## More demos for tf-idf and similarity measures

You'll find useful demos (front ends to SOAP Web Services, also available) of tf-idf and similarity measures provided by IULA, Universitat Pompeu Fabra, Barcelona, one of our partners in META-SHARE (http://www.meta-share.eu/) as follows:

### 1) tf-idf

http://lod.iula.upf.edu/resources/109

The demo itself is at:

http://ws04.iula.upf.edu/soaplab2-axis/#statistics_analysis.tfidf_row

You can input your own text either by URL or by browsing and uploading. Input files are expected to be in plain text format. *The demo works best when given more than one document in the same file*. Use '-----.' (5 hyphens followed by a full stop) on a line by themselves to separate documents in the same file. I suggest you just concatenate some of the files we were using in the workshop (radon, amnesty, tanks, …), separated of course by '-----.' into one file for upload. On the '109' link above you will also find URLs for example inputs (these are big files so will take some time) and for example outputs.

When the process has completed, click on the blue output link to see the output. The first output you will see will be hyperlinks for text0..textn. Clicking on one of these will give you all the words in that file (however, a stoplist appears to be used). The larger the word, the greater its tf-idf score. Clicking on a word will give you a bar chart and statistics for total frequency of occurrence (collection frequency of occurrence), term frequency (frequency of occurrence in the document) and tf-idf scores. If the file is large you may have to scroll down to see the chart and stats.

Remember to *remove* your results once finished, to be polite to other users.

### 2) similarity measures

http://lod.iula.upf.edu/resources/429

(based on work by Ted Pedersen)

The demo itself is at:

http://ws04.iula.upf.edu/soaplab2-axis/#statistics_analysis.text_similarity_row

Here, you give as input two plain text files, but this time one document per file. It is best to click the 'verbose' radio button to get the relevant statistics. Clicking on the 'similarity' result link will just show a single real, but if you have clicked the 'verbose' button then you will see at the end of the Report output several scores according to various measures. Dice is one that is often used (a variant of the equally popular Jaccard similarity measure). You will see listed there also the score for the **cosine** measure discussed in the workshop. Lesk is often used when measuring similarity of word senses (word sense disambiguation task). Remove results when finished.

For the interested (or those doing relevant final year projects), there are many more Web Services for various tasks at: http://lod.iula.upf.edu/types/Service