Not Google

Kyle Russell
13831056

Contents

1.0 Crawling

This section will give you everything you need to know about crawling and indexing in NotGoogle. Details on how workers and URL's can be added, keywords, crawler settings, starting/stopping the crawler and creating/opening indexes will be explained.

1.1 Interface

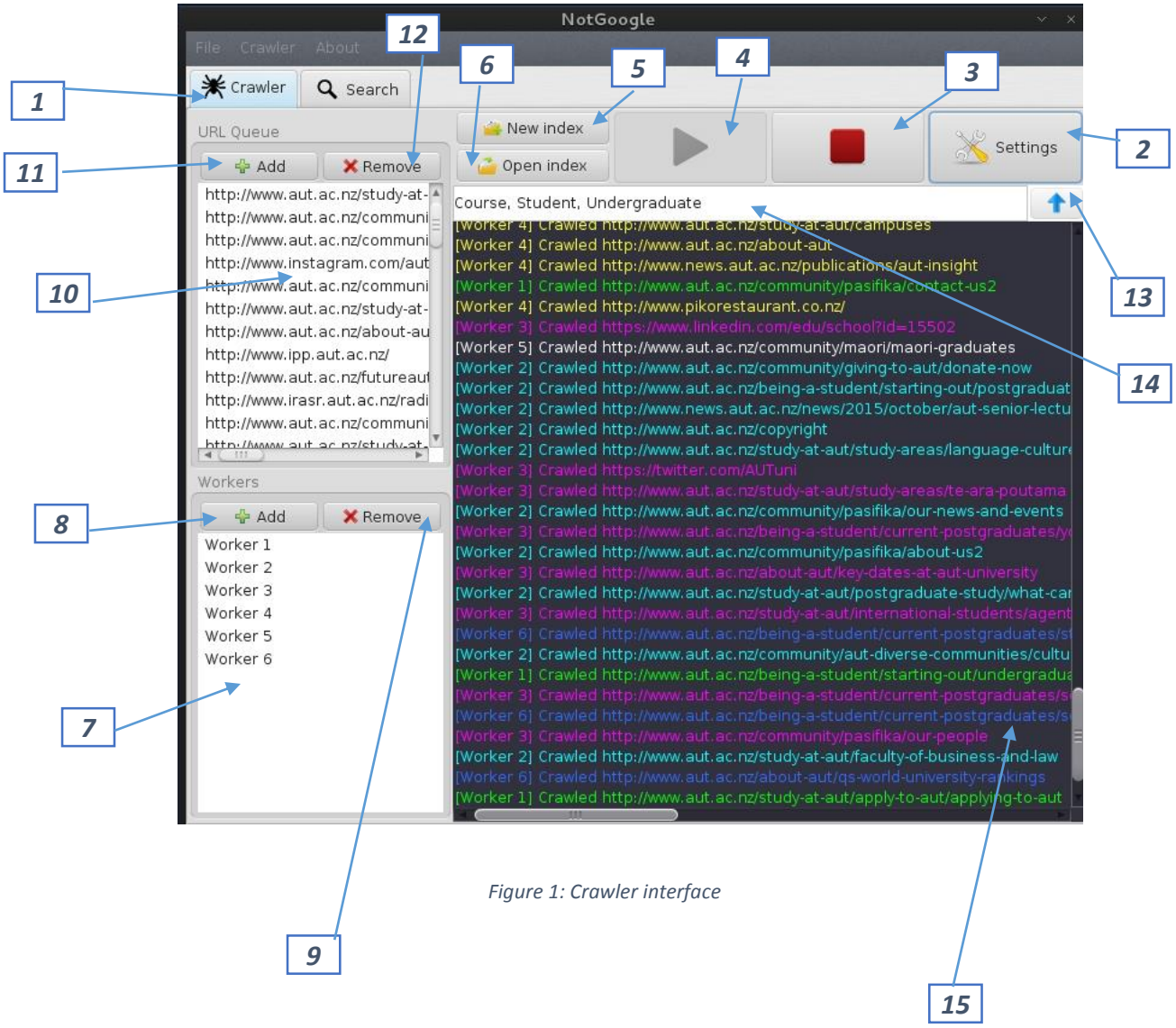

Figure 1: Crawler interface

1) View tab menu, includes crawler view (current) and search view (second)
2) Settings button, configure the crawlers settings (see section 1.5)
3) Stop button, stops the crawler when crawling
4) Start button, starts the crawler
5) New index button, shows dialog to create a new index for the crawler
6) Open index button, shows dialog to open an existing index for the crawler
7) Worker list, shows the names of the workers being used by the crawler
8) Add worker button, shows dialog to create a new worker (section 1.2)
9) Remove worker button, shows dialog to remove a worker
10) URL list, queue of URL's that the crawler has yet to visit
11) Add URL button, adds a URL to the queue
12) Remove URL button, remove a URL from the queue
13) Update keywords button, set crawler keywords to those in keyword bar (14)
14) Crawler keywords bar, the keywords being used by the crawler
15) Worker output, shows what URL's are being crawled by each worker

