Dublin City University
School of Computing
CA4009: Search Technologies
Laboratory Session 3

November 2019

Module Coordinator: Gareth Jones

Laboratory Tutors: Tianbo Ji, Abhishek Kaushik, Yasufumi Moriya, Procheta Sen

1 Introduction

This laboratory extends the study of using the standard information retrieval test collections from Laboratory 2 for the development of search systems. This laboratory focuses on a formal exploration of the impact of varying the k_1 and b parameters of the BM25 probabilistic retrieval model.

You will make use of the TREC data files from Laboratory 2 with the trec_eval programme, and search the indexed TREC document collection using the Lucene search engine via the search interface used in the previous laboratory and a custom written script.

2 Laboratory Reports and Submission

Similar to Laboratories 1 and 2, you should create a electronic report file for this laboratory.

- Include the title and date of the laboratory, and your names at the beginning of your report.
- For each activity described below, enter your answers into your report file, making clear which section your response relates to.
- At the end of the laboratory session you should upload your report to the CA4009 loop page via the link for this laboratory.
- If you do not finish the exercises to your satisfaction, you can complete the assignments in your own time, and submit a revised report via loop. The latest date for submission of the extended report is one week after this laboratory session.

3 Investigating BM25 Parameter Settings

3.1 Background

Your use of the BM25 algorithm for ranking the TREC collection documents using Lucene via the online search tool has, so far, used a single set of values for the parameters k_1 and b. As explained in lectures, varying the values of k_1 and b changes the weights of search terms calculated using the BM25 algorithm.

These changes in term weights produce consequential changes to the matching scores between the query and each document in which the terms have been weighted using BM25.

The changes in matching scores can on some occasions change the relative matching scores of the documents and result in changes in the ranks of the documents retrieved for a query. If you compare the rank of documents which are relevant to the current query for ranked lists created using BM25 with different k_1 and/or b values, you will almost certainly find some of them at higher or lower positions in the ranked lists. The positioning of relevant documents higher or lower in retrieval lists will result in different precision and recall values as calculated by trec_eval, and by implication user satisfaction with the different systems which produced these lists.

The objective of modifying the BM25 weights by varying k_1 and b is to maximise expected retrieval effectiveness. Essentially this means that the parameters are adjusted to increase the average rank of relevant documents with a corresponding reduction in the rank of non-relevant documents[1].

In order to select the best values of k_1 and b, the values of these parameters can be adjusted for a series of retrieval runs using an IR test collection representative of the operational retrieval task. The retrieval behaviour for each run can be evaluated using the trec_eval program that you installed in the previous lab. The values of one or more of the experimental metrics (Precision, MAP, Recall, etc.) are typically compared for different parameter values, with the parameter values which produce the best metric values selected as the operational settings.

3.2 Basic Experimental Procedure

For this part of the laboratory, you will carry out retrieval runs for BM25 ranking using the Lucene search engine (`https://lucene.apache.org`) via the search interface used in the first two laboratories, with the information retrieval test collection you downloaded in the last laboratory consisting of:

- the Lucene index of the TREC documents,
- TREC topics sets,
- relevance information for these topics contained in the qrel file.

This is an initial exercise in carrying out information retrieval runs using a standard test collection and then evaluating the results using trec_eval.

What you need to do:

- For the interactive interface from the previous labs available at: `http://136.206.48.37:8084/IRModelGenerator/`, select the "Search" option and then select "Batch execute TREC queries" .

  This offers you 3 sets of different sets of TREC search topics. The available topic sets are:

  - TREC 6 (query ids 301-350),
  - TREC 7 (query ids 351-400),
  - TREC 8 (query ids 401-450).

[1] Increase the rank here means put the document nearer the top of the list.

These are three separate sets of 50 topics in TREC format, numbered 301-350, 351-400 and 401-450. These are formatted in the TRE format described in the notes for the previous laboratories.

You should see a combo box in the interface where you can select which TREC query set you wish to use. Choose one set of topics for your first set of experiments. You can repeat the procedure for the other topic sets afterwards.

These are the same topics that you downloaded in the laboratory 2, divided into separate sets of 50 topics. Although the topic files are actually stored on the server for this exercise, It would be useful for your understanding of what is going on here, for you to manually inspect a few of the queries again to re-familiarise yourself with the structure of the each topic statement. Note that when executed using the procedure in this part of the laboratory, the server makes use of only the Title field of each TREC format topic statement for the batch execution.

- Select search with the "BM25" ranking option. To operate this, you also need to set values of k_1 and b. For your first run, you can use the standard values used in your initial explorations with BM25 in the previous lab, $k_1 = 1.2$ and $b = 0.75$.

  For further runs, you need to carry out experiments changing these parameters and explore the effect on the MAP (mean average precision) reported by trec_eval for these parameter variations. The details of the BM25 algorithm in the lecture notes on Text Retrieval should help you to determine suitable values of k_1 and b to explore.

