Automatic citation extraction from URLs

phiresky

2019-08-21

Introduction

pandoc-url2cite allows you to instantly and transparently cite most papers directly given only a single URL.

You simply add a URL of a publication, and it will replace that with a real citation in whatever CSL [1] style you want. This means you can avoid dealing with Mendeley [2] or Zotero [3] and keeping your Reference Manager database and bibtex file in sync, especially when collaborating with others.

Minimal Example

Here is a minimal example:

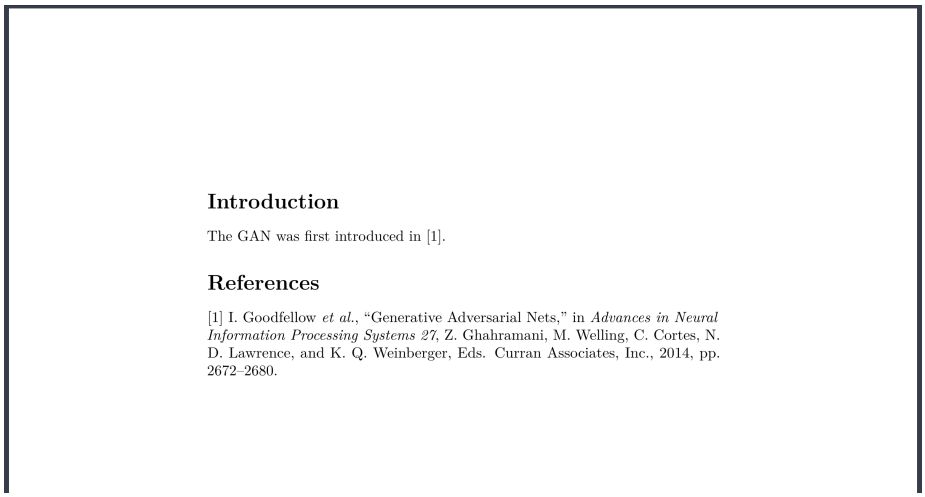
minimal.md

```
# Introduction

The GAN was first introduced in [@gan].

# References

[@gan]: https://papers.nips.cc/paper/5423-generative-adversarial-nets
```

Compiling this file with this command

```
pandoc --filter=pandoc-url2cite \
    --filter=pandoc-citeproc \
    minimal.md \
    --csl ieee-with-url.csl \
    -o minimal.pdf
```

