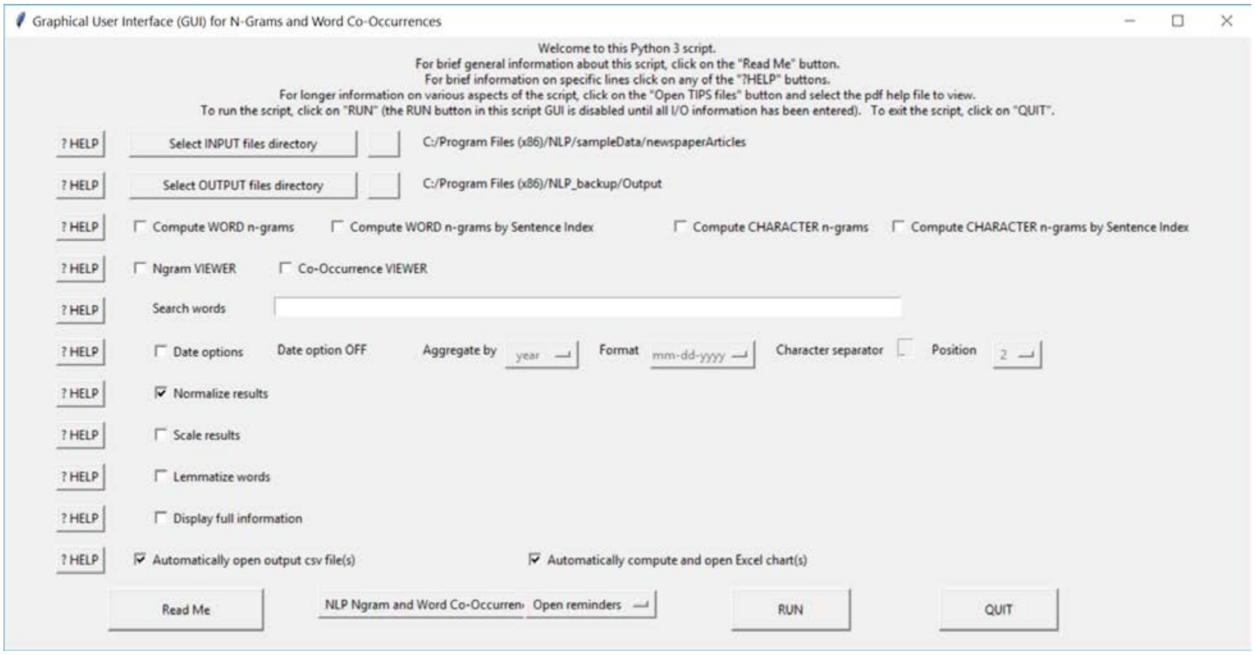
NLP Ngram and Word Co-Occurrence Viewer

A GUI for N-grams, Ngram and Word Co-Occurrence Viewer

The NLP Suite provides a specialized GUI for computing and plotting N-grams and for viewing N-grams and word co-occurrences (the NLT Ngram and Word Co-Occurrence Viewer).



Graphical User Interface (GUI) for N-Grams and Word Co-Occurrences

Welcome to this Python 3 script.
For brief general information about this script, click on the "Read Me" button.
For brief information on specific lines click on any of the "?HELP" buttons.
For longer information on various aspects of the script, click on the "Open TIPS files" button and select the pdf help file to view.
To run the script, click on "RUN" (the RUN button in this script GUI is disabled until all I/O information has been entered). To exit the script, click on "QUIT".

Select INPUT files directory: C:/Program Files (x86)/NLP/sampleData/newspaperArticles
Select OUTPUT files directory: C:/Program Files (x86)/NLP_backup/Output

□ Compute WORD n-grams □ Compute WORD n-grams by Sentence Index □ Compute CHARACTER n-grams □ Compute CHARACTER n-grams by Sentence Index
□ Ngram VIEWER □ Co-Occurrence VIEWER
Search words
□ Date options Date option OFF Aggregate by year Format mm-dd-yyyy Character separator Position 2
☑ Normalize results
□ Scale results
□ Lemmatize words
□ Display full information
☑ Automatically open output csv file(s) ☑ Automatically compute and open Excel chart(s)

Why not use the Google Ngram Viewer?

