**The project report submitted by:**

Hedaya Ahmed El-Attar

ID: 171111

**Under the supervision**

**of**

Dr. Islam El Shaarawy

**Problem:**

* Build a search engine for Arabic Wikipedia pages and retrieve the results that match queries.

**System architecture:**

* **Front-end:** Java GUI – Java Console.
* **Back-end:** Apache Lucene.

**Implementation details:**

* **Phase 1:**
* Download Lucene demo.
* Running the demo on Windows Command Line “Prompt” for Lucene docs by using the following commands:

java -cp C:\Users\Owner\Desktop\lucene\core\lucene-core-7.5.0.jar;C:\Users\Owner\Desktop\lucene\demo\lucene-demo-7.5.0.jar org.apache.lucene.demo.IndexFiles -docs C:\Users\Owner\Desktop\lucene

java -cp C:\Users\Owner\Desktop\lucene\core\lucene-core-7.5.0.jar;C:\Users\Owner\Desktop\lucene\queryparser\lucene-queryparser-7.5.0.jar;C:\Users\Owner\Desktop\lucene\demo\lucene-demo-7.5.0.jar org.apache.lucene.demo.SearchFiles

* **Phase 2:**
* Download Arabic Wikipedia html files.
* Running the demo again for about 10 articles of Arabic Wikipedia.
* **Phase 3:**
* Evaluate the performance of indexing for 1000 articles:
* Create a java new project
* Add external JARs of Lucene (analyser- core- demo queryparser).
* Open the classes of indexing and searching of Lucene.
* Create folder the contains 1000 Arabic Wikipedia.
* Add this folder to the project.
* Import org.apache.lucene.analysis.ar
* Running both codes and recording that:

**Time of indexing:** 121200 milliseconds (1000 articles), 791530 milliseconds (60,000 articles).

**Index size:** about 6 MB (1000 articles), 415 MB (60,000 articles). **User happiness:** Free, without GUI.

* **Phase 4:**
* Improve the performance of the search engine by some edits on the source-code:

**Edit the run configuration arguments of java virtual machine to: -Xms512M**

**-Xmx1024M**

This will increase the heap size of JVM.

A screenshot of a cell phone

Description automatically generated

**Edit the default number of hits per page and set it to the max which is 1000.** This will decrease indexing and the query time.

**A screenshot of a cell phone

Description automatically generated**

**Merge segments until there's <= maxNumSegments especially because the index is static. So**, the total size of the index will be less than the size of the starting index.

**A screenshot of a cell phone

Description automatically generated**

**Increase the Lucene buffer size which is the index writer. So,** it can move data for RAMs faster to be processed which improve the speed.

**A screenshot of a cell phone

Description automatically generated**

**The Final result is improving the speed of indexing** about 1.627 times **and indexing size.**

**Time of indexing:** 1098 milliseconds (1000 articles), 486476 (60,000 article).

**Index size:** about 270 MB (60,000 articles).

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

* **Phase 5:**
* Developing a complete search engine with GUI and 60,000 Arabic Wikipedia articles.

**A screenshot of a cell phone

Description automatically generated**