalefont
setfont
72 720 moveto
(This is a sample text in PostScript) show

newpath
144 576 moveto
144 648 lineto
288 648 lineto
288 576 lineto
closepath
stroke

showpage
```

Introduction to OCR

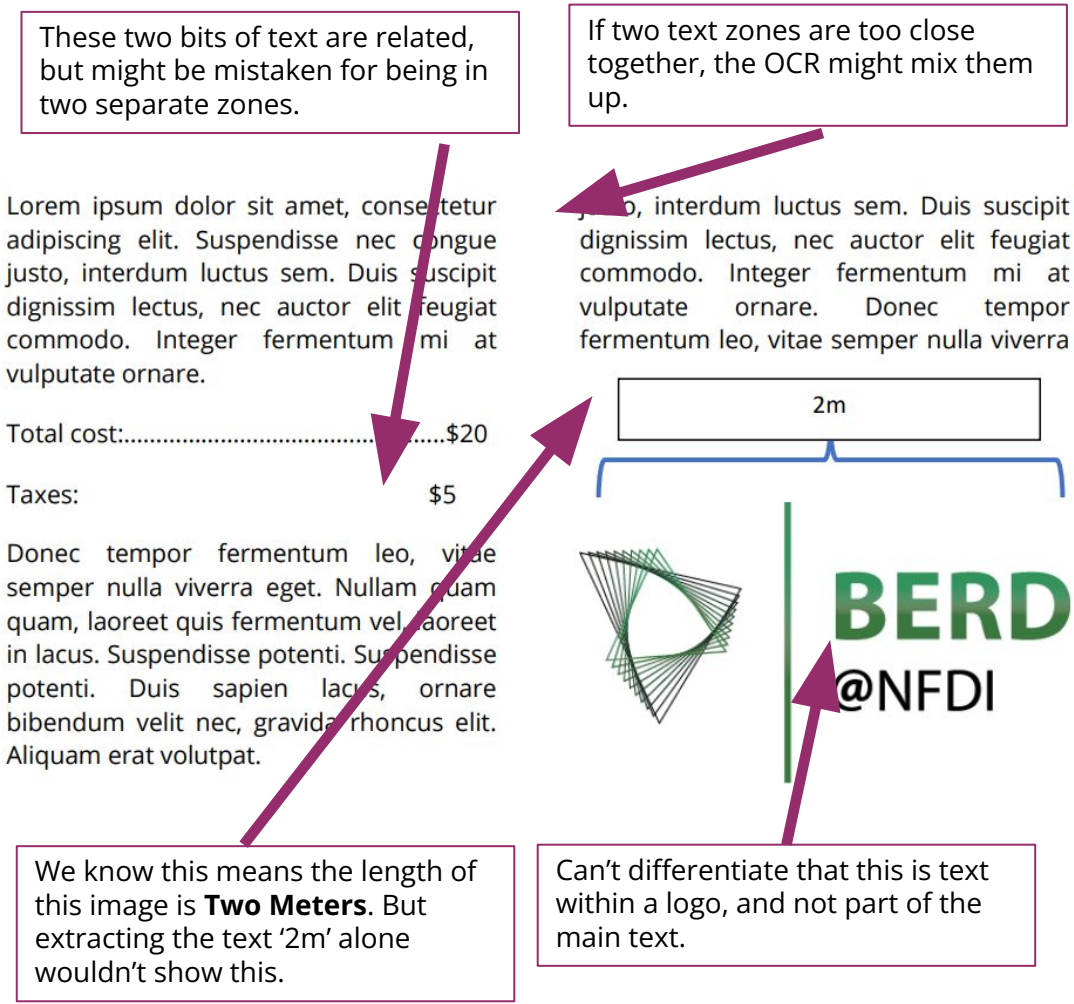
- **Optical Character Recognition (OCR)**
- Stage One: Break apart the document into text zones (otherwise text across two columns would be read as one line!)
- Step Two: In each zone, scan across with a box and guess the probability that the given set of pixels makes any possible letter.
- Step Three: Based on spelling, grammar, and what characters comes before and after a given letter, update those probabilities.
- Output: The most likely character sequence (including white spaces).





What can make OCR fail?

- Poor image quality, bad resolution makes it hard to tell similar shaped letters.
- Special characters like A vs Ä.
- Complex text zones, or large spaces between related texts.
- When text only makes sense in relation to an image (i.e.: annotations on a map)



How to Maximise OCR success

- If you are having problems with text extraction, the following pre-processing steps might help.
- Image de-noising and sharpening.
- convert colour image to black-white to make text easier to read.
- Adaptive thresholding is especially good for old/poor quality scans.

