README

How it works

Scrapers use patterns in data to collect and map website data to a different, more usable format for further processing. In the case of childcarefinder.gov.au this pattern is in the way the website serves its search results.

Search results are served per postcode/suburb via the url.

For example, the results for Surry Hills, 2010 in NSW are surfaced via the url: https://www.childcarefinder.gov.au/search/nsw/2010/surry+hills - note the location of the state, postcode, and suburb in the url - this is all repeated for all postcodes in the government database.

This allows the script to iterate via a full list of postcode and suburb names. This is readily available data.

The process is then to read in this large dataset of postcodes =>

1. For each postcode, scrape the results, paginated, associated with the suburb

    - Each centre in the search list is then added to a further deeper scrape (we get the centres url)
    - These results are queued, and processed at most five at a time to not overwhelm the website with requests

2. From the data in 1. each centre is scraped in a further iteration, here we get fee, vacancy, and last updated data on the centre specific page.

This process more or less continues for all centres. This will output a data shape along the lines of:

```
{
  "title": "Nakara School Council After School Hours Care",
  "id": "3575683220",
  "link": "https://www.childcarefinder.gov.au/service/nt/0810/nakara/nakara+school+c
  "state": "NT",
  "suburb": "Nakara",
  "postcode": "0810",
  "type": "Outside School Hours Care",
  "vacancy": true,
  "email": "nakara.ashc@ntschools.net",
  "phone": "0889279823",
  "address": "Goodman Street, NAKARA NT 0810",
  "fees": null
}
```

Getting started for development

Ensure `node` and a version of `chromium` is installed on your system. The `node` binary is available here https://nodejs.org/en/download/

Once node is successfully installed the project dependencies also need to be installed.

```
# In a terminal window
# to install dependencies
$ npm i
```

Assuming no errors, quick start running via:

```
$ npm start
```

This will begin the scraper running against all Australian postcodes. This will take some time to complete - expect upwards of an hour to complete the full list of childcare centres.

Making changes

To validate changes you'll need to:

- update the source
- rebuild the project; run `npm run build`
- and then run the script; run `npm start`

If you don't follow these steps you may not see changes reflected

Navigating the source

All source is written in typescript and is required to be built by `tsc` to be run on node (or run `npm start`).

Logic is split out into three core tasks:

1. Parsing the postcodes of Australia (from `src/data/postcodes_newlines`)
2. Scraping the childcare finder website for childcares for each postcode (see `src/run-postcode.ts`)
3. Scraping each individual postcode for fee / contact information (see `src/run-centre.ts`)

Each of these files has annotations to make the control flow easier to follow.