# IN3066/INM305: Information Retrieval

The deadline for this assignment is **18th May 2025 @ 5.00pm**

## 1. Introduction

This assignment requires you to demonstrate your knowledge of IR software by identifying a problem, building an application to solve that problem and undertaking an evaluation of that software using appropriate technologies. In this document we specify the coursework tasks you need to undertake (section 2), the coursework deliverables (section 3), the submission process and requirements (section 4) and the marking/grading criteria (section 5). Please read this document carefully and make sure you follow the instructions given.

## 2. Coursework tasks

You need to undertake three main tasks (recall any work done on a project proposal).

### 2.1 Background and motivation

This element of the assignment is assessed via the project report.

In your report, you must provide a problem description (e.g. commercial or research) and explain the domain in which your problem sits (e.g. web search, health search etc). You need to explain who the beneficiaries are i.e. your users, who undertake searches in your chosen domain. You need to provide a literature review giving background on the work done in terms of the problem you have identified, any algorithms utilized and any articles which provide insight into your domain etc. You need to provide an overview of your test collection, indicating where you got it from and what it consists of e.g. text, information needs/topics and relevance assessments (if topics and relevance assessment are absent, you will need to create them yourself). Please ensure that you provide citations to all material you reference. Please note, that you are NOT allowed to use generative AI tools to produce your collection.

### 2.2 IR application: development process and documentation

This element of the assignment is assessed via the project report and source code.

You are required to build an IR application on SOLR using an appropriate programming language to address the problem you have identified. Python is preferred, but there are different ways to make connections with SOLR and depending on the problem you address e.g. create a web application using JavaScript. Please speak to the module leader for any advice on approaches to take.

You need to document the decisions you made when building your IR application in the report, making reference to your source code. You should structure this part of your report as follows:

* Indexing: You must provide details on how you indexed your test collection, justifying your choices. Explain your schema design and how you treated the text in the test collection. How did you create the schema in SOLR? For the text, did you use JSON or XML format and use the SOLR indexer to create the index, or did you use libraries such as NLTK to parse html text to create JSON/XML describing choices such as stop work removal or stemming? Please explain the challenges you had when doing this, and what you did to overcome those issues.
* Search: How do your search algorithms work? What search models did you use and what libraries did you draw on to build your search? How did you process queries? How did you handle and present or record search results? Please explain your choices and how the search algorithms are designed and built to meet the needs of users in your domain. If you are comparing and contrasting search models and algorithms please make sure that you list them all. Make reference to any code or libraries that you used to build your search application e.g. machine learning libraries available as part of python.
* All source code should be appropriately commented, and laid out using appropriate software engineering standards. Code must be syntactic and grammatically correct and appropriately organised e.g. re-using code in functions as and when necessary.

In your report you must make reference to the source code when describing and reflecting on the technical work that you undertook. You must make a clear link between the problem you have identified and the technical work you have carried out. If necessary, make reference to the appropriate literature.

### 2.3 Evaluation of IR application

This element of the assignment is assessed via the project report.

What did you do to the test collection to make it complete e.g. did you produce topics and relevance assessments that relate to the text to complete the test collection? If you produced topics, explain how you came up with the ideas or the sources you used to inform your ideas. If you produced relevance assessments, explain how you produced these from the topics reflecting on the type of relevance and the criteria you used to judge relevance (see section 5 of the session 7 lecture materials). Note that you can re-use topics and/or relevance assessments produced by conferences such as TREC as and when appropriate – please approach the module leader for advice on these issues.

You need to explain and justify your evaluation method. For example if you are undertaking batch based experiments you will need to explain the laboratory style of evaluation and the process you used. Alternatively, if you have built a user interface, you will like focus more on a operational evaluation – did you evaluate the results yourself or did you recruit users? If you do recruit users please talk to me beforehand, as you’ll need to use ethics materials such as information sheets and consent forms.

Explain the results of your evaluation – which must focus on relevance. What metrics did you use e.g. precision, recall etc? Don’t explain standard methods, just provide a citation and reference to where you obtained that information. If you undertook an operational evaluation, did you use an information seeking model to evaluate the searches and if so, how? For example, if you recruited users what did they say about the relevance of the output presented to them on the user interface. Please reflect on the work you did when building the application in the light of the evaluation results. State the limitations of your work, and what you think you would do differently if you were to undertake the project again. If appropriate, please provide ideas for future work.

### 2.4 Final note

References for all parts of your work are required e.g. citations in the report, source code from external sources, software libraries etc.

### 3. Coursework Deliverables

### 3.1 Report

You must submit a report describing the tasks you carried out (see section 2) to the appropriate link. You must split your sections into: *1. Background and motivation, 2. IR application: development process and technical documentation, 3. Evaluation of IR application. 4. References.* Please do not use any other structure for your submission.

Requirements for UG and PG students are different:

* UG students – IN3066: Your report must not exceed THREE A4 pages, excluding references, using Arial 11 font or similar (single-spaced).
* PG students – INM305: Your report must not exceed SIX A4 pages, excluding references, using Arial 11 font or similar (single-spaced).

