CE706 - Information Retrieval

Assignment 2

Alba García Seco de Herrera

June 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 indicated.*

The context of your task

To properly evaluate a system, your test information needs must be germane (relevant) to the documents in the test document collection, and appropriate for predicted usage of the system. Given information needs and documents, you need to collect relevance assessments. This is a time-consuming and expensive process involving human beings (in this case you). For tiny collections, exhaustive judgments of relevance for each query and document pair can be obtained. For large modern collections, it is usual for relevance to be assessed only for a subset of the documents for each query. The most standard approach is **pooling**, where relevance is assessed over a subset of the collection that is formed from the top *k* documents returned by many different IR systems (usually the ones to be evaluated).

The Document Collection (dataset) for this assignment you will use the dataset that you used in the first assignment (the *Shakespeare* dataset (https://www.elastic.co/guide/en/kibana/5.5/tutorial-load-dataset.html).

Your task

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

- **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 searching as effective and efficient 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** information needs covered by the collection and then compose a sample query for each.

- **IR systems (15%)** You are going to compare 2 IR systems. In the **first assignment,** you built an IR system, that would be your *system 1*. For your *system 2*, 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.

- **Pooling (15%)** You will construct your pool by putting together the **top 5** retrieval results from your 2 IR systems (your original from assignment 1 and the newly created one). You need to do this for **each** of your three queries. In the next step, you will judge every document in this pool.
    - **N.B.** Documents outside the pool are automatically considered to be irrelevant (Sparck Jones and van Rijsbergen, 1975)

- **Assessing relevance (20%)** your task is to make a binary judgement for each document on their relevance (relevance/non-relevance) and explain why.
    - For each information need pair (query) you need to assess if each document in the pool is relevant or not (if it satisfies the information need).

- **Evaluation (20%)** Once you have a test collection you can explore the effect of each IR system on the evaluation results. To do that you need to identify a suitable metric. Use P@3 and R@3 as the metric of choice for this assignment.

- **Plagiarism detection (10%)** Finally you need to choose one of your retrieval systems for retrieval for plagiarism detection, i.e., to retrieve documents that may be the sources of plagiarism for a suspicious document. Which metrics you will use to make your decision and why? According to the metric you choose which system you will use.

Tasks in summary: Using the dataset from assignment 1, decide on 3 pieces of information you want to learn from the dataset. Use your original IR system from assignment 1 and a modified version to retrieve the answers from the dataset. You will then create a pool and assess the relevance of the documets in the pool given each of the queries. Finally, you will compare both systems in terms of P@3 and R@3.

You will have noticed that the percentages above only add up to **90%**. This is because one of the important aspects of the project is that your work should be **well documented**. **10% of your mark will come from this.** The report should contain:

- Design and design decisions/justifications of your overall architecture
- The actual ground truth data that make up your test collection (i.e. queries with their matching documents)
- Evaluation results
- Discussion of your solution focusing on the comparison of both systems.

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**)

The submission should be submitted as a single *pdf 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 – SU - Information Retrieval 2023

Assigment 2

Student ID

Test collection (Task 1)

Include here the selected information needs and how they will be represented as a query.

Information need	**Query**

IR systems (Task 2)

Include here the details of your two IR systems and the difference between them.

Pool method (Task 3)

For each method retrieve the top 10 documents. Therefore for each query, you will have a maximum of 10 documents. Fill the following table with the ID of the top 10 documents that are retrieved for each of the queries by each of the systems:

	Query 1		**Query 2**		**Query 3**	
Rank	**System 1**	**System 2**	**System 1**	**System 2**	**System 1**	**System 2**
1	*doc id…*					
2						
3						
4						
5						
# different documents						

Relevance assessments (Task 4)

You need to construct a more complete description of the information being sought for each query. Notice that only containing the same words is not a valid criterion.

Fill the following table with the complete description of the relevance criteria and the ID of the relevant documents

	Description	ID of relevant documents
Query 1		
Query 2		
Query 3		

Evaluation (Task 5)

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.

	System 1		System 2	
	P@3	**R@3**	**P@3**	**R@3**
Q1				
Q2				
Q3				

Plagiarism detection (Task 6)

Include the discussion of your solution focusing on the comparison of both systems.