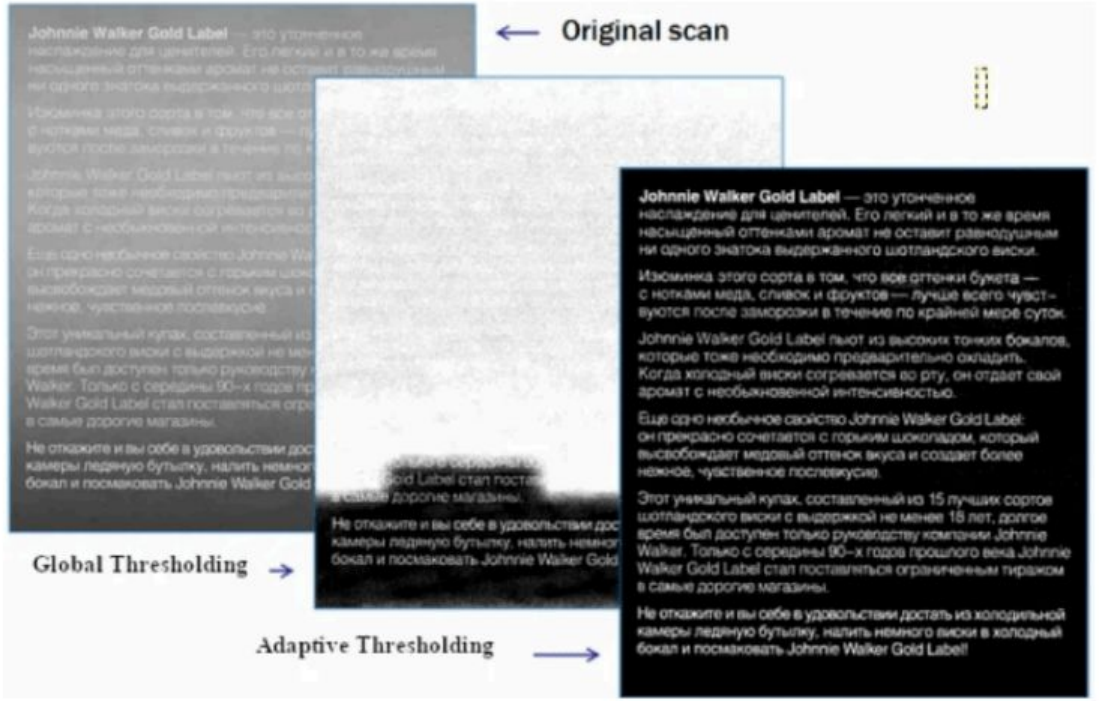


Image Credit: https://ubiai.tools/democratizing-text-extraction-exploring-the-best-open-source-ocr-solutions/

Selecting your OCR solution

Training an OCR model is very complex, so almost always you should go with a pre-trained solution. Fortunately, there are many options. Here's a short list of many common options, but there are many more. The Unit readings will feature an article comparing some of these options.

- Amazon Textract and Google Cloud Vision: Highly accurate, but not free. Amazon/Google offers some free credit to try before buying. Best for common commercial cases (invoices, contracts, etc).
- Tesseract: Open source and free, but known to have lower accuracy, unless the image of the text is very good and the use case is very simple (i.e.: none of the problems discussed above).
- EasyOCR: Similar to Tesseract, but for python only.
- Class exercises will demonstrate using the free options, but we can discuss how you might trial the paid options.

Turning PDFs into Research Data

Unit Three: Part Three

Keyword Extraction

Instructor: John ‘Jack’ Collins (2024)

Why Keyword Search?

- Once you have extracted the **‘text blob’** from your document, the data you are interested in might be in only a part of the document. By searching for keywords, you might be able to isolate the relevant part of the document (i.e.: a section heading, or the text before and after a keyword appears).

- Depending on your use case, keyword searching might be all you need. For example, **counting the number of times a keyword** is mentioned, **classifying topic of document** by keyword mention.

Some things to know when keyword searching

- Consider using **Stemming and/or Lemmatization** on your text blob. This will pre-process the text such that words like “thinking” become “think”. Always use a Stemmer or Lemmatizer which is specific to the language of the text.

- Consider searching for multiple versions of your keyword. I.e.: “Flood” and “Flooded.” Also consider with and without special characters in case of OCR mis-reads.

- In some cases, the document might have start/stop characters which help with extracting structured data. For example, tables may have column headers and table annotations.

Turning PDFs into Research Data

Unit Three: Part Four

AI for Data Extraction

Instructor: John ‘Jack’ Collins (2024)

Why AI?

- Several studies have been found certain AI algorithms can perform a range of data extraction tasks with text.
- For example, **BERT** can be used to categorize open text passages into topics based on training you provide.
- In this section, we are interested in extracting data that is in the text, but we want to convert it to a structured format. For example, developing a table of contractor names and contact details by reading scans of contracts.
