Computing the SemanGit Dataset

Dennis Kubitza
Matthias Böckmann
Damien Graux

For version 1.1 of the converter

Contents

1 Before you start

1.1 What is the SemanGit RDF Dataset ?

The SemanGit RDF Dataset is cumulative turtle-dataset [1] of publicly accessible meta-data on GitHub. It is computed from datasets provided by `www.ghtorrent.com` [1] and follows the SemanGit ontology which we develop and provide under `https://github.com/SemanGit/SemanGit/tree/master/Documentation/ontology`. For documentation, additional information and references please visit our website `www.semangit.de`

1.2 Licenses

For following these instructions and using the generated RDF, you will need to comply with different license agreements.

- The required input datasets are licensed by the GHtorrent project [1]. The according license agreements can be found here: `http://ghtorrent.org/faq.html`
- The processing of the datasets requires the usage of the Java Runtime Environment. Depending on the distributor you choose different license conditions may apply.
- The processing of the dataset requires the usage of software developed by us. The software can be obtained from `https://github.com/SemanGit/SemanGit`. The software is subject to the MIT License:

[1] `https://www.w3.org/TR/turtle/`

1.3 Warnings

- The computations described in the following parts are requiring around 1.3 TB of free space. Running out of space may harm your system.
- We recommend doing all of the following steps on one of the following:
  - A separate empty/new partition without any important files
  - An external hard drive without any important files
  - A virtual Machine with at least 1.3TB of free space
- The computation requires Java Runtime Environment 1.8 or later to be installed.

2 Download a monthly dump

1. Visit `www.ghtorrent.org`
2. Go to *Downloads* > *Downloads*
3. Click on mysql-XXXX-XX-XX to download the selected dump

3 Extract the dump

1. Create a new folder with an arbitrary name

   ```
   mkdir foldername
   ```

2. Extract your downloaded .tar.gz to the new folder

   ```
   mv path/to/downloaded/dump foldername
   tar xfz mysql-xxxx-xx-xx.tar.gz
   ```

Note: For the further processing it is essential that the extracted .csv files all end up in the same reserved folder which shouldn't contain any other files.

4 Download the Converter

1. Visit `https://github.com/SemanGit/SemanGit`
2. Click on **Converter**
3. Click on **Clone or download**, Download the .zip and extract it to an arbitrary location.
4. After extraction, you can delete the downloaded .zip file.

Note: Step 2 ensures that you get the latest functioning release. If you want to use the experimental version (not recommended), you can visit `https://github.com/SemanGit/Converter` instead.

5 Execute the Converter

Warning: The following steps require that you have about another 800GB of free space left on the partition where you extracted mysql-XXX-XX-XX. Having not enough space might harm your system. If you notice that you run out of space during the processing, open your systems Task Manger and quit the JAVA Environment. You have to perform a manual clean-up afterwards for freeing the allocated space, by deleting all files which end on .ttl

1. Make sure that you have a working Java Run-Time Environment. You can find instructions on checking your JRE here: `https://www.java.com/en/download/help/version_manual.xml`

2. Open your systems Console/Terminal or Shell.
3. Navigate to the location where you downloaded the Converter */out/artifacts/converter_main_jar*
4. You can now execute

   ```
   java -jar converter_main.jar path/to/extracted/files [options]
   ```

   from your console.

   - **path/to/extracted/files**: The path where your extracted .csv files (**from Step 2**) are located.
   - **options** A list of all supported options can be obtained from `https://github.com/SemanGit/SemanGit` $>$ *Converter* @ $XXXXXXX$ or in the **readme.md** contained in downloaded Converter folder.

References

[1] Georgios Gousios. The ghtorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 233–236, Piscataway, NJ, USA, 2013. IEEE Press.