Google Scholar Analyzer Manual

Table of Contents

1 Introduction

1.1 what it is

Google Scholar Analyzer (GSA) is a small python program I built in an attempt to know researchers, whose academic profiles are available on Google Scholar website, in a way that is much more straightforward. I found it increasingly necessary when I was trying to search for a PhD program while having no clues as to what kind of supervisors I would be interested working with in the future. Although scrolling down scholars' CVs or publications does help, most of time that process takes time and there appears to be no guarantee that the key information will be absorbed or saved (people usually do not take notes here) for future reference. For people who are smart and take notes, however, if you are not equipped with so-called photographic memory, or the ability to remember like an elephant, this program might still come in handy for you.

Anyway, I believe GSA is of high practical values. At one time, I would like jokingly to call it something like "Find you supervisor", but since it literally benefits anyone who want to get key information out of a Google scholar, not just PhD seekers, the current name will be better.

Running this program will get you a txt file saved in an auto-created folder named after the scholar's affiliation so that scholars from the same organization will be put together. The file contains the following information:

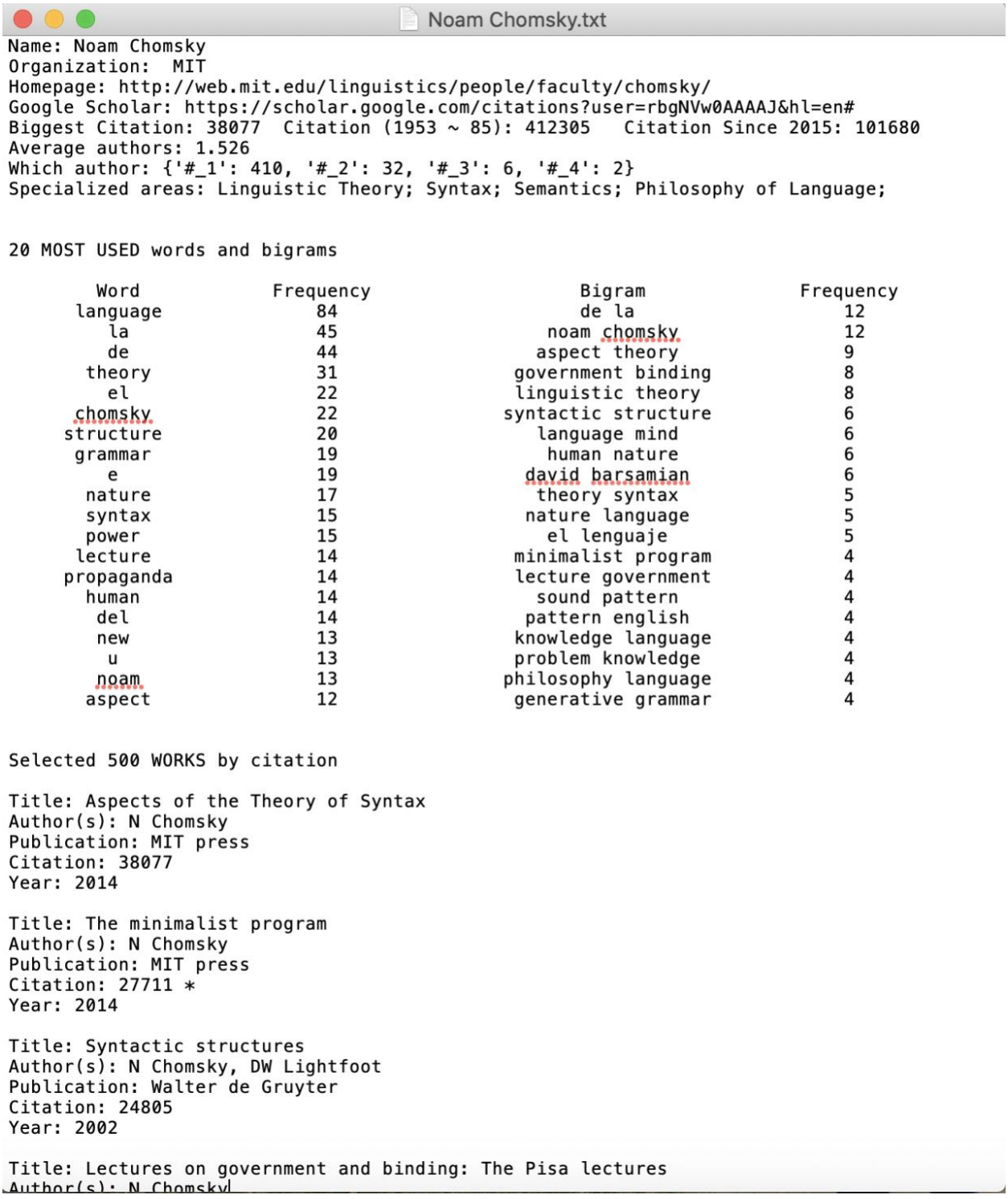
1. Basic info about the scholar, including name, organization, homepage link (if provided), the Google Scholar link, brief citation report (biggest citation, time-specified citation of all time, citation since a recent year, in our case since 2015), contribution (by looking at the average authors of all works extracted and the current scholar's positions in them) and specialized areas.
2. Most used one token or two (usually token = word) in the titles of all works or selected most cited 500 works of the interested scholar. Users can specify the number of most used tokens at their own discretion. The default value for that is 20.
3. A breakdown of all the scholar's works or most cited 500 works if he/she has more than 500 works available on Google Scholar website. This will first come with the specified number of works extracted, the average author(s) of these works, statistics on the contribution of the current scholar as to which author he/she is exactly in, and a full list of works with titles, authors, citation, and year for later lookup.

