Recuperación de Información

Práctica 4. Indexación

Santiago Álvarez Valdivia

Ioannis Efthymiou

November 10th, 2023

Program Usage

This program (Practica4.jar) is meant to read information about Simpsons episodes (in a particular format) and create a Lucene index from the information.

An invocation of the program has three parameters: the operation to perform, the kind of data to index, and the directory where the csv files to be indexed are.

There are two operations available, **index** (which creates a new Lucene index from scratch) and **append** (which adds documents to an index created previously).

There are two data types known, **episodes** and **scripts**, which contain a summary of episode information and the complete information of what the characters of a Simpsons episode say, respectively.

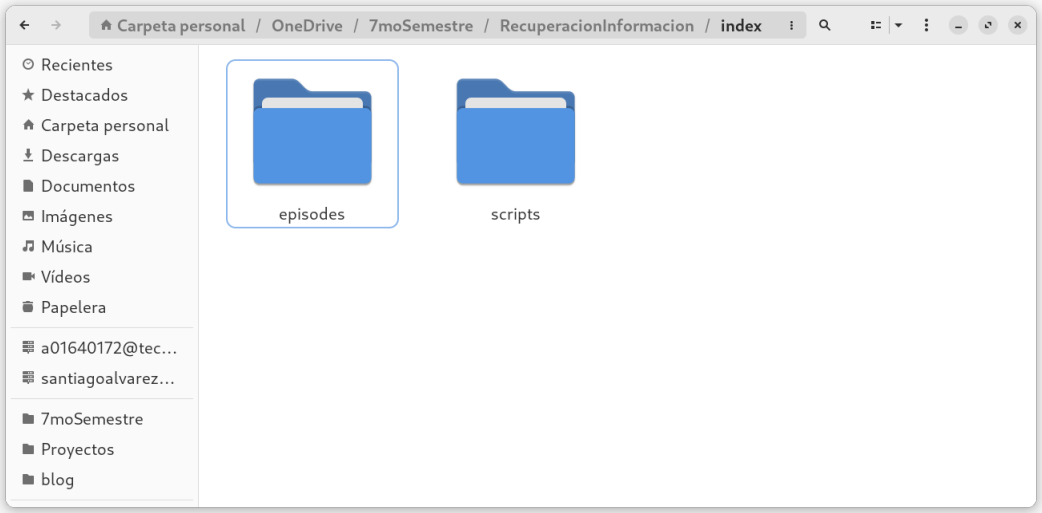
Then a typical invocation of the program would look like this:

```
java -jar Practica4.jar index episodes ./TestData/CapitulosUnidos
```

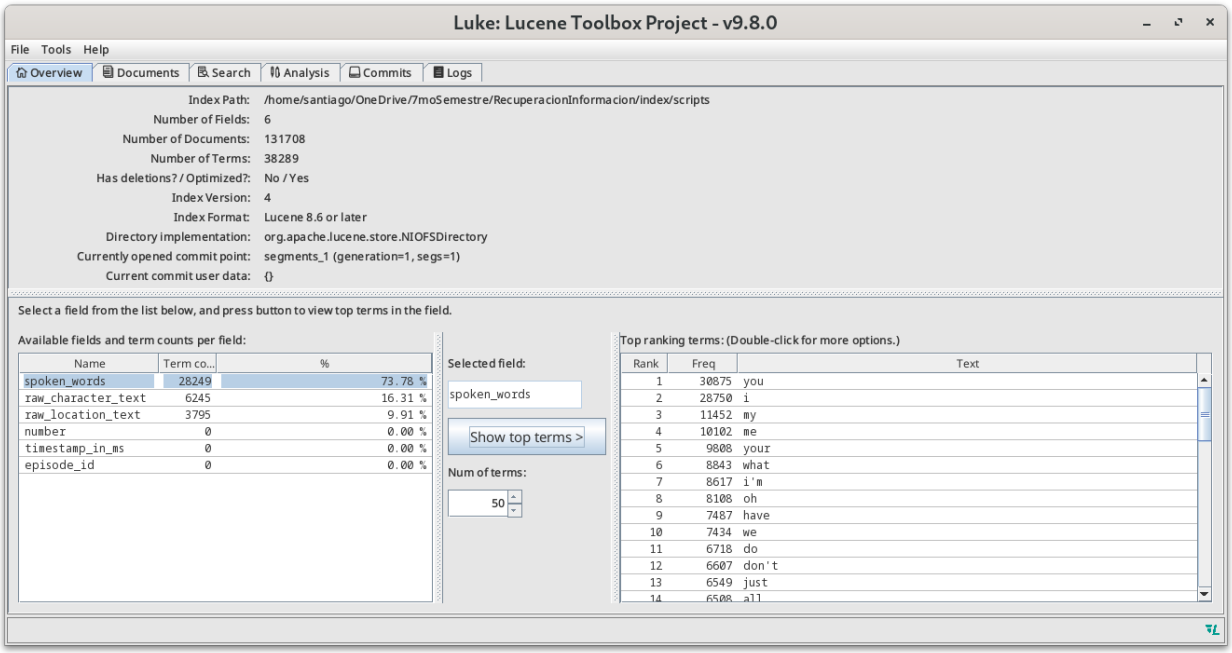
which would create a new index for the files in CapitulosUnidos, which are expected to have the format for episode summaries. In the case of the append operation, an index must already exist in the location where a new one would be created, otherwise we will get an error when the index is tried to be accessed.

