

Additional exercises: increasing the context of an earlier concordance

Sometimes, the first way in which a particular concordance is generated turns out to be insufficient for steps that become later necessary. In this additional exercise, we will consider a case I recently encountered. Using/tweaking code from the book, a student recently wrote an R script to generate a huge concordance for a lemma such that both the preceding and the subsequent context consisted of six words, which were tab-separated from the match in the middle. This output is exemplified in the following table (available as <additional_11_increase-concordance.txt>).

PRECEDING	MATCH	SUBSEQUENT
verbatim·copies·of·this·license· document,·	but	·changing·it·is·not·allowed··Preamble·
if·the·Program·itself·is·interactive·	but	·does·not·normally·print·such·an·
you·distribute·them·as·separate· works··	But	·when·you·distribute·the·same· sections·
in·spirit·to·the·present·version,·	but	·may·differ·in·detail·to·address·
KIND,·EITHER·EXPRESSED·OR·IMPLIED,· INCLUDING,·	BUT	·NOT·LIMITED·TO,·THE·IMPLIED· WARRANTIES·
INABILITY·TO·USE·THE·PROGRAM· (INCLUDING·	BUT	·NOT·LIMITED·TO·LOSS·OF·DATA·
hope·that·it·will·be·useful,·	but	·WITHOUT·ANY·WARRANTY;·without·even· the·

These data were then annotated for a large number of features, but after quite some morphological and syntactic annotation, it then turned out that the six-word context was insufficient for some of the intended semantic annotation. At the same time it was not straightforwardly possible to just re-run the concordance and retrieve more words because the results' order had undergone many changes. In order to not write a new script and re-do all the annotation that was already available, the problem was how to retrieve more context material for each already annotated match. (We will only consider the simplest case here, where the corpus from which the concordance was created consists of only one file.)

Assignment 1

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

- (i) The script loads the above concordance table into a data frame (from <additional_11_increase-concordance.txt>)
- (ii) The script
 - loads the original corpus file into a vector;
 - cleans up whitespace by deleting tabs and excess whitespace;
 - merges the whole corpus into a character vectors of length one;
- (iii) The script 'takes' each preceding context from the above concordance table and
 - locates all matches of the current preceding context in the corpus vector;
 - extracts from the corpus vector the 300 characters before the current match (i.e., the 300 characters that end with the preceding context in the above table);
 - extracts from the corpus vector the 300 characters after the current match (i.e., the

- 300 characters that begin with the subsequent context in the above table);
 - stores the new preceding and subsequent context(s) in a list. (NB: the script can handle cases where the search for the preceding context returns more than one match, where the user will then decide which of the retrieved matches is the correct one.
- (iv) The script saves the new preceding and subsequent contexts into a text file such that
- the first column of the output contains the number of matches of each preceding context;
 - the second column contains the (first) preceding 300-character context candidate that is supposed to replace the preceding context in the above table;
 - the third column contains the (first) subsequent 300-character context candidate that is supposed to replace the subsequent context in the above table;
 - if there was more than one match for a preceding context, then the columns 4 and 5, 6 and 7, 8 and 9, etc. contain the next preceding and subsequent context candidates (so that a user can then identify the right context) to be pasted.

To maybe get a better understanding of what this amounts to, cf. <additional_11_increase-concordance_output.ods>. The cells B2:D9 contain the input concordance, and what the above script is supposed to create is the content of the cells A12:B19 and D12:D19 (I only left the cells C12:C19 empty so the preceding and subsequent contexts are aligned and one can see better how D13 is the new, much longer, context compared to D3, etc.).