Quantitative Text Analysis
Exercise 4: Comparing Texts

24 July 2014, Essex Summer School

Kenneth Benoit and Paul Nulty

In today's lab we will continue comparing texts.

Instructions

1. **Detecting collocations**

   (a) Load the inaugural corpus using `data(inaugCorpus)`. Use the `collocations` command on the two George Bush speeches to inspect the top 50 collocations. Now try it with using the χ^2 measure instead of the default likelihood ratio measure.

   (b) Try the same thing for Obama's speeches.

2. **Document similarity**

   (a) Compute the cosine similarities between the Obama and Bush speeches (you should select these using the corpus subset commmand, and then create a dfm for this subset). Follow the model from class.

   (b) Compute a Euclidean distance for the same set.

   (c) Extra credit: convert the cosine similarity into a distance, and the distances from the previous two into a vector, and plot them against one another.

   (d) Convert the dfm objects to a binary feature matrix, and recompute both distances (as per the Choi et al paper).

3. **Resampling texts**. Here we will extract the 2009 Obama inaugural address using `subset`, and reshape it into a sentence-level corpus. Then we will extract the vector of sentences using `getTexts`, and resample it.

   (a) Extract a subset of the `inaugCorpus` set for Obama's 2009 inaugural adress.

   (b) Reshape this into a sentence-level corpus using `corpusReshape`.

   (c) Extract the texts to a character vector object using `getTexts`.

   (d) Produce a "possible" 10-sentence speech from Obama using this command:

   ```
   paste(sample(obamaSentenceVector, size=10, replace=TRUE), collapse=" ")
   ```

   (e) Repeat the last step several times and observe the texts that result. Do these sound like Obama?