Project Portfolio – Source data

A small collection of sample data extracted from Stack Overflow with close to 14000 posts (out of more than 16 million) and 33000 comments is provided for use in the portfolio project. In addition, data to aid the construction of inverted indexing is included. Find below instructions on how to load this data into your own database as well as a description of what it contains and hints on why this calls for a much needed better database schema.

To load the data to your own database

The data is provided in the (database-dump) file **stackoverflow_universal.backup**[1] and you can create your own database and load the data into that new database using the following two psql commands (assuming that you current directory is where the file is located:

```
psql -U postgres -c "create database stackoverflow"
psql -U postgres -d stackoverflow -f stackoverflow_universal.backup
```

Notice that this is a way to go, if you are using your own local Postgres database server. If you alternatively (or in combination) prefer to use the remote server on rawdata.ruc.dk, you can use the following command (replace X in rawX with your group number):

```
psql  -h rawdata.ruc.dk -p 5432 -U rawX -W -f stackoverflow_universal.backup
```

On rawdata.ruc.dk you must use the default database rawX already created for your group – you don't have privilege to create databases.

The content that you load with **stackoverflow_universal.backup** includes the 4 tables shown in figure 1.

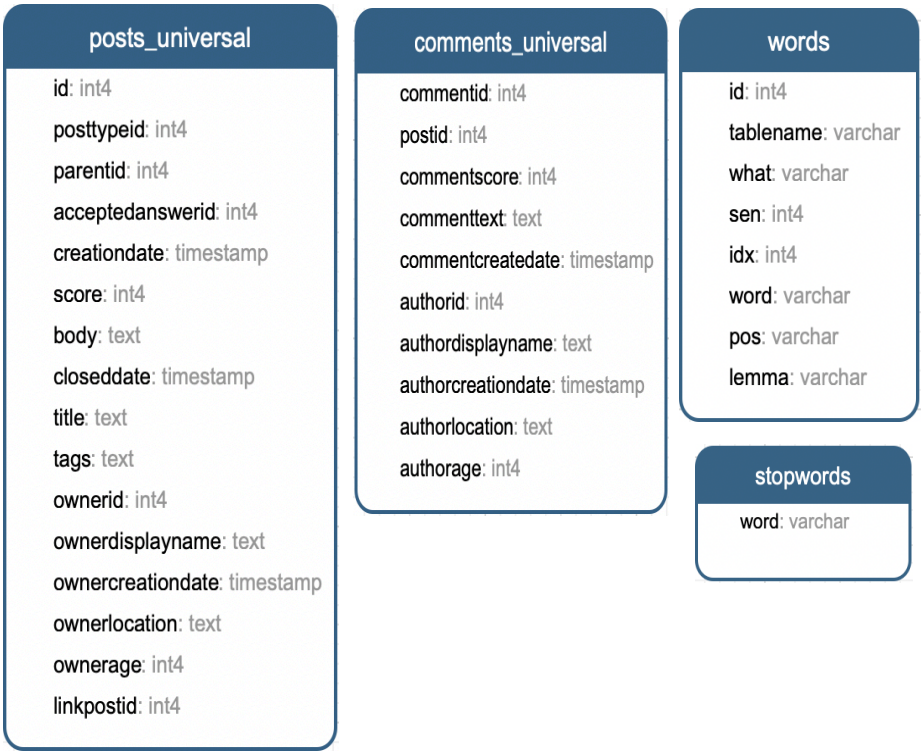

Figure 1: The four tables included in **stackoverflow_universal.backup**

[1] Or alternatively you can use the SQL script file **stackoverflow_universal.sql** in the exact same way as the **backup** version. Differences are that the **sql** script is easier to read (you can see exactly what's going on) while the **backup**-version is much faster to load.

The two tables posts_universal and comments_universal

These two tables provide the data downloaded from the stackoverflow site.

The attribute *posttypeid* specifies the type of the post with the value either 1 or 2 corresponding to either question or answer. A question post will not have a parent (thus *parentid* is null), while an answer post always refer to a parent post via *parentid*, which is a reference to the question post that the answer post refers to.

The words table (description)

To simplify the indexing for Information Retrieval on the textual columns in the stackoverflow data, some data preparation has already been done and provided in the table **words**. This table, shown with a small excerpt in figure 2, can be used as source to build various kinds of inverted indexes. As it is it also comprises a combined inverted index for **posts** and **comments**. The column **id** is the id of the **comment** or **post** where the **word** appears. The columns **tablename** and **what** are used to distinguish what an entry refers to (either the **title** or **body** in **posts** or the **text** column in **comments**). The columns **sen** and **idx** together specifies the position of the word, by sentence and word number. In addition, the table provides a lemmatization[2] of words with the column **lemma** and a specification of the word category (part-of-speech) with the **pos** column. The specifications in the latter are given using the Penn Tree Bank tags[3]. From the excerpt shown in figure 2, we see that sentence 8, in the **body** of the **post** with **id** 60496, includes the phrase "when something goes wrong" as the last 4 of 15 words in that sentence (must be the last 4 words due to the punctuation mark as word 16).

id	tablename	what	sen	idx	word	pos	lemma
60496	posts	body	8	12	when	WRB	when
60496	posts	body	8	13	something	NN	something
60496	posts	body	8	14	goes	VBZ	go
60496	posts	body	8	15	wrong	JJ	wrong
60496	posts	body	8	16	.	.	.

Figure 2: Small excerpt from the table words

The most important part of this table is the inverted index it provides on the post table. You can consider to use additional data from this table, but this will not be a requirement in the portfolio project.

The stopwords table

The stopwords table is a single column table listing a set of words commonly considered as stop words[4].

[2] https://en.wikipedia.org/wiki/Lemmatisation
[3] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[4] https://en.wikipedia.org/wiki/Stop_words