SciDetect Documentation

Nguyen Minh Tien
Minh-Tien.nguyen@imag.fr
Cyril Labbé
first.last@imag.fr

March 2015

Revision History

Version	Date	Author	Comment
1.4	13-02-2015	MT	Initial deployment
1.41	17-02-2015	MT	Added support for XML and XTX
2.0	25-02-2015	MT	Added multiple configurable parameters

Contents

1 Installation Requirements

A runnable program for the SciDetect software is implemented in:

`ScigenChecker_Local.jar`

It can be used as a stand-alone Java program. The program component requires `Java SE 6` or higher, with an additional libraries for pdf converter(included in `lib/`); Furthermore the configuration file (`config.txt`) and directories for log files (`logs` and `detaillogs`) are required by the client.

2 Usage

2.1 Command line client-side

SciDetect program is included in a runnable JAR file. The program is started by invoking:

`$java -jar ScigenChecker_Local.jar <parameters>`

Where `<parameters>` stands for a combination of one or more of the following command line options:

`-c <path_to_check>` gives the path to the directory (or file) that need to be checked;

`-l <log_filename>` gives path and name of the log file (defaults to `/logs/start_time.xls`);

`-d` Save detail log (optional, default false).

Typical use:

`$Java -jar ScigenChecker_Local.jar -c /tien/Test_demo -l /tien/Test_log.xls -d`

2.2 Supported file types

At version 2.0 `ScigenChecker_Local` currently supports .PDF and two specific Springer xml format namely `.XML` for `A++` format `.XTX` for PDF extraction of PDF files

3 Configuration

A configuration file (`config.txt`) should be accessible by the program. It should be found in the same directory with the `ScigenChecker_Local.jar`. The config file contains following information:

3.1 Path to sample folder

`samples data/samples`

This is used to set the directory where samples of texts produced by known generator can be found. This directory contains one directory per *classes*. One directory contains examples that are representative of its class. In a standard release, the `data/samples` directory contains four subdirectories with texts generated by the following generator:

`http://thatsmathematics.com/mathgen/` (dir `data/samples/Mathgen`);

`https://bitbucket.org/birkenfeld/scigen-physics` (dir `data/samples/Physgen`);

`http://www.nadovich.com/chris/randprop/` (dir `data/samples/Propgen`);

`http://pdos.csail.mit.edu/scigen/` (dir `data/samples/SCIgen`).

New subdirectories can be added. This can be done for two purpose:

1. add a corpus that represents fairly enough a particular field. By setting appropriate threshold, this will flag papers that appeared to be too far from that field.
2. In case new a generator appears, new samples (pdf) can be added in a new subdirectory (in `data/samples`) containing a representative corpora of the new class.

3.2 Threshold configuration

```
Threshold_Scigen        0.48    0.56
```

A line starting with `Threshold_Dirname` is used to define thresholds needed to take decisions to assigned tested texts the class for which examples can be found in the directory `Dirname`. There should have one line (i.e. two Thresholds) per classe. These values are 2 real numbers between 0 and 1. The smallest one is use to take the decision to assigned the tested paper (almost certainly) to the class. The second one is used as a threshold for suspicion for containing parts of generated text.

The previous example (concerning Scigen class) has the following meaning. Given distances from the tested text to its nearest neighbour in the set of samples (i.e. texts found in the Scigen dir):

If the distance is greater than 0.56, then it is reasonably believable that this is a genuine article.

From 0.56 to 0.48, there is a chance that this article or part of this article is Scigen generated.

If the distance is less than 0.48, there is a very high chance that this is an automatic Scigen generated article.

If new samples are added to the sample folder, the threshold configuration should also be added, if not the default-threshold values are used (0.48 and 0.56).

3.3 Path for log files

```
Default_log_folder      logs/
Default_detail_log_folder       detaillogs/
```

These lines are use to set the default log folder and a default detail log folder (see section ?? for more information). In case the path to a log file is not set (no -l parameter), the log file will be saved in the default log folder under the name: `time_date.xls` (e/g: 09:46 25.02.2015.xls means the check was started at 9:46 on 25/2/2015).

3.4 Max text length

```
Max_length      30000
```

This set the max length in character (including white space char) for a text to be eligible for classification. This parameter is used in order to avoid miss classification: when an article is too long, this cause the characteristic of the article to becomes too generic and very long paper may be misclassified (without splitting misclassification rate: ??).

The default value is set at 30000 characters (about 15 pages). A longer text will be splitter into several part which are tested individually.

4 Make use of detail logging

The detail log (parameter -d) stores all the distances from the text under test to all other samples in the sample set (i.e. all texts in all directories found at `/data/sample`). This can be use to get a more detail look at the results.

For example: an article returned with a distant to the nearest neighbour that barely pass the threshold. Turning on the detail log for that article and checking the results may help the decision.