1.2 Workers

The spider in NotGoogle uses workers called 'spider workers' to crawl the web concurrently. Each worker is distinguished by a name where the names of created workers are displayed in the crawler list.

Figure 2: add worker dialog

To **create a new worker** click the 'Add' button in the worker pane fig. 1 (8). You will be prompted with the dialog in figure 2. In the text field provided you can specify the workers name, it must be unique (not listed in the worker list) and you can also choose a colour for the workers output. To choose a colour for the worker click the colour button and choose your colour. Click OK when you are done and your worker will be added to the worker list.

To **remove a worker** click the 'Remove' button in the worker pane fig. 1 (9). If you have already selected a worker from the worker list then this worker will be deleted automatically for you. Otherwise you will be prompted with the remove worker dialog as shown in figure 3. The worker name of the worker you enter in the text field provided in figure 3 must be an existing worker.

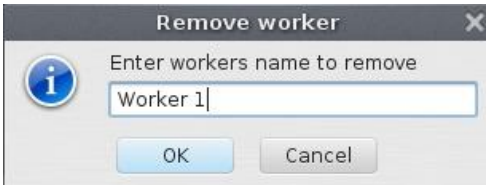

Figure 3: remove worker dialog

1.3 URL's

The spider's list of unvisited URL's is seen in the 'URL Queue' pane list fig. 1 (10) and the crawler requires some seed URL's before starting.

Figure 4: add URL dialog

To add a URL we can click the 'Add' button in the URL Queue pane fig. 1 (11), and you will be prompted with the 'Add URL' dialog shown in figure 4. You will be asked to enter a valid URL as demonstrated in figure 4. If you enter an invalid URL you will be notified. When you are done click OK and your URL will be added to the queue and will be displayed in the URL list.

Figure 5: remove URL dialog

To remove a URL click the 'Remove' button in the URL Queue pane fig. 1 (12). If you have a URL already selected then the URL will be removed automatically from the queue, otherwise you will be prompted with the 'Remove URL' dialog shown in figure 5. The name you enter must be an existing URL in the queue otherwise you will receive an error. Once you're done click the OK button to remove the URL you have entered.

1.4 Keywords

The crawler uses keywords specified in the keyword bar fig. 1 (14), to ensure the pages it parses are relevant to the terms you have entered. Later once you have crawled, you can use these keywords to search for. Keywords are delimited by a comma and optionally a space. The keyword update button fig. 1 (13) is used to update the crawler with keywords, if you are actively crawling and want to change the crawlers keywords then you can click the update keyword button and the crawler's keywords will be updated to those in the keyword bar and you must have at least one keyword specified.

1.5 Settings

The crawlers configurations can be changed by clicking the 'Settings' button fig. 1 (2). You can also open this dialog by selecting it from the Crawler menu:
Crawler->Preferences. You can change any of the fields in the configurations window and once done, you can click OK and the changes will be saved. The following include descriptions on each of the fields:

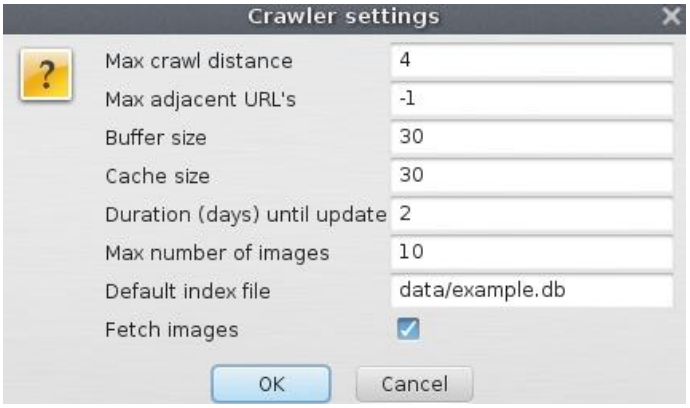

Figure 6: Crawler configurations window

Max crawl distance is the maximum distance that the crawlers will go, to ensure we don't index too much this value should be low.

