Above, we explored the notion that semantic differences between words, or differences between senses of one word, can be identified on the basis of the collocates of the word or words. In this assignment, you will look at the probably most widely-cited example of a homonym, namely bank. The idea is to determine the collocates that are most strongly attracted to the noun lemma BANK, i.e., bank and banks. In order to determine these collocates, however, you must retrieve all matches for the word forms bank and banks, but you also need to be able to retrieve the frequencies of all collocates of BANK and determine the overall corpus size, which is best done by at the same time generating a frequency list. (Note, this procedure of generating a vector with the whole corpus or a frequency list on the side is not useful when the corpus is large. In such cases, it is better to compartmentalize tasks and generate a frequency list of a large corpus separately and load it when required.)

Assignment

Write a script that has the following characteristics and performs the following operations:

(i) The script prompts the user to choose the BNC files <C:/qclwr/inputfiles/corp_bnc_sgml_*.txt> (files 1 to 4).

(ii) The script loads each of these files, converts them to lower case, and retrieves only the lines containing sentence numbers.

(iii) While going through all corpus files, the script
    - generates a frequency list of all corpus files;
    - retrieves all collocates of matches of bank or banks as words that occur up to four words preceding or following bank or banks.

(iv) For each of these collocates of bank or banks, the script then determines
    - its observed frequency, its expected frequency, and whether the observed frequency is larger or smaller than, or equal to, the expected frequency;
    - computes two statistics for this co-occurrence relation: chi-square (as above) and mutual information MI, which is defined as the logarithm to the base of two of the quotient observed frequency by expected frequency.

(v) All this information is then stored in a data frame for subsequent analysis. Inspect the first 50 matches to see which senses of BANK you think are represented by the collocates.

Load the file <C:/qclwr/scripts/semantics_3-assignment_colls.r> and <C:/qclwr/outputfiles/semantics_3_colls.txt> when you are done and compare your solution with them.