CedTMart Manual for Data Preprocessing

This manual provides a guideline how to perform different tasks including cleaning, converstion, partitioning RDF data in processing phase.

Required Software

****

- Java 1.6 or higher
- Any Linux distribution

Instructions

****

Step 1: Download CedTMart.Preprocessor.rar file from the CEDAR website

Step 2: Unrar the file

Step 3: If your data is not N3 Serialization format then Run allWithoutCompression.jar by "**`java -jar allWithoutCompression.jar`**". It will return the menu shown below.

```
----------------------------------------
|--------        CTM V2        --------|
----------------------------------------
Using parameter : nsPath - /projets/cair/cedar/_ns
Using parameter : invalidPath - /projets/cair/cedar/_invalidTriple
Using parameter : compressedPath - /projets/cair/cedar/_compressed
Using parameter : psPath - /projets/cair/cedar/_ps
Using parameter : posPath - /projets/cair/cedar/_pos
Using parameter : rdfPath - /projets/cair/cedar/__rdf
Using parameter : indicatorPath - /projets/cair/cedar/_indicator
Using parameter : comparePath - /projets/cair/cedar/_compare
Using parameter : n3Path - /projets/cair/cedar/__n3
----------------------------------------
----------------------------------------
Type number to execute :
        Clean all existing processed data - 10000
        RDF to N3 converter - 101
        N3 Reader/Partitionner(PS) - 102
        Predicate Reader/Splitter(POS) - 103
        Compressor for PS files (external script)  - 104
        Comparator for S/O arrays - 106
        Distributor of compressed files - 107
        Exit - -1
#
```

Step 4: **Select 101** for the conversion. It will convert your dataset on RDF/XML into N3 format

Step 5: Exit by typing -1

Step 6: You will find a folder called "__n3" (**it is double underscore**) folder under the destination root folder. If not, then create one.

Step 7: Copy your converted N3 data to the __n3 folder.

Step 8: There are two script files **`compress.sh`** and **`sort.sh`** . Make sure they are executable. Add "X" to make them executable

Predicate Partitioning Steps

Step 9: If your data is not N3 Serialization format then Run allWithoutCompression.jar by "**`java –jar allWithoutCompression.jar`**". It will return the menu shown in Step 3.

Step 10: Run Predicate Partitioner by selecting "102".

Step11: The system will ask to choose number of threads. Type how many threads you want (4 threads are recommended). Wait for the task to be completed.

Merging Steps

Step12: At the end of partition, the system will ask "Do you want to merge each predicate into single file right now?". You should type **Y**.

Step13: Once the merge is done, you should exit by typing "**-1".**

Step14: You will find the "_ps" folder which is generated automatically. Remove all subfolders in the "_ps" folder

Predicate Object Partitioning Steps

Step 15: If your data is not N3 Serialization format then Run allWithoutCompression.jar by "**`java –jar allWithoutCompression.jar`**". It will return the menu shown in Step 3.

Step 16: Run Predicate Partitioner by selecting "103".

Step17: The system will ask to choose number of threads. Type how many threads you want (4 threads are recommended). Wait for the task to be completed.

Sort Steps

Step 18: Run "sort.sh" (make sure the script file is executable).

Compression Steps

Step 19: Open "compress.sh" file and change the maximum and minimum heapspace for JVM in "-Xms -Xmx" for both "SO" and "OS" compression. **By default it is 2 GB and 8 GB for -Xms and -Xmx respectively**.

Step 20: Run "compress.sh" (make sure the script file is executable).