Max adjacent URL's is the maximum number of links that can be added to the URL queue by a node that has been crawled. To add all valid links found, use -1.

Buffer size is the size of the crawler's buffer. Newly found pages are added to the buffer to be indexed, these pages are indexed in batches to ensure we aren't accessing the database constantly. This value should be chosen sensibly to preserve a small memory footprint and make as infrequent DB access as possible.

Cache size is the size of the crawler's cache. Frequently accessed pages are added to the crawler's cache to avoid unnecessary index lookups. A large cache size will result in less frequent index lookups but a larger memory footprint.

Duration (days) until update is the number of days before old pages should be re-visited.

Max number of images is the max number of images a crawler can fetch and index. This value is ignored if **Fetch images** is unselected.

Default index file is the path of the file for the database/index.

Fetch images should be selected if you want to fetch and index the images of pages.

1.6 Indexes

The crawler uses a database as an index to store page data for searching and later lookups. By default you are provided with an index that has some sample data and can be found in data\example.db. Additionally you can create or open your own NotGoogle indexes.

Figure 7: Create index dialog

To create an index click the 'New index' button fig.1 (5). You can create an index from the menu: File->New Index. You will be prompted with a dialog shown in figure 7 to create an index. Here you will be required to enter the name of the index where the file will be placed inside the 'data' directory. You may also select whether or not to make this newly created index the default index. Doing so will update the crawler's configuration and from now on this index will be used. This can be changed anytime in the settings window (section 1.5). Once you're done click OK and your index will be created. You will be operating on this index during this session and can be changed by opening an index.

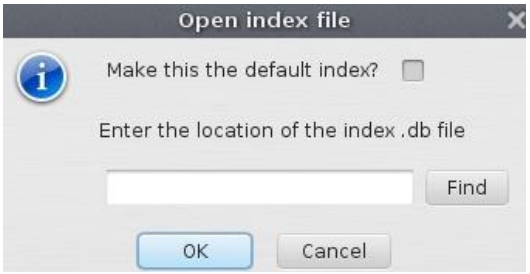

Figure 8: Open index dialog

To open an index click the 'Open index' button fig. 1 (6) where you will be prompted with the dialog in figure 8. You may also open an index from the menu: File->Open Index. You will be required to enter the (absolute or relative) path of the index file. If you are unsure of this path and want to locate the file then you may use the 'Find' button where you will be given a file choosing dialog where you can easily locate the file. The index file entered must be a NotGoogle index .db file. Similarly if you want to make this index your default you may select "Make this the default index?" and your crawler configurations will be updated. When you finished click the OK button and you will now be using this index file for this session.

1.7 Start/Stop crawler

Now that we have covered how to manage workers, URL's, settings, keywords we can now explain how to control the crawler. **Before starting the crawler** you must ensure that you have at least one seed URL, one worker and one keyword to search for. Additionally you must be using an index, by default you are provided with one but you may also open or create one.

To start the crawler click the start arrow button fig. 1 (4). Your crawler will now begin crawling and worker output will be displayed in the output pane fig. 1 (15). Additionally you can start the crawler from the crawler menu: Crawler->Start.

To stop the crawler click the stop square button fig.1 (3) or select it from the crawler menu: Crawler->Stop. The crawler must have been started otherwise these function will be unavailable. You may resume crawling by simply clicking the start button again.

2.0 Searching

By now you have been able to manage and run the crawler and should have a decent index built. The index that has been built from crawling will now be used in the search engine. This section will explain how you can perform searches with the NotGoogle search engine.

