Computerized Text Analysis: Classwork 2
Basic Descriptive Text Statistics

Kenneth Benoit

In this class we will continue to work with some of the texts that you set up in QDA Miner in Exercise 1. We will use QDA Miner to explore these texts and generate term-document matrixes, and matrixes that tabulate word-frequencies across user-created variables. You will need to open the UK manifesto project that we created yesterday. To do this, select 'Open an existing project' from the dialog box when QDA Miner starts (or choose 'Project → Open') and select the UK manifesto project that you saved yesterday.

If you have problems locating or re-opening the project from yesterday, then you can instead create a new project, and add the UK manifesto files from the texts folder on the S: drive. In this exercise we will view a word-frequency matrix created from the documents.

1. To be able to tabulate the word-frequencies with variables, we first need to code some documents with variables of interest, so, if you have not already done so, create a variable for 'PARTY' and code some of the texts according to the party indicated in their filename. There is no need to code all of the texts.
2. Open Wordstat by choosing 'Analyze → Content Analysis. The main QDA Miner window allows the management of cases, variables, and codes, and the Wordstat module is specifically designed for text analysis.
3. A window will open asking you to choose how you want to analyze the cases. In this case, we want to analyze all of the text in the documents, in relation with the PARTY variable, so choose 'Documents' and 'all text' in the upper section and 'other variables' and PARTY in the lower section.
4. Examine the options available in the Wordstat window. The 'frequency' and 'cross tab' windows allow you to view a word frequency matrix.
5. By default, an English exclusion dictionary is selected. Unselect this dictionary to view the raw word frequency matrix.
6. Use the Options section to pre-process the text. Experiment with excluding low-frequency terms. Exclude terms that occur fewer than 5 times. Exclude terms that occur in fewer than 3 cases. Exclude terms that occur in more than 80% of cases. After changing each option, when you return to the cross-tab or frequency window the word frequency matrix will be recalculated.
7. Perform stemming on the texts by selecting the pre-processing box and choosing the Porter stemmer.
8. Select the 'LaverGarry' categorization dictionary. You can examine the terms included in the dictionary by clicking on the folder beside each concept to expand it. Wordstat will sum the frequencies of words under each concept to give a high level impression of the document's focus.
9. Examine the word and concept frequencies in relation to PARTY in the cross-tab window. Consider how the raw frequency counts are affected by how many documents you have available for each variable. If many more documents are available for one variable than another, the 'column percent' option might be more relevant.