Information Retrieval

Lab Assignment 1

Paris Mavromoustakos
Partick De Kok

November 11, 2012

1 Introduction

In this first lab assignment, we had to cope with both indri 5.3 and trec_eval applications' environment and use them to index given sets of documents, apply certain queries and evaluate the results generated.

We Installed indri 5.3 and trec_eval on Mac OS X version 10.7.5 using the Terminal application, which we further used to run all the commands needed for this assignment.

2 Indexing

First of all, after downloading both sets of documents (LAT & GH95) from the link given in Blackboard, we unzipped the document files contained in both folders and added them all together in one single folder. This was the folder to be indexed, using IndriBuildIndex found in folder /buildindex of the indri installation directory. In order to index this folder, we first needed to create the "parameters.xml" file which defines the details of the indexing procedure. The parameters.xml file looks like this:
$< parameters >$
$< index > index\_outputDIR < /index >$
$< corpus >$
$< class > trectext < /class >$
$< path > index\_inputDIR < /path >$
$< /corpus >$
$< stopper >$
$< word > word1 < /word >$
$< word > word2 < /word >$
$< word > word3\ word4 < /word >$
$< /stopper >$
$< stemmer >$
$< name > stemmer\_name < /name >$
$< /stemmer >$
$< /parameters >$

Where, *index* is the directory where the indexing results will be saved, *class* defines how the documents will be processed (they will be considered as "trectext" documents in this example), *path* defines the folder containing the documents to be indexed, *stopper* incudes possible stopwords (for example, word1, . . . ,word4) and *stemmer* defines the stemming method. It is essential to point out that *stopper* and *stemmer* are not obligatory, and they can be excluded from the document.

To start indexing the documents, all we needed to do was run *IndriBuildIndex parameters.xml* in the Terminal. For this command to run properly, the present working directory should be /buildindex and the parameters.xml file should be stored in that directory aswell. While running, the terminal will print out the status of the indexing process, including the time it takes to run the process and the number of documents indexed. Indexing both sets of documents (LAT & GH95) took us 1:26 minutes and created 169477 indexed documents.

3 Checking the indexing results

In order to check our indexing results, we ran the command *dumpindex index_resultsDIR s* while working in the indri installation directory. The *index_resultsDIR* is already known from the previous step (included in the parameters.xml file) and the parameter *s* at the end of the command represents "status", meaning that it will print general info regarding the indeing results. The output we got for indexing both sets of documents without using stopwords or stemming, is the following:
Repository statistics:
documents: 169477
unique terms: 335273
total terms: 88270885

However, if we run the command using *v* instead of *s*, we would get a comlpete "vocabulary" table of all unique terms, which would look like this:
the 5071448 164890
to 2249861 159030
of 2210672 161030
a 2136668 160832
where the first column represents a unique term, the second column represents that term's total number of appearances in the set of documents and the third column represents the number of documents in which that term was found. In the table above, we see the 4 most "popular" unique terms in our index.