2.1 Interface

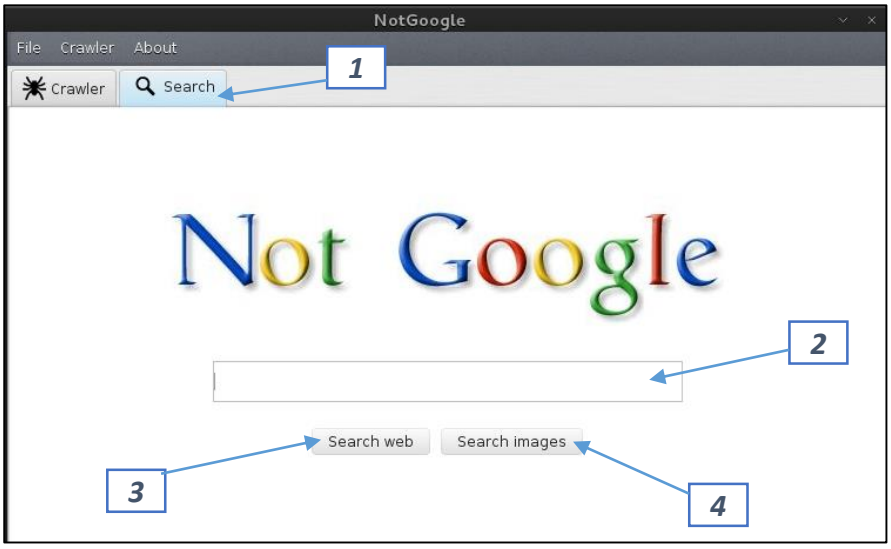

Figure 9: Search interface

1) The Search tab of the view tab pane. To view the search view you must be on the search tab.
2) Search bar. Enter your keyword to search in this field.
3) Search web button, searches for web results and takes you to the web results view
4) Image search butt, searches for image results and takes you to the image results view

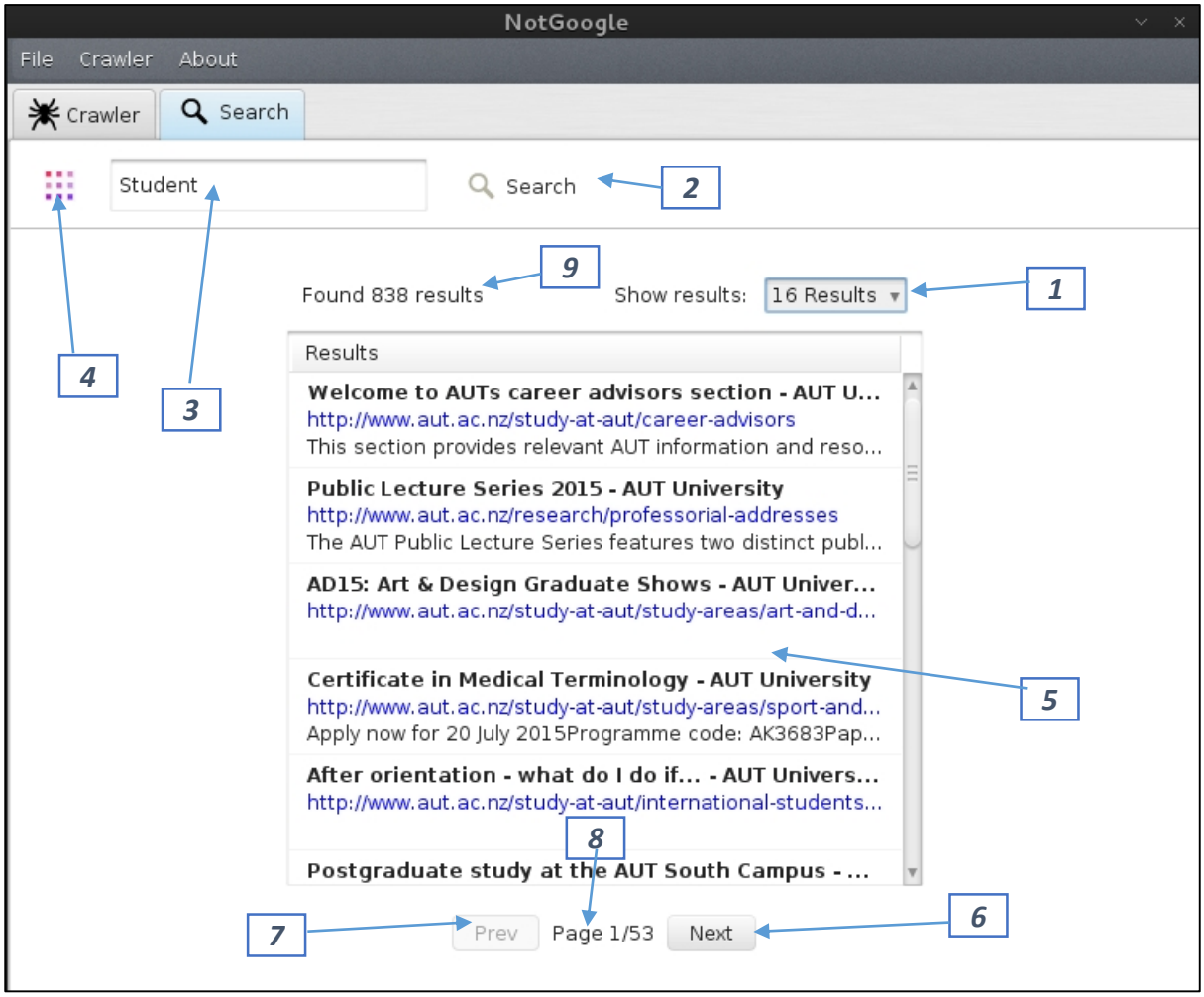

Figure 10: Web search view interface

1) Filters number of results per page
2) Allows searching from the results view
3) Search bar, can be changed and then searched from
4) Search view button, takes you back to the search view
5) Web results list
6) Next page
7) Previous page
8) Page info: Current page/Max page
9) Number of results found from the search