For all submissions, margins should be at least 2.5cm on all sides. Take care not to exceed this page limit, as additional pages will not be marked. Do not include a table of contents, cover page, or cover sheet. Do however include your name, student no and module code in the document header (i.e. Author, Student No, IN3066 or INM305). Acceptable formats for submission are word documents (.doc, .docx) or PDF.

### 3.2 Source code

You will need to submit your source code in both text form and its original executable form i.e. a text file with your source code and a zip file with your product package containing your IR application (see section 4 below for details on how and what to submit).

In terms of the text file for your source code, you must provide all the code that you have produced to build your application e.g. curl commands to build a schema, python programmes to index and search your test collection. Please provide a clear organisation for your source code file, providing page numbers and sections so that you can make clear reference to appropriate parts of your code when describing and evaluating code elements in your report.

Please note that you will not be assessed on how well your searches do when evaluating your IR application. Information needs will have varying levels of difficulty and therefore the retrieval effectiveness will vary between systems. The assessment will be used to measure your understanding of IR application development and evaluation methods in information retrieval.

## 4. Submission

Please refer any queries about the coursework to Andrew MacFarlane. Email address is: andym@city.ac.uk. The deadline for this assignment is **18st May 2025 @ 5.00pm**, though the relevant Moodle submission area described below.

### 4.1 Main coursework submission area

Your submission must include these files (NB zip files ARE NOT allowed):

* A single .pdf or word file containing your coursework report.
* A single text file with all the source code written by you (not the source code of libraries that you have used, re-used code, or code that was generated automatically). NB: this file is just for various checks, including checks for plagiarism; it does not substitute for your submission of the product package, which is submitted separately. This should be between 10-20 pages of text.
* A README file, which informs the marker of the contents that you have submitted and where to find them, and how to install and test your package.
  + You should include details of the folder structure and all files submitted in the zip file submitted to the product package submission area.
  + For web-based products, provide both the means to install the product and a URL for a hosted version that markers can access (where possible).

Note – do not submit your product package submission to this link – see below.

### 4.2 Product Package submission area

Please create a zip archive of your IR application, and submit it to the product package submission area. Your Product Package submission must include files with all code/images/data/etc needed for your "product" to run, using the required folder structure for it to run. Note that you do not need to submit the text for the test collection, you can just provide a link to the source in the report. However, if you did produce topics and relevance assessments, then you should submit them in this archive. In the case of laboratory evaluation approaches, there should be sufficient information for the marker to re-run the evaluations.

## 5. Assessment and Marking scheme

The final marks for the assessment are allocated using the following criteria:

* Background and motivation (5 marks)
* IR application: Development process and source code (70 marks)
* Evaluation of the IR application (25 marks)

General guidelines for what is expected in your assessed work are as follows:

### 5.1 Students undertaking IN3066

|  |  |  |
| --- | --- | --- |
| **Class** | **%** | **Description** |
| 1st | 70+ | Strong or comprehensive ability to identify a clear IR problem to address in a given domain for appropriate beneficiaries is demonstrated. Choice of test collection is fully justified for the problem at hand. Relevant literature is identified and strong or comprehensive synthesis of ideas is demonstrated. Strong or comprehensive knowledge of the development of IR applications to solve the identified problem is demonstrated. Evaluation of results demonstrates strong or comprehensive knowledge of appropriate evaluation methodologies. Reflection of relevance assessment in evaluation demonstrates strong or comprehensive knowledge of the difficulties in satisfying information needs (resolving ASKs). The learning outcomes will be realised in full. |
| 2:1 | 60-69% | Some critical ability to identify a clear IR problem to address in a given domain for appropriate beneficiaries is demonstrated. Choice of test collection is justified for the problem at hand. Relevant literature is identified and sound synthesis of ideas is demonstrated. Sound knowledge of the development of IR applications to solve the identified problem is demonstrated. Evaluation of results demonstrates sound knowledge of appropriate evaluation methodologies. Reflection of relevance assessment in evaluation demonstrates sound knowledge of the difficulties in satisfying information needs (resolving ASKs). Understanding of several of the learning outcomes (if not necessarily all) is of a high standard. |
| 2:2 | 50-59% | An attempt is made at to identify a clear IR problem to address in a given domain for appropriate beneficiaries. Choice of test collection is partly justified for the problem at hand. Relevant literature is identified and some synthesis of ideas is demonstrated. Adequate knowledge of the development of IR applications to solve the identified problem is demonstrated. Evaluation of results demonstrates an adequate knowledge of appropriate evaluation methodologies. Reflection of relevance assessment in evaluation demonstrates an adequate knowledge of the difficulties in satisfying information needs (resolving ASKs). Understanding of several of the learning outcomes is demonstrated, but does not provide a full answer (some important material is missing) and/or provides some information that is incorrect/inaccurate. |
| 3rd | 40-49% | Limited attempt is made in identifying a clear IR problem to address in a given domain for appropriate beneficiaries. Choice of test collection is justified for the problem at hand, but is limited. Some relevant literature is identified and limited synthesis of ideas is demonstrated. Limited knowledge of the development of IR applications to solve the identified problem is demonstrated. Evaluation of results demonstrates limited knowledge of appropriate evaluation methodologies. Reflection of relevance assessment in evaluation demonstrates a limited knowledge of the difficulties in satisfying information needs (resolving ASKs). Learning outcomes will be realised in part, but there are substantial limitations to the work. |
| Fail | 39% or lower | No attempt is made to identify a clear IR problem to address in a given domain for appropriate beneficiaries. Choice of test collection is not justified for the problem at hand, or is very limited. No relevant literature is identified and no synthesis of ideas is demonstrated. No knowledge of the development of IR applications to solve the identified problem is demonstrated. Evaluation of results demonstrates no knowledge of appropriate evaluation methodologies. Reflection of relevance assessment in evaluation demonstrates no knowledge of the difficulties in satisfying information needs (resolving ASKs). Learning outcomes will not be realised. |

