**CE706 - Information Retrieval**

Alba García Seco de Herrera

May 2023

# Plagiarism

*You are reminded that this work is for credit towards the composite mark in 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

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.

**Your task**

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

* **Indexing (20%)** The first step for you will be to obtain the *Shakespeare* dataset (<https://www.elastic.co/guide/en/kibana/5.5/tutorial-load-dataset.html>) that you find in the Elasticsearch tutorial to which you will apply your information retrieval knowledge. The *Shakespeare* is mainly text but also has some internal structure { a perfect use case for Elasticsearch. Once you have obtained the dataset, upload it with full text to Elasticsearch. If you run into problems using the upload script provided in the labs, 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.
* ﻿ **Tokenization and Case folding (10%)** The next step should be to transform the input text into a normal form of your choice. For this task you are required to use Elasticsearch’s build-in analyzers or other libraries (as learned in Lab 2) to tokenize the document and perform case folding to the tokens.
* ﻿**Selecting Keywords (20%)** One aim of your system is to identify the words *and* phrases in the text that are most useful for indexing purposes. For this task you are required remove words which are not “useful”. E.g. very frequent words or stopwords. Apply *tf.idf* as part of your selection and weighting step. (*Hint*: the stopword removal, n-gram extraction can be done with Elasticsearch’s build-in tokenizer and tf.idf scores can also be configured using Elasticsearch similarity module.)
* **Stemming or Morphological Analysis (10%)** Writing word stems to the database rather than words allows to treat various inflected forms of a word in the same way, e.g. *bus* and *busses* refer to exactly the same thing even though they are different words.
* **Searching (10%)** Once you have indexed the collection you want to be able to search it. You can do that on the command line (like in Lab 1), but it would be easier to do it in an an interactive system such as Kibana’s dev tool. The task is to create 3 textural queries that the user might come up and write the corresponding Elasticsearch queries.
* **Working with Elasticsearch API (10%)** Finally the 10% will be given if you could make everything work with the Elasticsearch API.

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/justifications of your overall architecture
* A description of the document collection you have chosen
* 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.

# Software

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

# Submission

You should submit:

* Report (**use the template below**)
* Code

The completed task should be submitted 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.*

**CE706 - Information Retrieval SU 2023**

**Assigment 1**

Student ID

# Instructions for running your system

*Include here instructions to run your system and control each individual component. This could be as simple as start Elasticsearch and Kibana if you are not using Elasticsearch API. You may include screenshots to clarify.*

# Indexing

*Include here the details of how you download your datasert and index it including any issue that you had and how did you face it. Explain which documents have you selected for your experiments.. You may include screenshots to clarify.*

﻿Tokenization and Normalisation

*Include here the details of how you did this step including any issue that you had and how did you face it.* ***Present examples to show how your system works****, e.g., if you use elastic analyzers, you can show how the analyzer works by given sample text input (remember we did this in Lab 2).* ***Present examples for each of the aspects where this step went well. Also include examples for when it when wrong*** *and how you could solve it. You may include screenshots to clarify.*

﻿Selecting Keywords

*Include here the details of how you did this step including any issue that you had and how did you face it****. Present examples for each of the aspects*** *where this step went well by given sample text input (remember we did this in Lab 2). Also include examples for when it when wrong and how you could solve it.You may include screenshots to clarify.*

# Stemming or Morphological Analysis

*Include here the details of how you did this step including any issue that you had and how did you face it.* ***Present examples for each of the aspects where this step went well*** *by given sample text input (remember we did this in Lab 2).* ***Also include examples for when it when wrong*** *and how you could solve it. You may include screenshots to clarify.*

# Searching

*Include here the details of how you did this step including any issue that you had and how did you face it. You may include screenshots to clarify.*