GSA is freely open to anyone who might find it useful. Anyone with programming background can modify the codes and thus revise the program to function as they want without notifying me, although it would be appreciated if you do.

The program is mostly user-friendly, so no programing knowledge is expected. Here.

1.2 how the saved file looks like

Below is how a demo looks like after running the program:



2 Instruction

2.1 Getting started

To use GSA, you need to have Python version 3 .5 or above installed on your computer. To do it, simply go to here, and download the latest version.

The following two packages are also required:

- NLTK. The instruction for the installation is available here. The current codes involve use of "stopwords" from NLTK data to avoid counting words like pronouns, function words, be verbs and the likes (the full list can be seen below). Users can either choose to download "all-corpora" by implementing the command *nltk.download()* right after importing NLTK package, or just copy and paste the following equation and replace the command in the codes stopwords = *stopwords.words('english')* with it.

  stopwords = ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]

- selenium. Users are referred to here to install this python package along with a driver associated with it. In our codes, we are using *webdriver.Chrome()*. Users are able to change this command at their discretion according to their most used browser. You can also see this if you encounter any difficulty installing Chrome Driver. For Mac user, you can go to **Preference's Advance setting** in Safari, and enable Develop menu in the end. Click Develop right above the browser, turn on **Allow Remote Automation**, change the command *wd = webdriver.Chrome()* into *wd = webdriver.Safari(),* and you are good to go!

2.2 How to use

There are two values that users need to input first when running GSA source codes:

1. The url link of the interested scholar on Google Scholar website, such as: https://scholar.google.com/citations?user=rbgNVw0AAAAJ&hl=en# (Noam Chomsky). Please note that the link has to be the one without clicking any buttons or links because that might change the link.
2. The root where you want to save the extracted file into your computer, such as: "/Users/wzx/Downloads/Google Scholar Profile/" (on mac).

Users can also change at their discretion two values as follows:

3. The number of most frequently used one token or two in the titles of all the scholar's works to display. The default value is 20.
4. As demonstrated above, if users have not installed the NLTK data, they can replace the command stopwords = *stopwords.words('english')* with the provided equation.