Big Data Infrastructure
MapReduce Lab

Connect to the machine: ssh <username>@diufrm202

The folder **/bdi_2018/data/** on HDFS contains two folders, NYTimes_articles (that in turn contains 128 files) and btc09 (that in turn contains 2 files), respectively.

Use the Hadoop cluster to complete the following tasks.

1. Run `WordCount.java` giving it as input all the files contained in the HDFS directory `/bdi_2018/data/NYTimes_articles`. If you look at the output of the program you can notice that the program cannot deal with punctuation and other symbols. For example, the strings `"Monti"` and `"Monti,"` are considered different because of the comma ending the latter.

2. Fix `WordCount.java` to make it able to deal with such situations.

3. Fix WordCount.java to deal also with HTML-entities, that is, to substitute them with the character they represent.

The folder `/bdi_2018/data/btc09` contains quadruples of the form

```
<subject> <predicate> object <provenance>
```

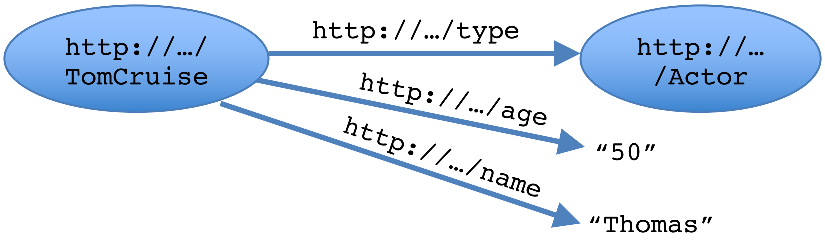
(the four fields are tab-separated) where `subject`, `predicate`, and `provenance` are URIs, while `object` can be either an URI `<object>` or a string (for example, `"46"`); in the latter case we say that `object` is ***a literal***.

If you take into consideration the first three components of each quadruple, you obtain the list of the edges of a graph: `<predicate>` is the label of the directed edge connecting `<subject>` to `<object>`.

For example,

```
<http://…/TomCruise> <http://…/type> <http://…/Actor>
<http://…/TomCruise> <http://…/age> "50"
<http://…/TomCruise> <http://…/name> "Thomas"
```

encodes the following graph.



With reference to the data we have just described, complete the following assignments:

4. Count the number of literals linked to each node and filter out all the nodes with less than five literals. For example, `<http://…/TomCruise>` has only two literals (namely `“50”` and `“Thomas”`) so it is filtered out.

5. Compute the ***in-degree*** and the ***out-degree*** of each node with at least 10 literals (the in-degree of a node is the number of edges ending in that node, while the out-degree of a node is the number of edges starting from that node).

For all the exercises use the HDFS directory `/bdi_2018/<YOUR_USERNAME>` to store the output of your programs.
Send all the code you produced together with the HDFS path in which you stored the outputs to paolo.rosso@unifr.ch.
The code must be well indented and commented.

DEADLINE: midnight 12th December 2018.

Good luck ☺