This results in the following output:
minimal.pdf



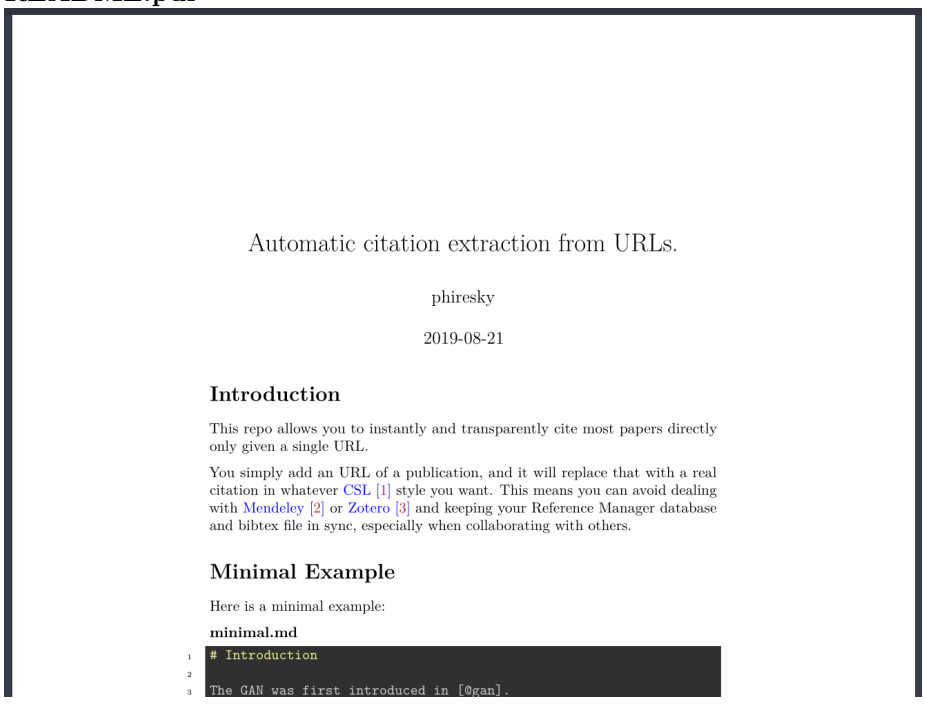
Introduction

The GAN was first introduced in [1].

References

[1] I. Goodfellow *et al.*, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

For a slightly longer example, you can look at this readme itself:
README.pdf



Automatic citation extraction from URLs.

phiresky

2019-08-21

Introduction

This repo allows you to instantly and transparently cite most papers directly only given a single URL.

You simply add an URL of a publication, and it will replace that with a real citation in whatever CSL [1] style you want. This means you can avoid dealing with Mendeley [2] or Zotero [3] and keeping your Reference Manager database and bibtex file in sync, especially when collaborating with others.

Minimal Example

Here is a minimal example:

minimal.md

```
# Introduction

The GAN was first introduced in [@gan].
```

Source README.md - Result README.pdf

How to Use

Install this package globally using `npm install -g pandoc-url2cite`.

Then, add `--filter=pandoc-url2cite` to your pandoc command (before pandoc-citeproc, see the minimal example above).

Alternatively, clone this repo [4] somewhere, then install the dependencies using `npm ci install`.

If you're not familiar with writing papers in pandoc, you can refer to e.g. this article [5]. It's pretty flexible, you can use templates from whatever conference you want, and you can still use inline latex code if you need it (and you are ok with not being able to convert your document to nice HTML or EPUB anymore).

Citation Syntax

url2cite allows multiple ways to cite:

1. (PREFERRED) Use the pandoc citation syntax for citations:

   ```
   The authors of [@alexnet] first introduced CNNs to the
   ImageNet challenge.
   ```

   More information about referencing specific pages etc. is in the pandoc manual [6].

   Then add the URLs with the usual "link reference" syntax to the bottom of your document in its own paragraph:

   ```
   [@alexnet]: https://...
   ```

2. Convert all links to citations

   Add `url2cite: all-links` to your yaml front matter [7]. This will cause all links in the document to be converted to references.

   You can still blacklist some links by adding `no-url2cite` to either the CSS class of the link (pandoc-only):

   ```
   [foo](http://example.com){.no-url2cite}
   ```

   or to the link title:

   ```
   [foo](http://example.com "no-url2cite").
   ```

How it Works

The main idea is that usually every piece of research you might want to cite is fully identifiable by an URL - no need to manually enter metadata like author, release date, journal, etc. Citation managers like Zotero already use this and enable you to automatically fetch metadata from a website. But then you still have a citation database somewhere that you may or may not be able to synchronize with different computers, but probably won't be able to add to the version control of your paper. There's hacks such as better-bibtex [8] to automatically generate and update diffable bibtex files – But that means you

now have two sources of truth, and since the export is one-way this leads to multiple contributors overriding each other's changes. pandoc-url2cite goes a step further: URLs are directly used as the cite keys, and the "bibliography file" is just an auto-generated intermediary artifact of those URLs.

pandoc-url2cite is based on the work of the Zotero [3] developers. Zotero has a set of "Translators" [9] that are able to extract citation info from a number of specific and general web pages. These translators are written in Javascript and run within the context of the given web site. They are made to be used from the Zotero Connector browser extension, but thankfully there is a standalone Translation Server [10] as well. To avoid the effort required to automatically start and manage this server locally, pandoc-url2cite instead uses a publicly accessible instance of this server provided by Wikipedia with a public REST API [11].

