Automatic citation extraction from URLs.

phiresky

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 style you want. This means you can avoid dealing with Mendeley or Zotero and keeping your Reference Manager database and bibtex file in sync, especially when collaborating with others.

Minimal Example

Here is a minimal example:

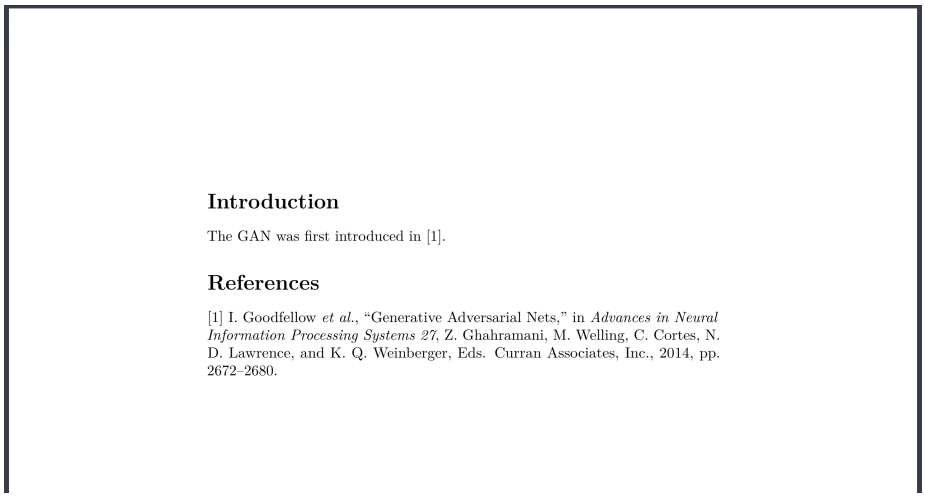
minimal.md

```
---
citekeys:
    gan: https://papers.nips.cc/paper/5423-generative-adversarial-nets
---

# Introduction

The GAN was first introduced in [@gan].

# References
```

Compiling this file with this command

```
pandoc --filter=pandoc-url2cite/index.js \
    --filter=pandoc-citeproc \
    minimal.md \
    --csl ieee.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: Source README.md - Result README.pdf

How to Use

Clone this repo somewhere, then install the dependencies using `npm ci install`. Then, add `--filter=pandoc-url2cite/index.js` to your pandoc command (before pandoc-citeproc).

If you're not familiar with writing papers in pandoc, you can refer to e.g. this article. 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).

How it Works

url2cite is based on the work of the Zotero authors. Zotero has a set of "Translators" 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 as well. To avoid the effort required to automatically start and manage this server locally, url2cite instead uses a publicly accessible instance of this server provided by Wikipedia with a public REST API.

To improve performance and to avoid problems if the API might be down in the future, all citation data is cached (permanently) as bibtex as well as CSL to `citation-cache.json`.

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.
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" sometimes used in Germany can't be described in CSL.

Related Work (Longer Example)

AlexNet [1] first introduced CNNs to the ImageNet challenge. [2]–[4] further improved on the results.

In [5] there is some other interesting stuff.

References

[1] 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.

[2] A. Zisserman and K. Simonyan, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014.

[3] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[4] 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.

[5] 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.