Google has a powerful freeware tool for viewing N-grams (https://books.google.com/ngrams). Why, then, bother writing a new tool for the NLP Suite? A case of re-inventing the wheel, and probably not as good as the Google wheel? Good question. But… as the Google Ngram Viewer url shows, the url contain the word *books*. Indeed, the Google Ngram Viewer applies to the millions of books digitized by Google (some 25 millions! https://www.edsurge.com/news/2017-08-10-what-happened-to-google-s-effort-to-scan-millions-of-university-library-books).

If you have your own corpus, made up of blogs, of your own in-depth interviews, you could not use the Google Ngram Viewer. This is why we created a Java tool (NGrams_CoOccurrences.jar), so that you can visualize n-grams with your corpus, just like the Google Ngram Viewer.

How does the NLP NGram and Word Co-Occurrence Viewer work?

The tool works in exactly the same way as the Google tool. You enter keyword searches and you search, with the option of viewing N-grams and/or word co-occurrences.

Ngrams search

The NGrams_CoOccurrences.jar is a Java routine that allows searches for Ngrams, i.e., key words (e.g., "nursery school" (a 2-gram or bigram), "kindergarten" (a 1-gram or unigram), and "child care" (another bigram) that occur in different documents within a selected time period (e.g., month, year). It works similarly to **Google Ngram Viewer** except this routine works on documents supplied by the user rather than on the millions of Google books (see https://books.google.com/ngrams/info).

The routine relies on the Stanford CoreNLP for lemmatizing words.

The routine will display the FREQUENCY OF NGRAMS (WORDS), NOT the frequency of documents where searched word(s) appear, as with Word Co-Occurrences.

Date in filename required

The NGrams part of the NGrams_CoOccurrences.jar routine requires date metadata, i.e., a date embedded in the filename (e.g., The New York Time_2-18-1872).

Normalization

Hovering over each data point in the Excel line chart will display the following information: the Group size (i.e., the number of all available documents at that specific data point, regardless of whether any of the documents contain any of the searched words) and the total number of documents in the corpus.

Inherent bias

THE INHERENT BIAS OF SUCH A SEARCH IS THAT A SPECIFIC EVENT (E.G., AN "ASSAULT") THAT IS MENTIONED REPEATEDLY IN A SINGLE DOCUMENT WILL LEAD TO A HIGH FREQUENCY, ALTHOUGH THE ACTUAL NUMBER OF DISTINCT "ASSAULT" EVENTS MAY BE LOW.

Word Co-occurrences search

The word co-occurrences part of the Java routine will allow searches for word co-occurrence, i.e., key words that occur together in the same document.

The routine uses a SET OF TEXT FILES in input and produce an Excel line chart in output. **The word co-occurrences part of the NGrams_CoOccurrences.jar routine DOES NOT require date metadata, i.e., a date embedded in the filename (e.g., The New York Time_2-18-1872).**

The routine will display the FREQUENCY OF DOCUMENTS where searched word(s) appear together in the same document, NOT the frequency of the searched word(s) as with NGrams.

Hovering over each data point in the Excel line chart will display the following information: the Group size (i.e., the number of all available documents at that specific data point, regardless of whether any of the documents contain any of the searched words) and the total number of documents in the corpus.

Inherent bias

THE INHERENT BIAS OF SUCH A SEARCH IS THAT A SPECIFIC EVENT (E.G., AN "ASSAULT") THAT APPEARED IN MANY DCUMENTS WILL LEAD TO A HIGH FREQUENCY, ALTHOUGH THE ACTUAL NUMBER OF DISTINCT "ASSAULT" EVENTS MAY BE LOW.

Input

The routine uses a SET OF TEXT FILES in input.

NGRAM AND WORD CO-OCCURRENCE SEARCHES DO NOT MAKE MUCH SENSE WITH A SINGLE FILE; A POINT WOULD BE PLOTTED!

Output

In output the Java routine produces an Excel line chart with hover-over effects.