All citation data is cached (permanently) as bibtex as well as CSL to `citation-cache.json`. This is both to improve performance and to make sure references stay the same forever after the initial fetch, as well as to avoid problems if the API might be down in the future. This also means that errors in the citation data can be fixed manually, although if you find you need to do a lot of manual tweaking you might again be better off with Zotero.

Limitations

1. Currently, extracting the metadata from direct URLs of full text PDFs does not work, so you will need to use the URL of an overview / abstract page etc. I'm not sure why, since this does work in Zotero. More info might be here [12].

2. Currently, this filter only works if you use pandoc-citeproc, because the citations are written directly into the document metadata instead of into a bibtex file. If you want to use natbib or biblatex for citations, this filter currently won't work. Using citeproc has the disadvantage that it is somewhat less configurable than the "real" LaTeX citation text generators and the CSL language has some limitations. For example, the bibtex "alpha" [13] style sometimes used in Germany can't be described in CSL.

   To make it work with biblatex, this script would need to write out a *.bib file somewhere temporarily and reference that in the latex code.

3. Some websites just have wrong meta information. For example, citationstyles.org has set "Your Name" as the website author in their Open Graph [14] metadata.

4. Using URLs directly as citekeys (e.g. `[@https://google.com]` does not work because of pandoc parsing, see this issue [15]. But it does work for DOIs: `As shown in [@doi:10.1037/a0028240]...`!

Related Work (Longer Example)

AlexNet [16] first introduced CNNs to the ImageNet challenge. [17]–[19] further improved on the results.

References

[1] Y. Name, *Citation Style Language*. [Online]. Available: https://citationstyles.org/

[2] *Mendeley - Reference Management Software & Researcher Network*. [Online]. Available: https://www.mendeley.com/?interaction_required=true

[3] *Zotero – Your personal research assistant*. [Online]. Available: https://www.zotero.org/

[4] phiresky, *Effortlessly and transparently add correctly styled citations to your markdown paper given only a URL: phiresky/pandoc-url2cite*. 2019 [Online]. Available: https://github.com/phiresky/pandoc-url2cite

[5] 1. O. 2. K. Fern and e.-R. F. 1. 3. comments, *How to use Pandoc to produce a research paper*. [Online]. Available: https://opensource.com/article/18/9/pandoc-research-paper

[6] *Pandoc - Pandoc User's Guide*. [Online]. Available: https://pandoc.org/MANUAL.html#citations

[7] *Pandoc - Pandoc User's Guide*. [Online]. Available: https://pandoc.org/MANUAL.html#extension-yaml_metadata_block

[8] E. Heyns, *Make Zotero effective for us LaTeX holdouts. Contribute to retorquere/zotero-better-bibtex development by creating an account on GitHub*. 2019 [Online]. Available: https://github.com/retorquere/zotero-better-bibtex

[9] *dev:translators [Zotero Documentation]*. [Online]. Available: https://www.zotero.org/support/dev/translators

[10] *A Node.js-based server to run Zotero translators. Contribute to zotero/translation-server development by creating an account on GitHub*. zotero, 2019 [Online]. Available: https://github.com/zotero/translation-server

[11] *Citoid/API - MediaWiki*. [Online]. Available: https://www.mediawiki.org/wiki/Citoid/API

[12] *Try translating PDF URLs based on URL · Issue #70 · zotero/translation-server*. [Online]. Available: https://github.com/zotero/translation-server/issues/70

[13] *Bibtex bibliography styles*. [Online]. Available: https://www.overleaf.com/learn/latex/Bibtex_bibliography_styles

[14] *Open Graph protocol*. [Online]. Available: http://ogp.me/

[15] *url as citekey/referencekey · Issue #308 · jgm/pandoc-citeproc*. [Online]. Available: https://github.com/jgm/pandoc-citeproc/issues/308

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017 [Online]. Available: http://doi.acm.org/10.1145/3065386

[17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Sep. 2014 [Online]. Available: http://arxiv.org/abs/1409.1556

[18] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9 [Online]. Available: https://ieeexplore.ieee.org/document/7298594

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778 [Online]. Available: https://ieeexplore.ieee.org/document/7780459