# CE306 - Information Retrieval (Resit)

Alba García Seco de Herrera

June 2021

## Plagiarism

*You are reminded that this work is for credit towards the composite mark in CE306, and that the work you submit must therefore be your own. Any material you make use of, whether it be from textbooks, the Web or any other source must be acknowledged as a comment in the program, and the extent of the reference clearly indicated.*

## Task 1

The idea of this task is that you apply the information retrieval knowledge you acquired during this term and put it into practice. You are already familiar with Elasticsearch. You also know the processing steps that turn documents into a structured index, commonly applied retrieval models and you know the key evaluation approaches that are being employed in IR. Now is a good time to put it all together.

This task comes in stages. Marks are given for each stage. The stages are as follows:

- **Obtain a dataset (10%)** The first step for you will be to obtain the *Signal Media One Million News Article* [1] dataset to which you will apply your information retrieval knowledge. The Signal Media One Million News Articles Dataset1 is a collection of news articles from a variety of sources that has been made available to the research community. Provide a detailed description of the dataset as well as the information on where to find and download it.
- **Indexing (10%)** Once you have obtained the dataset, upload it with full text to Elasticsearch. If you run into problems using the upload script provided, then feel free to use your own approach. You might also want to start loading a small sample of documents first before using the full collection.
- **Searching (10%)** Once you have indexed the collection you want to be able to search it. You can do that on the command line but it would be much better to have an interactive system. You could start with Kibana for that but you are free to use other open source tools for your GUI.
- **Building a Test Collection (10%)** Imagine you would like to explore what search engine settings are most suitable for the collection you are indexing to make search as effective as possible. To start with this you should devise a small test collection that contains a number of queries together with their expected results. Identify *three specific events covered* by the collection and then compose *two sample queries for each* of these that you might reasonably expect a user to submit to find documents about this event.
- **Evaluation (30%)** Once you have a test collection you can explore different search engine settings to see what effect they have on the evaluation results. To do that you need to identify a suitable metric. Use P@5 as the metric of choice for this assignment. You can then vary different parameters. You could for example change the pre-processing pipeline by comparing a system that uses stemming with one that does not. However, this will require you to re-index the collection. Instead I suggest you try different retrieval models such as Boolean versus TF.IDF.
- **Engineering a Complete System (10%)** The final system should have control over all the individual components so that as the final result we have a complete search engine.

[1] https://research.signal-ai.com/newsir16/signal-dataset.html

You will have noticed that the percentages above only add up to 80%. This is because one of the important aspects of the project is that your work should be well documented and your code well commented. **20% of your mark will come from this.** The report should contain:

- Instructions for running your system
- Screenshots illustrating the functionality you have implemented
- Design and design decisions of your overall architecture
- A description of the document collection you have chosen
- The actual ground truth data that make up your test collection (i.e. queries with their matching documents)
- A short description and motivation of your evaluation methodology
- Evaluation results
- Discussion of your solution focussing on functionality implemented and possible improvements and extensions.

The report does not need to be long as long as it addresses all the above points.

You should submit:

- Report
- Code
- Powerpoint slides

## Software

The backend search engine to be used is *Elasticsearch*. Apart from that you are free to write code in any language of your choice and employ any open source tool that you find suitable.

## Task 2

Once you have completed and met the requirements in Tasks 1, produce four PowerPoint slides (or equivalent) to evidence the functionality of your work.

You should imagine these slides being used in a 5 minute presentation to an employer. They should emphasise proof of operation showing screengrabs of operational code and any problems encountered.

## Submission

You should submit:

- Report
- Code
- Powerpoint slides

The submission of all two completed tasks should be submitted as a single *zip file* via the electronic submission system. Please check the details of the submission deadline with the CSEE School Office.

*The guidelines about late assignments are explained in the students' handbook.*