CE306/ CE706 - Information Retrieval
Assignment 2: Elasticsearch & Evaluation

Alba García Seco de Herrera

8th March 2020

Plagiarism

You are reminded that this work is for credit towards the composite mark in CE306/CE706, 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.

The Context of your Task

Imagine you have just finished university and started a job with an organisation that is in desperate need of a new search engine that allows employees to search the document collection at hand. This is your chance to shine!

The Task

The idea of this assignment 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.

Crowdsourcing mechanisms are now routinely being used for labelling data for information retrieval and other computational intelligence applications. As part of your assignment, you will participate in a crowdsourcing task aiming to predict media memorability when watching short videos. You will need to participate in two task in the labs during Week 25. The first task will last 25 minutes and the second task will last 20 minutes. Notice that the second task will need to be complete within 24 to 72 hours after completing the first task. The following labs will be make available for this aim:

- Monday 16th March, 9 am to 11 am, CES lab 1;
- Tuesday 17th March, 11 am to 1 pm, CES lab 1;
- Wednesday 18th March, 9 am to 11 am, CES lab 1;
- Thursday 19th March, 1 pm to 4 pm, CES lab 4;
- Friday 20th March, 11 am to 2 pm, CES lab 1.

This assignment comes in stages. Marks are given for each stage. You may choose not to attempt some stages. The stages are as follows:

- **Indexing (10%)** The *Signal Media One Million News Articles Dataset*[1] is a collection of news articles from a variety of sources that has been made available to the research community. The first step for you

[1] `http://research.signalmedia.co/newsir16/signal-dataset.html`

will be to obtain the dataset and load it into 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. Note that the collection is provided in JSON format and each article contains different fields. Make sure that a user can decide which field to search (note that one of the fields is the publication date of the article).
- **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 ten specific events covered by the collection and then compose some sample queries that you might reasonably expect a user to submit to find documents about this event.
- **Evaluation (20%)** 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 (MAP, for example). 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. Alternatively, you might want to 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.
- **Crowdsourcing (20%)** This mark will only be consider you complete both crowdsourcing tasks. More details on the tasks will be provided in the labs on the designated dates.

You will have noticed that the percentages above only add up to 80%. The remaining 20% will come from a report that describes your work. 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.
- Short description of your crowdsourcing experience e.g. lessons learned, how to improve the user experience, etc.

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

You may work in pairs. Both members of a pair will get the same mark unless there is reason to do otherwise except for the 20% of marks assigned for crowdsourcing which will be individually assigned based only on the completition of both tasks. If you do work in a pair, then please make sure that you both submit an assignment and that this will be identical for both of you.

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.

Submission

The assignment, which counts for **20%** of the overall mark, should be submitted as a single *zip file* via the electronic submission system by **Friday, 9th April 2020, 11:59 (mid-day)**. *The guidelines about late assignments are explained in the students' handbook.*