# IN3066/INM305: Information Retrieval Coursework Proposal

The deadline for this summative assignment is **7th  March 2025 @ 5.00pm**

## 1. Introduction

Before starting your assignment you will have an opportunity to submit a proposal with the ideas that you have for your coursework for formative feedback. Whilst this is not mandatory, we strongly recommend that you write and submit the proposal so that you can get some advice on how to proceed. Details of what to submit are given below.

## 2. Problem and domain

You will need to identify a problem to address, which can either be commercial (e.g. a digital library application) or research (e.g. compare and contrast different retrieval algorithms). You must identify the domain in which the problem exists e.g. web search, enterprise search, health search, legal search etc. Explain why this is important and worth pursuing, and how significant the problem you have identified is. What is the end goal of your work? You can address either a technical issue or one focused on users. In either case, you need to identify beneficiaries i.e. your target users such as the general population (as in web search) or professional search (as in legal or systematic review search). Identify literature that provides the necessary background to the work you undertake. This requires you to undertake searches on systems such as Google scholar or computing focused digital libraries such as the ACM portal, where the relevant articles can be found. Provide citations to the literature you have found, and put the references into a references section in Harvard format.

## 3. Test Collection

Given the problem you have identified, the domain of the problem and the beneficiaries (i.e. your users), you will need to choose a test collection. Examples of sources for test collections are given in the lecture materials for session 3, section 2.1. Please review this material thoroughly, and ensure that you choose a collection that will meet your project needs, that is appropriate for your domain and of interest to your users in terms of resolving their information needs.

Please provide an overview of the collection that you have chosen, and explain why it is useful for the problem you have identified. Explain the structure of the collection and what it consists of. Does the text have fields and if so, is it using a meta-data standard? Do you have any topics for your test collection and relevance assessments for that text. You will have this for TREC collections, but not for other types of collections in most cases. Please think about how to source your topics in the case of the later (perhaps TREC topics) or how you will generate them yourself. If relevance assessments are not available, then you will need to generate them – we will cover this in the second part of the module.

You must provide citations and references to the sources for your test collection i.e. where the text, topics and relevance assessments can be obtained.

## 4. Methodology and tools

You will be building an IR application to solve the problem that you have identified, and you’ll need to think about the process, methods and tools that you will use in order to complete your assignment.

You will need to make a number of high level decisions before thinking about the methods and tools that you will use in order to develop your IR application:

* What type of application is appropriate for the problem you have identified? Is it experimental for a research problem or user focused? Will you build a user interface using something like JavaScript or will the application be batch based and run on the command line?
* You will need to undertake a literature review to identify the methods and tools you will use to build your application. What other methods have been used in the domain (if any) and what was the result (if you are doing a research project you may be able to identify a gap to address)?
* You need to be clear about the scope of your work, given the limited timeframe of around two months you will need to complete the project. Don’t try and solve a big problem, but undertake something that can be complete within the timeframe without it being trivial leading to poor marks.

When you have clear answers to these high level decisions you can consider detailed methods and tools to build your application:

* Provide a plan of work for implementing your system on SOLR such as:
  + How will you treat the collection? What do you need to do in order to index the collection in SOLR such as creating a schema, eliminating stop words etc.
  + How will you search the collection, given the information needs appropriate for the domain? What search models will you use, and will you need to do any pre-processing for these models e.g. learning to rank or language models?
  + Given the problem you have identified, what is the desirable end result? This will give you an indication of what to do when evaluating your software, given the methods presented in the 2nd half of the module.
  + What are the timelines for each element of your work? Break your work up into sub-tasks and provide an idea of how long each should take.
* What libraries and tools will you use above and over ones already provided by SOLR? For examples libraries such as NLTK for processing text or some of the machine learning libraries that are available on python.
* If you are doing a research project, what baseline will you use to test your ideas against?
* What programming languages and/or frameworks will you use? The module is taught using python, but you may wish to use web based languages such as JavaScript to build web interfaces.

## 5. References

Please provide all the references for the citations to literature and sources that you have made in the proposal using Harvard format. Your references section should not be empty. You must provide references to all the material that you used to identify the problem, select the test collection and the methods and tools you used.

## Formative Assessment Submission

You should split your proposal into: *1. Project title, 2. Problem and domain, 3. Test collection, 4. Methodology and tools*, *5. References.*

Your proposal must not exceed TWO A4 pages, excluding references, using Arial 11 font or similar (single-spaced). Margins should be at least 2.5cm on all sides. Take care not to exceed this page limit, as you will not obtain feedback on additional pages. Do not include a table of contents, cover page, or cover sheet. Acceptable formats for submission are word documents (.doc, .docx) or PDF. Please refer any queries about the proposal to Andrew MacFarlane. Email address is: andym@city.ac.uk. The deadline for this assignment is **7th March 2025 @ 5.00pm**, through the relevant Moodle submission.

**Submission Process :**

**The following information on proposal 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 declaration on the left of the submission screen. This takes the place of the coversheet previously used for paper-based submissions.
* Once the deadline has passed coursework cannot be changed, nor can additional materials be submitted.
* Text beyond any specified word limit will not obtain feedback.
* 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.*