Assignment 5: Supervised Methods for Classifying and Scaling Texts

Kenneth Benoit

This exercise involves using the automatic document classification features of WordStat, using texts from movie reviews (files here) (Pang and Lee 2002) and then from Evans et al (2007) *amicus curiae* briefs (files here).

Instructions

1. Load the movie review texts into QDA Miner. After creating a new project, begin by loading the positive reviews, and use the spreadsheet editor to code all of these with under a new variable type - Sentiment - with the value POS. Then load the reviews from the negative folder and give them the variable value NEG. Make sure these are Categorical variable types.

2. Open WordStat with the parameters as follows: 'Analyse all text in relation with Variable SENTIMENT'.

3. Choose the automated text classification button (3rd from the left, bottom row, in the Crosstab panel)

4. Try the different options in the 'Learn and Test' panel and observe the results. Note the different options for performing cross validation.

5. Construct a systematic exploration of the parameter space with the experiment button on the history panel.

6. Repeat the experiment, but choose a much smaller set of examples. What is the relationship between the accuracy and the size of the training set?

7. Create a new project for the Evans et al amicus briefs. Import all of the texts in the "training" and "testing" folders. Create a variables for "SET" (training or test) and "Class" (petitioner or respondent).

8. Predict the category of petitioner versus respondent for the *amicus* briefs using only the training briefs. You can choose which documents to predict from the 'Apply' tab by selecting 'list of documents' and 'edit list'.

9. Experiment with feature selection to see if predictive accuracy can be improved.

10. Now we will work with the Wordscores text scaling algorithm, using examples from Laver, Benoit and Garry (2003). I recommend that you use R for this, along with the quanteda library, which contains the austin package.

    To install the libraries you need, follow the instructions at `https://github.com/kbenoit/quanteda` under the 'How to Install' section.

    We will use two sets of files:

    - The example from Table 1 of LBG (2003). This data is built in to the `austin` library in R so if you are using R to do this, then you will not need to load this file in at all. If you are using some other software, then you can dowlonad the file `LBGexample.csv`. This file is in .csv (comma separated value) format and can be loaded directly into Stata or your spreadsheet.
    - The Irish 2010 budget speeches data, available when you type

```
library(quanteda)
data(iebudgets)
dtm <- dfm(subset(iebudgets, year==2010))
```

The R file `http://www.kenbenoit.net/courses/tcd2014qta/exercise5.R` has all of the commands you will need to implement the steps outlined below, including the installation of the austin library. Follow these steps:

(a) Start Rstudio and open the exercise8.R file in Rstudio

(b) Install the `quanteda` library as per the instructions in the github website linked above.

(c) Estimate the wordscores model for the LBG (2003) example, and inspect the results. Follow the code for precise instructions. Here you will be using the reference scores set at -1.50, -0.75, 0.00, 0.75, and 1.50 for reference texts r_1 through r_5 respectively. Score the virgin text and compare your results to LBG (2003) Table 1.

(d) Run the wordscores scaling procedure on the Irish 2010 Budget speeches. Here we will use the 5th text (Cowen, the FF Prime Minister) as one reference text, and the 6th text (Kenny, the FG opposition leader). We will score all words in the Cowen and Kenny texts, and then score all texts as if they were purely "virgin" documents.

(e) Inspect the word scores as shown in the .R file.

(f) To run the naive Bayes example using amicus briefs, you can try the example in the README.md file from `http://github.com/kbenoit/quanteda`.