**CE706 - Information Retrieval**

Alba García Seco de Herrera

February 2021

# 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

Researchers, clinicians, and policy makers involved with the response to COVID-19 are constantly searching for reliable information on the virus and its impact. This presents a unique opportunity for the information retrieval (IR) and text processing communities to contribute to the response to this pandemic, as well as to study methods for quickly standing up information systems for similar future events.

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.

**Scenario**: In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the *COVID-19 Open Research Dataset (CORD-19)*[[1]](#footnote-1). CORD-19 is a resource of over 181,000 scholarly articles, including over 80,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in information retreival and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

**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 dataset. Once you have done so upload a **sample of 1000** articles with full text to Elasticsearch (the simplest thing is to use the first 1000 documents). You will work with the **metada.csv** file provided by the challenge.
* ﻿**Sentence Splitting, Tokenization and Normalization (10%)** The next step should be to transform the input text into a normal form of your choice. This should include the identification of sentences, bullet points and cells in tables.
* ﻿**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. Your system should remove words which are not “useful”. E.g. very frequent words or stopwords. You should also identify phrases suitable as index terms. Apply *tf.idf* as part of your selection and weighting step.
* **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, 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 Graphical User Interface (GUI). **Note** that the each article in the collection contains different fields. Make sure that a user can decide which field to search (**Hint:** one of the fields is the publication date of the article).
* **Engineering a Complete System (10%)** The final system should allow a user to have control over all the individual components, so in the final result we will have a complete search engine, not disperate code.

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 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.*

**CE706 - Information Retrieval 2021**

**Assigment 1**

Student ID

# Instructions for running your system (Engineering a Complete System)

*Include here instructions to run your system and control each individual component. 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.*

﻿Sentence Splitting, Tokenization and Normalization

*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. 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. 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. 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.*

1. <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge> [↑](#footnote-ref-1)