INFS7410 Project - Part 3

Preamble

The due date for this assignment is 24 October 2019 at 17:00 Eastern Australia Standard Time.

This part of the project is worth 15% of the overall mark for INFS7410. A detailed marking sheet for this assignment is provided at the end of this document.

Aim

Project aim: The aim of this project is to implement a new information retrieval method based on the knowledge you have acquired during the course, evaluate it and compare it to other methods in the context of a real use-case.

Project Part 3 aim

The aim of part 3 is to:

- Use the evaluation infrastructure setup for part 1.
- Implement a new information retrieval method.
- Evaluate, compare and analyse the developed method against baseline methods.

The Information Retrieval Task: Ranking of studies for Systematic Reviews

Part 3 of the project considers the same problem described in part 1: re-rank a set of documents retrieved for the compilation of a systematic review. A description of the wider task is provided in part 1.

What we provide you with (same as part 1)

We provide:

- for each dataset, a list of topics to be used for *training*. Each topic is organised into a file. Each topic contains a title and a Boolean query.
- for each dataset, a list of topics to be used for *testing*. Each topic is organised into a file. Each topic contains a title and a Boolean query.
- each topic file (both those for training and those for testing), includes a list of retrieved documents in the form of their PMIDs: these are the documents that you have to rank. Take note: you do not need to perform the retrieval from scratch (i.e. execute the query against the whole index); instead you need to rank (order) the provided documents.
- for each dataset, and for each train and test partition, a qrels file, containing relevance

assessments for the documents to be ranked. This is to be used for evaluation.

- for each dataset, and for test partitions, a set of runs from retrieval systems that participated to CLEF 2017/2018 to be considered for fusion.
- a Terrier index of the entire Pubmed collection. This index has been produced using the Terrier stopword list and Porter stemmer.
- a Java Maven project that contains the Terrier dependencies and a skeleton code to give you a start. **NOTE:** Tip #1 provides you with a restructured skeleton code to make the processing of queries more efficient.
- a template for your project report.

What you need to produce

You need to produce:

- correct implementations of the method you are considering
- correct evaluation, analysis and comparison of the developed method, including comparison with the methods implemented in part 1. This should be written up into a report following the provided template.
- a project report that, following the provided template, details: an explanation of the method you have implemented (with your own words), an explanation of the evaluation settings followed, the evaluation of results (as described above), inclusive of analysis, a discussion of the findings.

Required methods to implement

In part 3 of the project you are required to implement a method that you believe will provide results better than the baselines considered in part 1. For the method you implement, consider the re-ranking task, that is, use your method only to re-rank the documents that are provided in the topic file, and do not insert into the ranking any document that was not initially retrieved and provided in the topic file.

We provide two options to tackle this part of the project. You can choose to follow either of the two options.

Option A: your own choice

Implement a method of your choice, including proposing a radically new method. For example you may consider a query expansion technique (be considerate about the amount of expansion though, and the time required to run large queries), a learning to rank approach, or a different retrieval model, among other possible choices. Your method can consider using relevance information in an iterative fashion (e.g., see below for choice B), but only for a document that has been already placed into an iterative ranking mechanism.

Option B: relevance feedback

Implement the following method, making decisions on your own as you see best fit given what you have learned in the course, with respect to the details of the implementation and of the settings. The method is Relevance Feedback, using either of **(1) BM25** (the full formula with RSJ weight), **(2) Binary Independence Model** (where you would use relevance feedback information to set the relevant statistics), **(3) The Rocchio Algorithm**.

To implement the relevance feedback method, do the following: **(1)** start from the BM25 baseline you consider in your experiments; **(2)** the first document BM25 ranks, is the first document your system ranks; **(3)** for every subsequent rank i+1, to decide which document should be placed, consider the documents at rank 1 to i and acquire their true relevance labels from the qrels; then compute the scores of the documents you have not ranked yet according to the Relevance Feedback method you have chosen; **(4)** place at rank i the document with the highest score identified at point 3; **(5)** continue with this process until all documents have been re-ranked.

Note that once a document has been placed at rank i, you should not change its position depending on its relevance label.

In other words, the process describe above consists of presenting the first document from your baseline to an imaginary user, then gather its relevance assessment for that document. Then, identify the next document to present to the user, by considering the feedback given so far. Once the next document is displayed, feedback is again gathered and used, along with all feedback provided up until now, to identify the next document to show to the user. This proceeds in an iterative way.

General considerations

When tuning any method, tune with respect to MAP using the training data portion only.

We strongly recommend you use and extend the Maven project provided for part 1 to implement your method.

In the report, detail how the method was implemented, including which formula you implemented.

What queries to use

For part 3, we ask you to consider the queries for each topic created from the title field of each topic. For example, consider the example (partial) topic listed below: the query will be `Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries` (you may consider performing text processing). This is the same query type used in part 1.

```
Topic: CD008122

Title: Rapid diagnostic tests for diagnosing uncomplicated P. falciparum
malaria in endemic countries

Query:
1. Exp Malaria/
2. Exp Plasmodium/
3. Malaria.ti,ab
4. 1or2or3
5. Exp Reagent kits, diagnostic/ 6. rapid diagnos* test*.ti,ab
7. RDT.ti,ab
8. Dipstick*.ti,ab
```

Above: example topic file

Required evaluation to perform

In part 3 of the project you are required to perform the following evaluation:

