

Socializer

Distributed social network event extractor

User Guide

Radek Burget



Socializer User Guide

Configuration

The necessary configuration includes the credentials to the social network accounts that are used as data sources, the configuration of the underlying storage and the Spark cluster parameters.

Credentials

Set Facebook and Twitter credentials in `configuration.xml`.

```
<property>
  <name>twitter.consumer.key</name>
  <value>key</value>
</property>
<property>
  <name>twitter.consumer.secret</name>
  <value>secret</value>
</property>

<property>
  <name>facebook.app.id</name>
  <value>id</value>
</property>
<property>
  <name>facebook.app.secret</name>
  <value>secret</value>
</property>
```

HBase

Copy content of your `hbase-site.xml` into project's `hbase-site.xml`. This file can be found in your HBase installation folder.

Spark

Optionally you can specify some Spark properties into `configuration.xml`. These properties won't be overridden by `spark-submit` arguments. See all available properties at <https://spark.apache.org/docs/latest/configuration.html>.

Cluster

Set number of partitions of you cluster in `configuration.xml`. This number means how much nodes cluster contains.

```
<property>
  <name>cluster.partition_num</name>
  <value>1</value>
</property>
```

Usage

Before the usage, the target RDF repository must be set up in the cluster as described in the [Timeline Analyzer Installation Guide](#). The target HBase tables are automatically created by Halyard tools or they may be created manually.

Creating tables

Created HBase tables has `_halyard / _hgraphdb` suffix. You don't have to write this suffix when using application because it will be automatically added in the code.

Sometimes, there it is not possible to create required tables via `spark-submit` job.

If you are not able to create the table (`spark-submit` job freeze on start), use this builder:

```
java -cp socializer-1.0-jar-with-dependencies.jar cz.vutbr.fit.xtutko00.TableBui
```

Downloading data

Submit built jar into Spark cluster with `spark-submit`. The following command-line options are recognized:

- `-hc, --hbase_client <arg>` HBase client (halyard or hgraphdb)
- `-ht, --hbase_tablename <arg>` name of the HBase table (without suffix)
- `-s, --sources <arg>` file with sources, each on new line (line format: T/F:sourceName)

An exaple run:

```
$spark-submit --class cz.vutbr.fit.xtutko00.MainSpark socializer-1.0-jar-with-de
```

Where the `sources.txt` file contains the list of social network accounts that should be downloaded.

Acknowledgements

This work was supported by the Ministry of the Interior of the Czech Republic as a part of the project Integrated platform for analysis of digital data from security incidents VI20172020062.