ingenieur
wissenschaften
htw saar

Hochschule für
Technik und Wirtschaft
des Saarlandes
University of
Applied Sciences

Information Retrieval
Winter 2019/2020

Prof. Dr.-Ing. Klaus Berberich
Telefon: 06 81 58 67-243
klaus.berberich@htwsaar.de

Programming Assignment 2

The programming assignment will be discussed on **December 5**. To obtain bonus points, you have to submit your solution via Moodle by **December 3 at 12:00 (noon)**. Please submit your solution, consisting of source code files and possibly libraries, **as one zip archive**. Teams of up to three students are allowed.

Aufgabe 2.1 SQLite and Xerial (0.5 Points)

Install SQLite (https://www.sqlite.org) on your system. Download the JDBC driver Xerial (https://github.com/xerial/sqlite-jdbc) in its most recent version. Familiarize yourself with how you can access a SQLite database using the command line and from within a Java program. You will find examples of how this can be done at the given URL. As a solution please submit the output of the `.version` command when run in your SQLite command line.

Aufgabe 2.2 Database (1 Point)

Create a SQLite database named `nyt.sqlite` to store the document collection. Create two tables within the database:

- `docs` stores meta data about the documents. It should contain the following attributes: (i) `did` as a unique document identifier, (ii) `title` as the title of the document, and (iii) `url` as the URL of the document.
- `tfs` stores how often a term occurs within a specific document. It should contain the following attributes: (i) `did` as the document identifier of the document, (ii) `term` as the textual representation of the term, and (iii) `tf` as the term frequency of the term within the document.

Choose suitable data types for each of the attributes. A list of data types available in SQLite can be found at:

https://www.sqlite.org/datatype3.html

As a solution, please submit the `CREATE TABLE` commands that you use to create the two tables.

Aufgabe 2.3 Importing Documents (1.5 Points)

Next, we will import documents into our database. To this end, you will extend the classes `Importer` and `Parser` from Programming Assignment 1.

- The method `importDirectory` in `Importer` should call the method `parse` in `Parser` for every file with suffix `.xml` that it encounters.
- The returned instance of `Document` should be added to our database. For each document, one row should be inserted into the table `doc`. For each distinct term from the document, one row should be inserted into the table `tfs`.

For instance, for the following document

- `id` : 23
- `title` : ABC
- `url` : http://www.nytimes.com/abc
- `content` : [a, b, c, a, b, a]

a single row

- 23, ABC, http://www.nytimes.com/abc

is inserted into `docs`. In the table `tfs`, the following three rows are inserted

- 23, a, 3
- 23, b, 2
- 23, c, 1

Please use the class `java.sql.PreparedStatement` to speed up the insertion of rows into our database. This allows precompiling SQL statements and batch inserts. Familiarize yourself with the method `addBatch` and `executeBatch`. Use a batch size of ten documents. Please submit your code as a solution.

Aufgabe 2.4 Computing Document Collection Statistics (1 Point)

Different retrieval models require different statistics about the document collection. In this exercise, we will create additional tables that contain such document collection statistics. We will make use of the command `CREATE TABLE ... AS (SELECT ...)` to create additional tables from the table `tfs`. Please create the following additional tables

- `dls` stores document lengths. It should contain the following attributes: (i) `did` as the document identifier and (ii) `len` as the total number of term occurrences within the document
- `dfs` stores document frequencies. It should contain the following attributes: (i) `term` as the textual representation of the term and (ii) `df` as its document frequency in the collection.
- d stores the total number of documents in the collection. It should have a single attribute (i) `size` and contain only a single row.

As a solution, please submit the `CREATE TABLE` commands that you use to create the three tables.