1. For all methods, train on the training set for the 2017 topics with respect to any parameter you may consider in your method (if you have multiple paramters, do not tune for more than 2 parameters) and test on the testing set for the 2017 topics (using the parameter value you selected from the training set). Report the results of every method on the training (the best selected) and on the testing set, separately, into one table. Perform statistical significance analysis across the results of the methods.
2. Comment on the results reported in the previous table by comparing the methods on the 2017 dataset.
3. For all methods, train on the training set for the 2018 topics with respect to any parameter you may consider in your method (if you have multiple paramters, do not tune for more than 2 parameters) and test on the testing set for the 2018 topics (using the parameter value you selected from the training set). Report the results of every method on the training (the best selected) and on the testing set, separately, into one table. Perform statistical significance analysis across the results of the methods.
4. Comment on the results reported in the previous table by comparing the methods on the 2018 dataset.
5. Perform a topic-by-topic gains/losses analysis for both 2017 and 2018 results on the testing datasets, by considering as baseline BM25.
6. Comment on trends and differences observed when comparing the findings from 2017 and 2018 results.

In terms of evaluation measures, evaluate the retrieval methods with respect to mean average precision (MAP) using `trec_eval`. Remember to set the cut-off value (`-M`, i.e., the maximum number of documents per topic to use in evaluation) to the number of documents to be re-ranked for each of the queries. Using `trec_eval`, also compute Rprecision (Rprec), which is the precision after R documents have been retrieved (by default, R is the total number of relevant docs for the topic).

For all statistical significance analysis, use paired t-test; distinguish between p<0.05 and p<0.01.

Baseline

If you have submitted to Part 1, please use the BM25 baseline that you have produced when submitting. If you have not submitted to Part 1, we have provided for you a BM25 baseline run you can use for comparison, and as base of your method (if you are following choice B). This is made available in Blackboard.

How to submit

You will have to submit 3 files:

1. the report, formatted according to the provided template, saved as PDF or MS Word document.
2. a zip file containing a folder called `runs-part3`, which itself contains the runs (result files) you have created for the implemented methods.
3. a zip file containing a folder called `code-part3`, which itself contains all the code to re-run your experiments. You do not need to include in this zip file the runs we have given to you. You may need to include additional files e.g. if you manually process the topic files into an intermediate format (rather than automatically process them from the files we provide you), so that we can re-run your experiments to confirm your results and implementation.

If your set of runs is too big, please do the following:

- Include in the zip the test run.
- Include in the zip the best train run you used to decide upon the parameter tuning.
- Create a separate zip file with all the runs; upload it to a file sharing service like dropbox or google drive (or similar), then make sure it is visible without login and add the link to it to your report. Please ensure that the link to the resources is available for at least 6 days after the submission of the assignment.

All items need to be submitted via the relevant Turnitin link in the INFS7410 Blackboard site, by 24 October 2019 at 17:00 Eastern Australia Standard Time, unless you have been given an extension (according to UQ policy), *before* the due date of the assignment. Note: appropriate, separate links are provided in the Assignment 3 folder in Blackboard.

INFS 7410 Project Part 3 – Marking Sheet

Criterion	**%**	**7** **100%**	**4** **50%**	**FAIL 1** **0%**
IMPLEMENTATION The ability to: • Understand, implement and execute IR methods	7	• Correctly implement the method specified in the report • If choosing option A, well motivated intuition of why the implemented method may work well for the task • If choosing option B, make sensible choices for parameter settings and implementation choice	• Correctly implement the method specified in the report • If choosing option A, the intuition of why the implemented method may work well for the task is not provided, or is weak • If choosing option B, choices for parameter settings and implementation choice are not justified/not sensible.	• No implementation, or largely incorrect implementation of the method described in the report
EVALUATION The ability to: • Empirically evaluate and compare IR methods • Analyse the results of empirical IR evaluation • Analyse the statistical significance difference between IR methods' effectiveness	6	• Correct empirical evaluation has been performed • Uses all required evaluation measures • Correct handling of the tuning regime (train/test) • Reports all results for the provided query sets into appropriate tables • Provides graphical analysis of results on a query-by-query basis using appropriate gain-loss plots • Provides correct statistical significance analysis within the result table; and correctly describes the statistical analysis performed • Provides a written understanding and discussion of the results with respect to the method • Provides examples of where the implemented method works, and where it does not, and why, e.g., discussion with respect to queries, runs.	• Correct empirical evaluation has been performed • Uses all required evaluation measures • Correct handling of the tuning regime (train/test) • Reports all results for the provided query sets into appropriate tables • Provides graphical analysis of results on a query-by-query basis using appropriate gain-loss plots • Does not perform statistical significance analysis, or errors are present in the analysis	• No or only partial empirical evaluation has been conducted, e.g. only on a topic set, or a subset of topics • Only report a partial set of evaluation measures • Fails to correctly handle training and testing partitions, e.g. train on test, reports only overall results
WRITE UP *Binary score: 0/2* The ability to: • use fluent language with correct grammar, spelling and punctuation • use appropriate paragraph, sentence structure • use appropriate style and tone of writing • produce a professionally presented document, according to the provided template	2	• Structure of the document is appropriate and meets expectations • Clarity promoted by consistent use of standard grammar, spelling and punctuation • Sentences are coherent • Paragraph structure effectively developed • Fluent, professional style and tone of writing. • No proof reading errors • Polished professional appearance		• Written expression and presentation are incoherent, with little or no structure, well below required standard • Structure of the document is not appropriate and does not meet expectations • Meaning unclear as grammar and/or spelling contain frequent errors. • Disorganised or incoherent writing.
BONUS *Binary score: 0/2* The ability to: • Identify highly performing IR solutions and settings	2	• The implemented method outperforms the BM25 baseline		• The implemented method does not outperform the BM25 baseline