Computerized Text Analysis: Classwork 9
Document Classification Algorithms

Kenneth Benoit

This exercise involves using the automatic document classification features of WordStat to experiment with feature selection and model evaluation **Instructions:**

1. Load the movie review texts into QDA Miner. 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
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. Choose a political classification task from the texts we have worked with previously (UK Manifestos and Irish Budgets), and see what accuracy you can achieve (e.g. predict Party Membership of speaker, or year of manifesto - you'll have to hand-label any variables you haven't already saved).
8. Instead of using raw word frequencies, apply and appropriate dictionary (i.e. sentiment for the movie reviews, Laver and Garry for political text) before running the classification functions. Do the dictionaries improve performance?