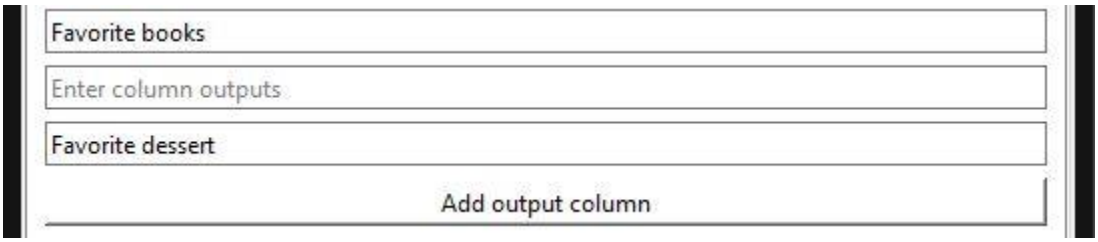
Some notes to better facilitate use of the web-scraping tool

On changing the output

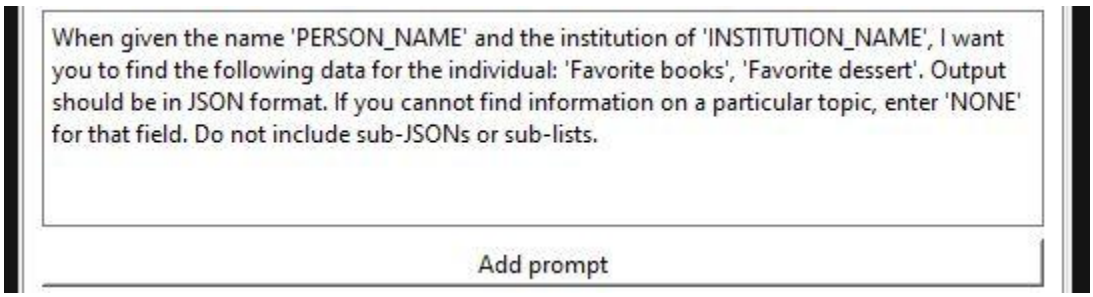
Using the tab entitled ‘Alter output’, the internal workings of the tool can be reformatted in several notable ways: the output columns of the Excel file, the prompts used by GPT when analyzing text taken from websites, and search terms used by the tool when gathering links. Furthermore, the name of the output file can be chosen in the textbox labeled with ‘Output file name’. (If no name is chosen, then the output file will take the name of the input csv and add “_output” to the end).

Columns of the output Excel file can be changed by adding new ones to the text input of the output column section (pictured below)



Any column can be entered here and that will appear verbatim on the output file. If there is a blank slot left, then that will be ignored. The output Excel for this set of columns would only have ‘Favorite books’ and ‘Favorite dessert’.

Prompts given to GPT can be changed in the prompt section, seen below.



This particular prompt was gotten by pressing the ‘Generate prompts’ button. This will generate prompts based on the inputted columns. In order to get the best results from GPT, the tool will fit a maximum of three requests per prompt. If there are four inputted headers

and prompts are generated by the tool, then there will be two prompts, one with three requests and the other with only one.

The text “PERSON_NAME” and “INSTITUTION_NAME” are stand-in variables that will be switched out with the pertinent name/institution from the input csv file, as each name is processed. For example, if the input csv file contains “Frodo Baggins” as a name and “Gentleman” as an institution, the prompt would automatically be converted to

> “*When given the name 'Frodo Baggins' and the institution of 'Gentleman', I want you to find the following data for the individual: 'Favorite books', 'Favorite dessert'. Output should be in JSON format. If you cannot find information on a particular topic, enter 'NONE' for that field. Do not include sub-JSONs or sub-lists.*”

when Frodo’s turn comes to be scraped. You are able to build or alter prompts as you see fit, maybe to add clarifying information for a specific prompt, but without requesting that the output be in JSON format the tool will very likely fail to output anything.

If you want the Excel output to have a certain column but not scrape any information for that column, simply include that column in the header section and keep it out of the prompt section.

If you want to include columns from the input csv, just put them in the header section and keep them out of the prompt section.

If processing starts and no prompts are selected, the tool will automatically generate them.

There are a few output columns that will not be included in automatically generated prompts:

- “email”, “emails”, “Email”, or “Emails”
    - Emails are scraped using a method separate from GPT (regular expression) and so aren’t included
- “Other key notes”, “other key notes”
    - Due to the origin of the tool, this is reserved specifically to hold information on patents and awards
- “Relevant links”, “relevant links”
    - This column holds all links used to scrape
- “Name”, “name”
- “Institution”, “institution”

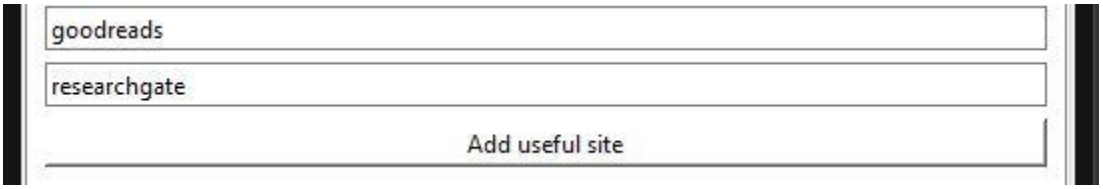
- Name and institution are required to be in the input csv, so if you want them in the output just include them in the header.

If you want to save an output format, you can do so by pressing the ‘Save formatting’ button. Saved formats can be loaded using the drop down menu.

IF there is no input whatsoever given and the drop down menu is set to its default text, then the scraping option for scientenometrics will be used, which this tool was originally developed for.

There is currently no way to delete saved formats (they can be overwritten by using the same name when saving) due to a fear of cluttering the GUI. If an output needs to be deleted, they can be found in the folder named ‘/saved_output_formats/’ where this tool is loaded from. They are saved as txt files.

Additional sites can be used for searching by adding them to the search section.



Specifying these as additional links will lead to additional search terms, i.e.

- Frodo Baggins Gentleman
- Frodo Baggins Gentleman goodreads
- Frodo Baggins Gentleman researchgate

Would be the searches used by the tool in finding good sites to scrape. If no additional sites are specified, then only the first search will be used.

Due to concerns on scraping and information validity, the following locations will be disregarded in scraping

- Facebook
- Instagram
- LinkedIn
- Twitter
- ratemyprofessors
- Coursicle

- YouTube
- Amazon
- Wiki (of any sort)
- .doc
- .pdf
- imgres

Basic flow of the tool

After the user interacts with the GUI and presses the ‘Process’ button, the tool will go through its internal scraping cycle.

It will start by loading the names/institutions in the input csv file. Then, for each name, searches will be performed to find good links. For each good link gotten, the text on that page will be analyzed and desired information will be retrieved, using the inputted prompts (or potentially regular expression). This information will then be written to the output Excel file and scraping will move on to the next person in the csv. This format of scraping then saving ensures that if the tool runs into some error late into the process of scraping, there will still be saved output of all the people it managed to scrape before the error. The log and table exists to keep users apprised of particulars of the process, as they happen.

For any questions relating to the tool, feel free to email me at evandzook@gmail.com