Ngrams and Word Co-occurrences Viewer

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 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).

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.

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.

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.

The routine uses a SET OF TEXT FILES in input and produce an Excel line chart in output.

NGRAMS DO NOT MAKE MUCH SENSE WITH A SINGLE FILE; A POINT WOULD BE PLOTTED!

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.

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.

The routine uses a SET OF TEXT FILES in input and produce an Excel line chart in output.

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