- The name of this type of NLP task is **“Named Entity Extraction”**

Names Entity Extraction (NEE)

- Before LLMs, deep-learners were used for NEE and were typically tuned on specific tasks using lots of data. Therefore, you might consider looking for NEE dee learners and using training data to make your own NEE model if you cannot use LLMs (this approach not covered in this course).
- If you can use LLMs (**sensitive data might not be allowed to be uploaded**), they can be leveraged for NEE tasks.
- I recommend looking at purpose-built solutions from the industry (listed later), but in this course we will show how to use Chat-GPT and examine some of the issues you need to understand when doing this.

Input

"Barack Obama was born on August 4, 1961, in Honolulu, Hawaii. He served as the 44th president of the United States from 2009 to 2017. Obama graduated from Columbia University and Harvard Law School."

Output

- **Person:** Barack Obama
- **Date:** August 4, 1961
- **Location:** Honolulu, Hawaii
- **Title:** 44th president
- **Organization:** United States
- **Date Range:** 2009 to 2017
- **Institution:** Columbia University
- **Institution:** Harvard Law School

Text source: https://en.wikipedia.org/wiki/Barack_Obama

- **Favourite Food: ???**

Data-from-Text Options

- **Amazon Comprehend:** Purpose built to extract data you define from a large body of documents. Not free, but Amazon offers free credit to get started.
- **Google NLP API**: Google provides code examples to help you engineer your own data-from-text extraction pipeline. Not free, but Google can provide free credit to trial.
- **Extracta, Parsio, Parseur**: Industry products tuned for common use cases like contracts and invoices. Good for reading scans of tables into dataframes.
- **Chat-GPT, Llama, Google Gemini**: Studies show they perform worse than purpose-build NEE Deep-learners, but these LLMs still perform well, are more versatile, and require relatively less investment of training and development.

Prompting LLMs for NEE

My Prompt

- Few shot training is a process for providing examples for the LLM to emulate.
- Providing an Output template with symbols like **$$** and **@@** will help you write code that can turn this into a table later. Use symbols that will not naturally appear in the text like commas.
- You can also include in the output template a section for page references, or extractions of the original text which support the extracted data.
- In this case, I can search for the outputted values and check that they do appear in the original text to provide some confidence this was not a hallucination.
- Source of error: sometimes the answer to a question may be complicated, have multiple true answers, or be

Read this example for how to extract data from text:

Input:

"John 'Jack' Collins is a PhD Student in Sociology at the Graduate School of Economic and Social Sciences. He holds a Bachelor's of Sociology with Honours from the Australian National University. Jack has a Master's degree in Data Science from James Cook University. His Master's project was regarding predictive modelling for student attrition from sub-tertiary courses in Australia. During his Master's studies, he also assisted in research projects regarding social attitudes and voting behaviour in Australia. Before starting PhD, Jack was a Senior IT Consultant specialising in data engineering, analytics and software development. Jack is interested in applying Data Science and IT to sociological research, particularly with regard to machine learning, analytics, and web applications."

Output:

$$Institution$$:@@Graduate School of Economic and Social Sciences@@
$$Discipline$$:@@Sociology @@
$$Qualifications$$:@@ Bachelor's of Sociology with Honours from the Australian National University@@,@@Master's degree in Data Science from James Cook University@@.

Prompting LLMs for NEE

- Few shot training is a process for providing examples for the LLM to emulate.
- Providing an Output template with symbols like **$$** and **@@** will help you write code that can turn this into a table later. Use symbols that will not naturally appear in the text like commas.
- You can also include in the output template a section for page references, or extractions of the original text which support the extracted data.
- In this case, I can search for the outputted values and check that they do appear in the original text to provide some confidence this was not a hallucination.
- Source of error: sometimes the answer to a question may be complicated, have multiple true answers, or be

My Input

Input:

"Johannes Lattmann is a PhD candidate in Political Science at the CDSS. He completed his B.A in Social Sciences at the Humboldt University in Berlin and his M.Sc in Political Science & Political Economy at the London School of Economics in London. During his studies, he worked as a research assistant at Humboldt, LSE as well as at the Social Research Center in Berlin. Johannes´ research is focused on the causal analysis of political behaviour. This includes polarization dynamics, voting decisions as well as protests and the occurrence of political violence. Methodologically, he mainly has utilized quasi-experimental designs for his research. However, he is also interested in the application of automated text analysis for political science. Furthermore, in his research Johannes has examined the consequences of centralized funding allocations and their economic and political impact on local regions."

Chat GPT Output

Output:

##Institution##:@@CDSS@@
##Discipline##:@@Political Science@@
##Qualifications##:@@B.A in Social Sciences from the Humboldt University in Berlin@@,
@@M.Sc in Political Science & Political Economy from the London School of Economics in London@@.

Turning PDFs into Research Data

End of Unit Three

Instructor: John ‘Jack’ Collins (2024)