Unsupervised Methods for Scaling Texts: Lab Exercise

Kenneth Benoit

This exercise involves using the automatic document scaling using correspondence analysis in WordStat and in R, and the Poisson scaling midel in R. For texts we will use the 2010 Irish budget speech corpus that you have worked with on previous days (and available here if you are missing them: (2010 Irish Budget debates).

Instructions

1. Correspondence analysis in Wordstat

   (a) In QDAMiner, reopen the project for the Irish 2010 budget speeches you have worked with in previous assignments.

   (b) Launch Wordstat, and choose the Analyze tab and the Correspondence Analysis button.

   (c) Explore the results and test using different options.

   (d) Now return to QDAMiner and add the file `2010_BUDGET_11_John_Gormley_Green_ENTIRE.txt`, which represents the Gormley speech with the discussion of water meter text.

   (e) Redo the correspondence analysis on all texts, and inspect the two Gormley positions. Now remove one of the Gormley positions and reanalyze. (You can "remove" a text by deselecting it in QDAMiner and relaunching Wordstat.)

2. Correspondence analysis in R

   (a) Launch an R session, and issue the command `library(quanteda)` to load the `quanteda` library and `library(austin)` to load the `austin` library. You should have installed these yesterday.

   (b) The debate over the Irish 2010 budget speeches are already available from the `austin` library as a data object (which is somewhat confusingly called "iebudget2009"). To access this you simply use the command

   ```
   data(iebudget2009)
   ```

   (c) Check the column names (variable names) and size of the dataset you've loaded using:

   ```
   colnames(iebudget2009)
   dim(iebudget2009)
   ```

   (d) You will need to install the `ca` library for correspondence analysis, which can be installed and loaded using the commands

   ```
   install.packages("ca", dependencies=TRUE)
   library(ca)
   ```

   (e) To fit and plot the correspondence analysis in two dimensions, use these commands:

   ```
   c1 <- ca(iebudget2009)
   # remove the "what" argument above to plot words as well
   plot(c1, what = c("none", "all"))
   # plot the speakers in one dimension, ordered
   dotchart(c1$colcoord[order(c1$colcoord[,1]),1],
            labels = c1$colnames[order(c1$colcoord[,1])])
   ```

   (f) Compare these results (from R) to those you obtained from Wordstat.

3. Poisson scaling in R

   (a) To set the orientation of the estimation you will need to note which column is Joan Burton (opposition anchor from Labour) and which is Brian Lenihan (Finance Minister).

   (b) Estimate the wordfish model using the following command. (The column indices in the "`dir`" vector refer to 11 for Burton and 4 for Lenihan.)

   ```
   wfm <- wordfish(iebudget2009, dir=c(11,4))
   ```

   (c) Summarize the plot the results:

   ```
   summary(wfm)
   plot(wfm)
   ```

   (d) Plot the $\hat{\psi}$ by $\hat{\beta}$ and interpret the plot. Use the following commands:

   ```
   plot(wfm$beta,wfm$psi,type="n")
   text(wfm$beta,wfm$psi,wfm$words)
   ```