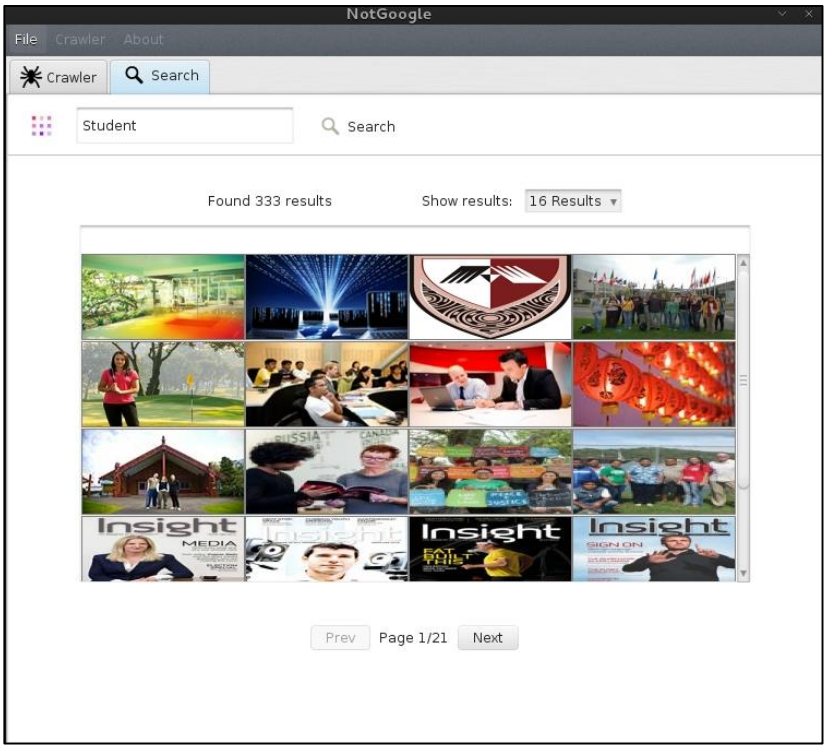

Figure 11: Image search results in results view

2.2 Performing a search

To perform a search in NotGoogle you must be in the search view, to do so click the search tab in the view tab pane fig. 9 (1). You will be displayed with the search view on figure 9. In this view you can enter search term in the search bar fig. 9 (2) and then choose whether you want to search for web results or image results. The keyword you enter must already be indexed otherwise you will see no results from the search.

2.3 Performing a web search

Once you have entered your search term click the 'Search web' button fig. 9 (3). If your search returns some results you will be redirected to the web search results view as shown in figure 10. The results of your search can be seen in the list of fig.10 (5). The list is ordered by PageRank and each entry contains the pages title, URL and description. You can click an entry to expand on the details of a result item where you will be given the dialog shown in figure 12.

Figure 12: Web result item dialog

This web result item dialog gives all the information about the selected page including its URL, parent URL (page that found this page), title, description, the date it was last updated and it's PageRank. Additionally you are provided with the buttons 'Open in browser' and 'Copy to clipboard'. The first button will attempt to open this page in your default browser. This functionality is only supported in Windows and you may need to enable extra privileges to launch the page in your browser. The second button will copy the pages URL to your clipboard.

2.4 Performing a image search

You can also search for images in NotGoogle. To perform an image search you will need to use the 'Search images' button fig. 9 (4) instead of the 'Search web' button fig. 9 (3). You will then be redirected to the image results view shown in figure 11 with your resulting images displayed in the table. These images are ranked in the order of their corresponding pages and are ordered as such. To expand on result item you may click on an image in the table and you will be shown the dialog in figure 13. This dialog provides extra information about the image including the image's width, height and URL. Similar to web result item dialogs, you are given an 'Open in browser' button and 'Copy to clipboard'. The first button will attempt to open the image in your default browser while the second button will copy the image's URL to your clipboard.

Figure 13: Image result item dialog