The Journal of Open Source Software

pdfsearch: Search Tools for PDF Files

Brandon LeBeau[1]

1 University of Iowa

DOI: 10.21105/joss.00668

Software

- Review
- Repository
- Archive

Submitted: 30 March 2018
Published: 27 June 2018

Summary

PDF files are common formats for reports, journal articles, briefs, and many other documents. PDFs are lightweight, portable, and easily viewed across operating systems. Even though PDF files are ubiquitous, extracting and finding text within a PDF can be time consuming and not easily reproducible. The pdftools R package (Ooms 2017), which uses the poppler C++ library to extract text from PDF documents, aids in the ability to import text data from PDF files to manipulate in R. The pdfsearch package (LeBeau 2018) is an R package (R Core Team 2016) that extends the text extraction of pdftools to allow for keyword searching within a single PDF or a directory of PDF files.

The pdfsearch package can aid users in manipulation of text data from PDF files in R and may also improve the reproducibility of the extraction and manipulation tasks. Users can search for keywords within PDF files where the location of the match and the raw text from the match are returned. This aspect of searching for keywords may be most useful for those conducting research syntheses or meta-analyses (Cooper 2017) that are more reproducible and less time consuming than current practice. Current research synthesis or meta-analysis practice involves the reading of each document to search for the presence of certain terms, phrases, or statistical effect size information to answer specific research questions. The improved workflow with the pdfsearch package would allow those conducting research syntheses, the ability to narrow down relevant portions of text based on the keyword matches returned by the package instead of looking at the entire text of the document. In addition, regular expressions could be written to search and extract statistical information needed to compute effect sizes automatically.

As an example, the package is currently being used to explore the evolution of statistical software and quantitative methods used in published social science research (LeBeau and Aloe 2018). This process involves getting PDF files from published research articles and using pdfsearch to search for specific software and quantitative methods keywords within the research articles. The results of the keyword matches will be explored using research synthesis methods (Cooper 2017).

The package vignette includes more information on this package. Included in the vignette are keyword searches within PDF documents and an exploration of the output from the package. The vignette also discusses limitations of the package. Below is example output of the package searching for the phrase “repeated measures” from Guo and Deng (2015).

References

Cooper, Harris. 2017. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*. Vol. 2. Sage publications.

Guo, Y., and A. Deng. 2015. “Flexible Online Repeated Measures Experiment.” *ArXiv E-Prints*, January.

```
# A tibble: 6 x 5
  keyword           page_num line_num line_text token_text
  <chr>                <int>    <int> <list>    <list>
1 repeated measures        1       24 <chr [3]> <list [3]>
2 repeated measures        2       57 <chr [3]> <list [3]>
3 repeated measures        2      108 <chr [3]> <list [3]>
4 repeated measures        2      110 <chr [3]> <list [3]>
5 repeated measures        2      125 <chr [3]> <list [3]>
6 repeated measures        6      444 <chr [3]> <list [3]>
```

Figure 1: Example output searching for "repeated measures" phrase in a single PDF document.

LeBeau, Brandon. 2018. *pdfsearch: Search Tools for PDF Files.* https://github.com/lebebr01/pdfsearch.

LeBeau, Brandon, and Ariel M Aloe. 2018. "Evolution of Statistical Software and Quantitative Methods."

Ooms, Jeroen. 2017. *pdftools: Text Extraction, Rendering and Converting of PDF Documents.* https://CRAN.R-project.org/package=pdftools.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.