Information Retrieval: Assignment 1

Prof. Toon Calders, Ewoenam Tokpo

Deadline: 12/11/2020

This project is to be executed in groups of 2 to 3 students. The choice of data for the project is open; students are allowed to choose any interesting dataset of choice for the project.

1 Introduction

The goal of this assignment is to use an existing library, Lucene [1] (Java) or PyLucene [2] (Python), to create a document store with retrieval functionality. From past experiences it became apparent that using Lucene in java is far more straightforward than using PyLucene of which the installation often goes wrong.

At the end of the project, it is expected that the operational structure and details of Lucene will be understood. In addition, you should be able to implement a simple information retrieval system with Lucene that can index, search and retrieve text documents.

2 Dataset

For the document store of this project, you will be using a dataset of choice. Some suggested datasets include Wikimedia dump and the Stackoverflow dump. Using large datasets like the stackoverflow dump will be considered a plus for your project. However, smaller datasets are also acceptable.

3 Assignment

The objective of this project is to get familiar with Lucene and to study the feasibility of using Lucene to add retrieval capabilities to large datasets. The requirements are as follows:

1. Give a description of the functionality offered by Lucene, such as the types of indices that are included, the different score models, etc. How does Lucene store the index, does it have spell correction? And other operational properties and information about Lucene.

2. Implement a document retrieval system with Lucene. This part of the project should highlight three main aspects:

   - Document analysis and indexing
   - Query processing
   - Document search and retrieval

[1] https://lucene.apache.org/

[2] https://lucene.apache.org/pylucene/

Some documents can be manually labelled in order to demonstrate the search and retrieval aspect of the system. One acceptable approach is to use the title of a text, for example, the title of a question in Stackoverflow as the query, and only index the remaining parts of the documents. In this way a ground truth can be generated rather easily.

You are encouraged to explore other interesting functionalities that can be implemented on the datasets using Lucene. The core aspect of your application is the search capability, not the interface, hence, a text-based interface can be used.

4 Deliverables

1. The code of your project. Do not include bulky software libraries or large datasets in emails. The preferred way to share code is via a link to a publicly available GitHub repository. Include the link in your report.

2. The report in pdf format, to be submitted via BlackBoard. Do not submit zip-files, word documents etc., only the submission of a single pdf file will be accepted. The report should be approximately 8 pages in length.

A Note on Plagiarism

There is absolutely nothing wrong with using existing materials, you will even be commended for not reinventing the wheel, as long as you are not violating the copyright of other authors. Nevertheless, it is expected from you to clearly indicate whenever you used material that was not created by yourself. Clearly indicate in your submissions which parts constitute original work, which parts are taken from other works, and which parts were adapted from external sources. These sources have to be properly acknowledged in all your submissions. Concretely, this means at least the following guidelines are observed:

- Papers, books, webpages, blogs, etc. that were inspected while making the assignment will be referenced in a separate section "References". Citations to these materials are included in the text where appropriate.

- Text fragments exceeding one sentence that are copied from other sources are clearly marked as such. You could for instance include quoted text, definitions, etc. in italics, followed by a reference. An example of how to do this: Bela Gip (2014) defines plagiarism as *"The use of ideas, concepts, words, or structures without appropriately acknowledging the source to benefit in a setting where originality is expected"*

  References: (at the end of the document) Gipp, Bela. Citation-based plagiarism detection. Springer Vieweg, Wiesbaden, 2014. 57-88.

- When using code from other sources, indicate so in the report, and in the source code. This could for instance be done by adding a comment with a reference to the source of the function for each function that was copied from another source. It is recommended to include a separate folder "sources" in your GitHub repository with the original files from other authors that you used. Include source in the message of your commits.