### 5.2 Students undertaking INM305

|  |  |  |
| --- | --- | --- |
| **Class** | **%** | **Description** |
| Distinction | 70+ | Strong or comprehensive ability to identify a clear IR problem to address in a given domain for appropriate beneficiaries is demonstrated. Choice of test collection is fully justified for the problem at hand. Relevant literature is identified and strong or comprehensive synthesis of ideas is demonstrated. Strong or comprehensive knowledge of the development of IR applications to solve the identified problem is demonstrated. Evaluation of results demonstrates strong or comprehensive knowledge of appropriate evaluation methodologies. Reflection of relevance assessment in evaluation demonstrates strong or comprehensive knowledge of the difficulties in satisfying information needs (resolving ASKs). The learning outcomes will be realised in full. |
| Merit | 60-69% | Some critical ability to identify a clear IR problem to address in a given domain for appropriate beneficiaries is demonstrated. Choice of test collection is justified for the problem at hand. Relevant literature is identified and sound synthesis of ideas is demonstrated. Sound knowledge of the development of IR applications to solve the identified problem is demonstrated. Evaluation of results demonstrates sound knowledge of appropriate evaluation methodologies. Reflection of relevance assessment in evaluation demonstrates sound knowledge of the difficulties in satisfying information needs (resolving ASKs). Understanding of several of the learning outcomes (if not necessarily all) is of a high standard. |
| Pass | 50-59% | An attempt is made at to identify a clear IR problem to address in a given domain for appropriate beneficiaries. Choice of test collection is partly justified for the problem at hand. Relevant literature is identified and some synthesis of ideas is demonstrated. Adequate knowledge of the development of IR applications to solve the identified problem is demonstrated.  Evaluation of results demonstrates an adequate knowledge of appropriate evaluation methodologies. Reflection of relevance assessment in evaluation demonstrates an adequate knowledge of the difficulties in satisfying information needs (resolving ASKs). Understanding of several of the learning outcomes is demonstrated, but does not provide a full answer (some important material is missing) and/or provides some information that is incorrect/inaccurate. |
| Fail | 49% or lower | Limited or no attempt is made to identify a clear IR problem to address in a given domain for appropriate beneficiaries. Choice of test collection is justified for the problem at hand, but is limited. Some relevant literature is identified and limited or no synthesis of ideas is demonstrated. Limited or no knowledge of the development of IR applications to solve the identified problem is demonstrated. Evaluation of results demonstrates limited or no knowledge of appropriate evaluation methodologies. Reflection of relevance assessment in evaluation demonstrates limited or no knowledge of the difficulties in satisfying information needs (resolving ASKs). Learning outcomes will not be realised. |

**Submission Process :**

**The following information on coursework submission re-emphasises the information in your programme handbook.**

* All submissions are by Moodle. No other form of submission will be accepted.
* Please note that you are **not** required to submit a coversheet when submitting by Moodle. Clicking the Submit button on the Assignment Submission screen indicates that you have read and agreed to the rules of submission set down by the university.
* Once the deadline has passed coursework cannot be changed, nor can additional materials be submitted.
* Text beyond any specified word limit will not be marked.
* Use of Generative AI tools on the module is forbidden.
* Plagiarism will not be tolerated under any circumstances and where found will lead to a formal investigation of your work and reference to the Academic Misconduct Panel. This might result in penalties ranging from mark deduction to withdrawal from the University. See your programme handbook for details on the nature of plagiarism and the department's policy.
* **IT IS ENTIRELY YOUR RESPONSIBILITY TO ENSURE THAT YOUR WORK IS SUBMITTED FULLY, CORRECTLY AND ON TIME.**
* It is therefore strongly recommended that you set yourself a 'hard' personal deadline for submission well in advance of the Moodle closing date.

**©City University**  
22nd January 2025

*This page does not appear to be using a stylesheet. It will be much improved if you use a recent browser that supports cascading style sheets.*