

Computerized Text Analysis: Classwork 2

Basic Descriptive Text Statistics

Kenneth Benoit

The objective of this class exercise is to continue working with the French manifesto texts from Classwork 1, as well as some texts from the Comparative Manifesto Project. 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 project that we created yesterday, or make a new one using the French texts. If some of the texts have not imported correctly, you can import them as follows:

1. Open the file with its associated program (a browser for html files, notepad for txt files, microsoft word for doc files)
 2. Highlight all the text by pressing ctrl-a, copy the text (ctrl-c), and paste it directly into the text viewing window in QDA Miner (ctrl-v).
- or
3. Choose to save the file as plain text, and then import this plain text file into QDA Miner by choosing "Cases, append from documents" (if you do this, you should delete the incorrectly imported version to avoid having duplicates, right-click on the document name and choose 'delete'.

Instructions:

1. Inspect the converted French texts from Day 1 in QDA Miner. If any encoding-related errors remain (for example the ligature in 'oeuvre' we saw yesterday), you may be able to manually fix them using the search and replace function under the document tab. You may also wish to remove converted formatting characters such as page numbers, footers, and headers. Since none of this information is part of the text we are interested in analyzing, we should remove it.
2. From the top menu bar, choose "Analyse, Content analysis" to open the WordStat content analysis tool. The word frequency matrix analysis is available under the 'frequency' tab. To see a matrix of word frequencies by document, choose the 'crosstab' tab, and choose tabulate with 'FILE'
3. Experiment with the options under the Dictionary and Options tabs. Create a lemmatized frequency matrix by selecting the French lemmatization dictionary.
4. Save this project, and make a new project using the UK manifesto texts. Check that the manifesto texts have imported correctly.
5. By default, each case is associated with two variables - the filename and the document itself. For the UK texts, we have two additional pieces of information - the year and the political party. To add this information to each text, we need to create new variables to represent these values
6. In QDA Miner, choose "Variable, add variable" from the menu bar. Create a new ordinal variable called PARTY with the values LIB DEM, LAB, and CONS. Once you have created this variable, you can add it to each case by selecting the case, and choosing the appropriate value for the variable from the drop-down menu in the VARIABLES sidebar. Do this for year too, using the variable type DATE.

7. Once you have added variables to the files, start wordstat again, and choose to analyze in relation with the variables PARTY and YEAR. Under the cross-tab section, you can now see the word frequency ratios between the parties and across years.
8. Inspect the matrix: by party and by year, using the raw counts and the ratios of word use per column.