Demo

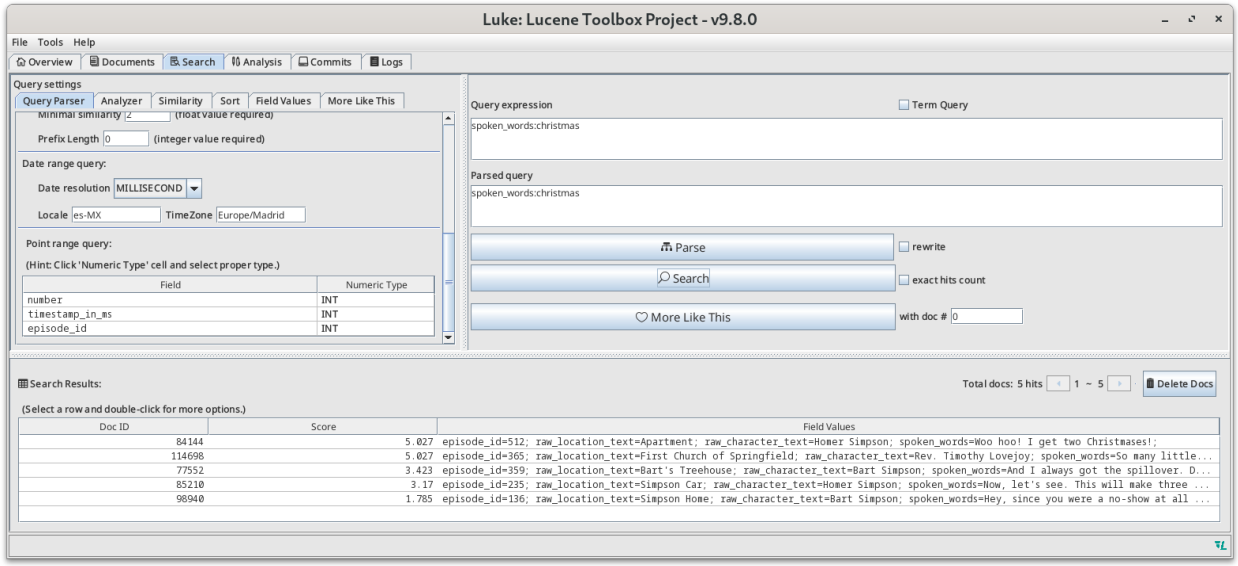
After running the program for both data types, we end up with this indexes



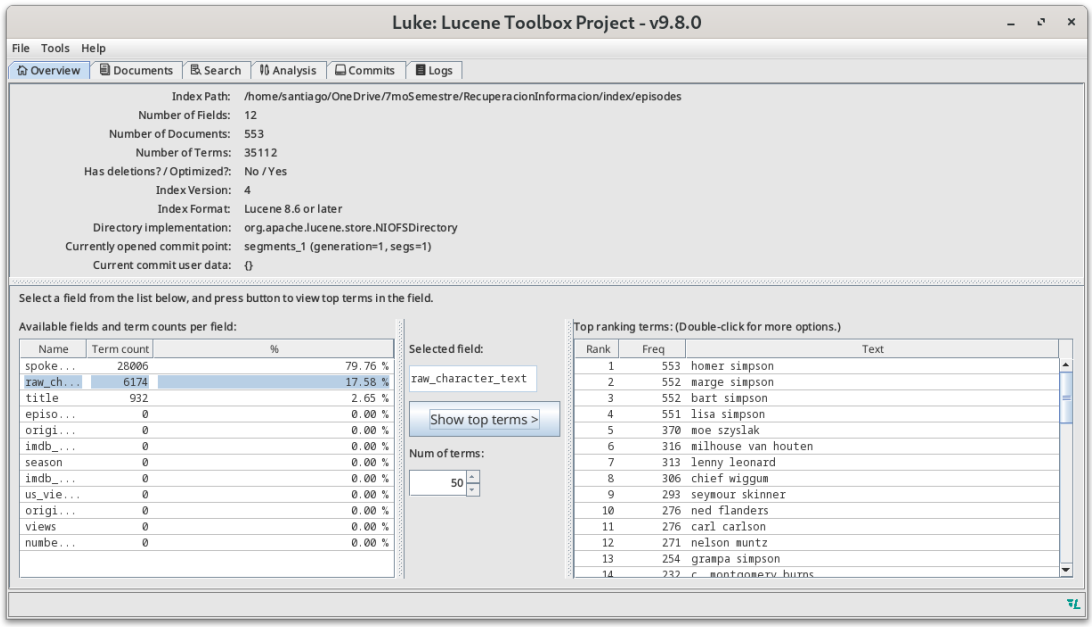
Opening the scripts index with Luke shows the indexed information



And we can perform searches on the indexed information



Meanwhile, this is how the episodes index looks like



We can also filter information on this index