- To carry out a run, enter the values of k_1 and b in the respective text boxes and click "Execute Queries". The interface will display a busy spinner while the server executes the search for each of the 50 queries in the batch you selected.

  After this finishes, the server will return a link to the generated results file. You need to download this file by clicking on the link. Save the results file in a local file on your computer. This file is in the same standard TREC results format as the ranked lists results file that you examined in laboratory 2.

- You then need to evaluate the retrieval results in the downloaded file. To do this, run trec_eval as in laboratory 2 with appropriate parameters, i.e. the *qrel* file (qrels.trec678.adhoc) and the retrieval results file that you saved locally.

  This TREC qrel file contains relevance information for all of the topics in the topic sets TREC 6, TREC 7, and TREC 8. trec_eval ignores relevance information for topics in the qrel files which do not appear results file, so you can use the same qrel for whichever set of topics that you are working with.

  The results for the run on the set of topics that you have run should then be displayed on the screen, in the same format as they were for your use of trec_eval in laboratory 2.

Once you have successfully carried out this procedure for a single pair of k_1 and b parameter values, you should carry out further runs using other parameter settings.

In your report give the key results that you obtain using trec_eval, and comment on the change in the values of the metrics in the trec_eval results that you observe when you vary k_1 and b.

3.3 Optimisation of Parameter Values

To set up an information retrieval system for a particular task, developers typically carry out a series of runs with test collection representative of the application task with a systematic variations of the k_1 and b values. They then choose the k_1 and b which give the best results as the system settings for their operational system.

For the next part of this laboratory, you will carry out and report on an investigation of this sort. To do this you will use a prepared script which enables you set multiple values of k_1 and b, and then run the script to carry out search runs for your selected parameter values and pass the output direct to trec_eval, and then to generate output a combined output file containing the MAP value for each search run.

What you need to do:

- Download the script `test.sh` from the laboratory 3 section of the CA4009 loop page.
- Either save this file directly into your folder containing the trec_eval program, or save `test.sh` into the default download folder and copy it to your folder containing trec_eval.
- To setup the script, first open `test.sh` in any editor.

  In `test.sh` you should see the line `for k in`. In this line you should put the values of k_1 which you want to calculate the MAP values for. You need to write these in a space separated manner (e.g. for `k` in `1 2 3`).

  Similarly in the second line `for b in` you can put the values of b which you wish to investigate.

  Once you have entered the values that you wish to investigate save the file `test.sh`.

  Suitable k_1 values are in the range 0.2 to 2, k_1 be assigned other values, but numbers in this range are known from published experimental results to generally be the best.

  b can take values in the range 0.0 to 1.0, you should be able to see why this is the case from the lecture notes.

- You can run the script using the command:

  ```
  sh test.sh <qrel_file_path> <result_file_name> <trec_dataset>
  ```

  where:

  `<qrel_file_path>` is the relevance file used previously in your operation of trec_eval
  `<result_file_name>` is the name you wish to give you to the output file of test.sh
  `<trec_dataset>` is the TREC topic set you wish to you - possible values are 6, 7 or 8

  This will generate a result `result_file_name.csv` file in your current directory.

  Open `result_file_name.csv` result.csv to inspect its content.

  Your should see Ia table of values. The first row of the file shows the b values. The first column shows the k_1 values. The other cells in the tables are the MAP value corresponding to the pairs of b and k_1 values.

- Your objective now is to plot the MAP values on graphs to creates visualisation of the effect of varying k_1 and b on the MAP value for your chosen set of topics.

  To do this load the result.csv file into Microsoft Excel. This will enable you to create line plots showing the relationship between k_1 and MAP and between b and MAP, and also to form a 3D surface plot with k_1, b and MAP as the axes.

  (You can also use any other tool to do line and surface plotting if you prefer.).

You should generate line plots for each set of MAP values for each set of k_1 and MAP values keeping b fixed and for each set of b values keeping k_1 fixed, and a single 3D surface plot showing the MAP values while varying both k_1 and b.

Based on the results in the graph you should be able to recommend the best values of k_1 and b to use for BM25 with your current chosen TREC test collection (6, 7 or 8).

Include your plots in your report, adding comments on what you observe in the lines on the graphs, your recommended values for k_1 and b, with a short explanation of why you chose these values. To do this analysis, you should again refer to the lectures notes describing the roles of k_1 and b in BM25 to consider what your results tell you about the best values of k_1 and b tell you about the current test collection.

- Repeat the above process for the other two sets of topics. Compare the new sets of results to your first plots. Do they show the same patterns or trends in the relationship between k_1 and b? Do they suggest different optimal k_1 and b values? If they do not, how would you choose the k_1 and b values using this document set for a search system based on BM25 when you have these three topics sets available as training data to help you determine the best values for k_1 and b?