Information Retrieval - Assignment 3

DEPARTAMENTO DE ELETRÓNICA, TELECOMUNICAÇÕES E INFORMÁTICA

MESTRADO INTEGRADO EM ENG. DE COMPUTADORES E TELEMÁTICA

ANO 2018/2019

Beatriz Coronha 92210

Pedro Santos 76532

1. Introduction

The goal of this assignment is to calculate performance metrics for the weighted retrieval method created in the other assignments using the Cranfield corpus. In this report will be presented the class diagram, the results and the instructions to run the code.

2. Class Diagram

The class diagram for the project is presented below (Figure 1) and it is in the folder of the assignment to be easier to analyse it.

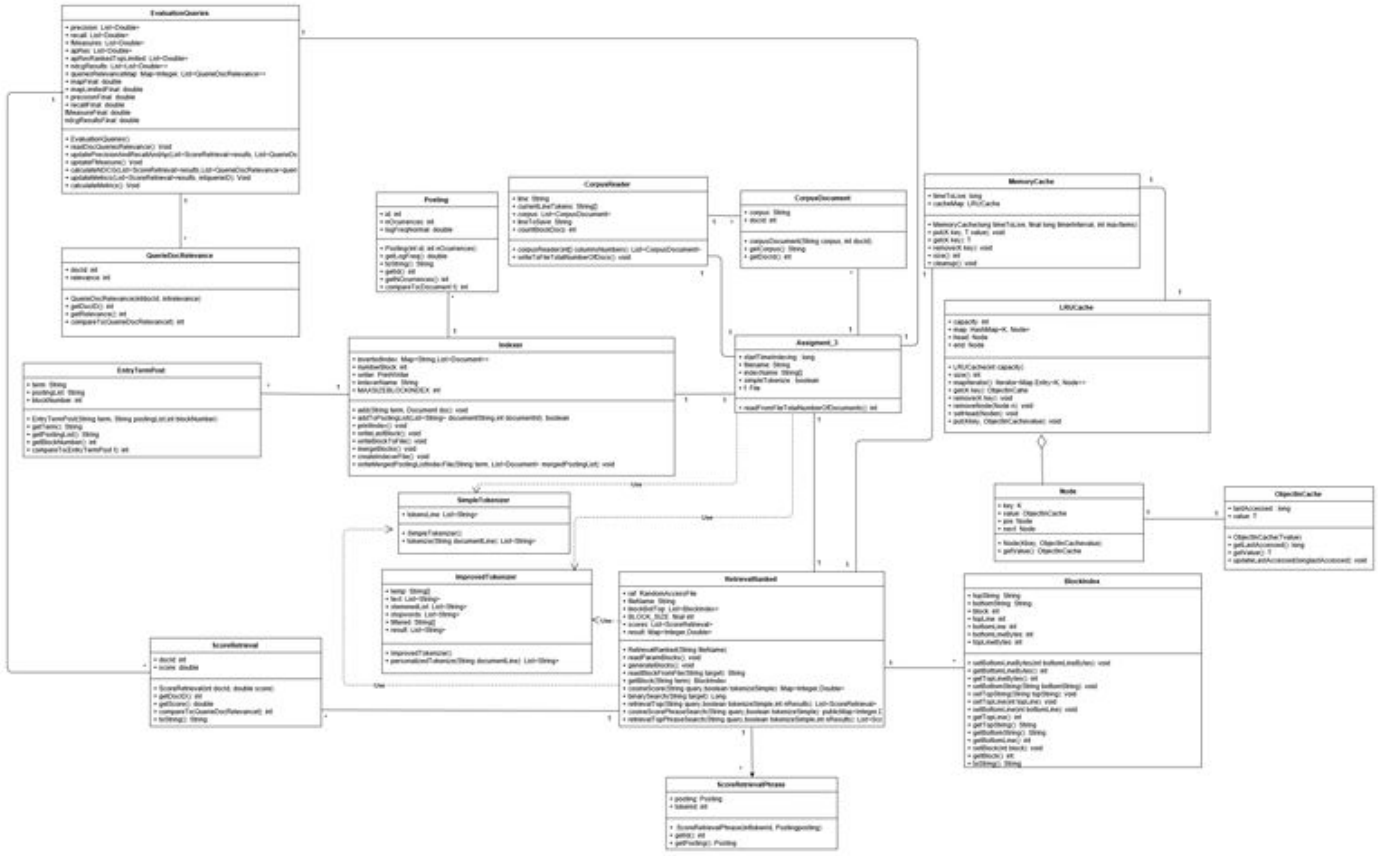

Figure 1: Class Diagram

3. Results/Tests

Simple Tokenizer:

Precision: 0.04
Recall: 0.66
F-measure: 0.08
Mean Average Precision: 0.31
Mean Precision at rank 10: 0.43
NDCG: 0.26

Query Throughput: 0.094 q/s
Query Latency: 10.67 ms

Improved Tokenizer:

Precision: 0.05
Recall: 0.72
F-measure: 0.09
Mean Average Precision: 0.34
Mean Precision at rank 10: 0.44
NDCG: 0.28

Query Throughput: 0.12 q/s
Query Latency: 8.37 ms

4. How to Run/Requirements

The only requirement that the user needs to worry about is the integration of the porter stemmer.

The first thing to do to be able to run this code is to install the porter stemmer, to do that is necessary to open the directory of the code in the terminal and run the following: mvn install:install-file -Dfile=lib/libstemmer.jar -DgroupId=org.tartarus -DartifactId=snowball -Dversion=1.0 -Dpackaging=jar. By doing that, the file “libstemmer.jar” is being installed with maven. The next step is to do a clean install, so

that the target will be cleaned before the installation. To do that, run this line in the terminal: mvn clean install in the “assignment_3” directory. Also, it is possible to install the dependency using netbeans.

The used tokenizer by default is the simple one, to change that it is necessary to change the variable “simpleTokenize” to false in the main class. If it is runned the program several times will create only the indexer the first time, so if we want to create a new indexer for example using another tokenizer it is necessary to delete from the folder all the files started by the word “indexer” and then run the program again to create the new indexer.