## Revision History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Author</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4</td>
<td>13-02-2015</td>
<td>MT</td>
<td>Initial deployment</td>
</tr>
<tr>
<td>1.41</td>
<td>17-02-2015</td>
<td>MT</td>
<td>Added support for XML and XTX</td>
</tr>
<tr>
<td>2.0</td>
<td>25-02-2015</td>
<td>MT</td>
<td>Added multiple configurable parameters</td>
</tr>
<tr>
<td>2.1</td>
<td>09-03-2015</td>
<td>MT</td>
<td>Separated text and corpus class</td>
</tr>
<tr>
<td>2.2</td>
<td>25-03-2015</td>
<td>MT</td>
<td>Added Web Service options</td>
</tr>
<tr>
<td>2.3</td>
<td>13-04-2015</td>
<td>MT</td>
<td>Added option to scan individual file</td>
</tr>
<tr>
<td>2.4</td>
<td>05-06-2015</td>
<td>MT</td>
<td>Bugs fix and improvements</td>
</tr>
</tbody>
</table>
Contents

1 SciDetect Local 3
  1.1 Installation-Requirements-Quick start ................................. 4
  1.2 Usage ........................................................................ 5
    1.2.1 Command line client .................................................. 5
    1.2.2 Supported file types .................................................. 5
  1.3 Configuration .................................................................. 6
    1.3.1 Path to sample folder .................................................. 6
    1.3.2 Threshold configuration .............................................. 6
    1.3.3 Path for log files ...................................................... 7
    1.3.4 Max-Min text length ................................................... 7

2 SciDetect Web Service 8
  2.1 Installation Requirements and Usage ..................................... 9
    2.1.1 Web application ....................................................... 9
    2.1.2 Web service client ..................................................... 9
    2.1.3 Usage ................................................................. 9
  2.2 Configuration ................................................................ 10
    2.2.1 Client Configuration .................................................. 10
    2.2.2 Server Configuration ............................................... 10

3 Extra information 11
  3.1 Make use of detail logging ................................................ 12
  3.2 Tuning/Setting Thresholds ............................................... 13
Chapter 1

SciDetect Local

This chapter is for the localize version of SciDetect where everything is performed on a local machine.
CHAPTER 1. SCIDETECT LOCAL

INSTALLATION-REQUIREMENTS-QUICK START

Installation  A stand-alone Java program, the documentation and the source code are available at the following URL: [http://scidetect.forge.imag.fr](http://scidetect.forge.imag.fr)

Requirements  The stand-alone Java program requires Java SE 6 or higher. It is also using an additional libraries for pdf converter (should be included in the `lib/` directory).

Quick start  The runnable program for the SciDetect software is packaged inside:

```plaintext
SciDetect_Local.jar
```

The following are needed:

- The configuration file (`config.txt`)
- The samples directory directories (`web/WEB-INF/data`
CHAPTER 1. SCIDETECT LOCAL

USAGE

1.2.1 Command line client

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

```bash
$java -jar SciDetect_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 containing the files to be checked or the path to the individual file that need to be checked;
- `-l <log_filename>` gives the name of the log file (defaults to `/logs/start_time.xls`);
- `-d` Save detail log (optional, default false).
- `-h` Show usage.

Typical use:

```bash
$java -jar SciDetect_Local.jar -c /tien/Test_demo -l /tien/Test_log.xls -d
```

1.2.2 Supported file types

At version 2.1 SciDetect_Local currently supports .PDF and two specific Springer xml format namely .XML for A++ format .XTX for PDF extraction of PDF files
A configuration file (config.txt) should be accessible by the program. It should be found in the same directory with the SciDetect_Local.jar. The config file contains following information:

1.3.1 Path to sample folder

```
# Where samples can be found
samples web/WEB-INF/data/samples
```

This is used to set the directory where samples of texts produced by known generators can be found. This directory contains one directory per classes (i.e. per known generator). One directory contains examples that are representative of its class. In a standard release, the web/WEB-INF/data/samples directory contains four subdirectories with texts generated by the following generator:

- [http://thatsmathematics.com/mathgen/](http://thatsmathematics.com/mathgen/) (dir data/samples/Mathgen);
- [https://bitbucket.org/birkenfeld/scigen-physics](https://bitbucket.org/birkenfeld/scigen-physics) (dir data/samples/Physgen);
- [http://www.nadovich.com/chris/randprop/](http://www.nadovich.com/chris/randprop/) (dir data/samples/Propgen);

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

1. Adding 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. When a generator appears, new samples (pdf) can be added in a new subdirectory (in data/samples) containing a representative corpora of the new class.

1.3.2 Threshold configuration

```
# Defining Thresholds for Scigen
Threshold_Scigen 0.48 0.56
```

A line starting with Threshold_Dirname is used to define thresholds. Thresholds are needed to take decisions to assigned tested texts to a class. Examples of each class 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 (i.e new dir), the threshold configuration should also be added, if not the default-threshold values are used (0.48 and 0.56).

1.3.3 Path for log files

```plaintext
# Set the default 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 3.1 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).

```
INDEX-53.txt is a Scigen 0.34236384 data/samples/Scigen/INDEX-scigen25.txt
INDEX-53.txt is a Physgen 0.47908222 data/samples/Physgen/INDEX-physgen7.txt
INDEX-011.txt is Genuine 0.60918242 data/samples/Scigen/INDEX-scigen41.txt
INDEX-013.txt is Genuine 0.61375975 data/samples/Scigen/INDEX-scigen25.txt
```

1.3.4 Max-Min text length

```plaintext
# the maximum, minimum size of a text
Max_length 30000
Min_length 10000
```

This set the max(min) 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: 0.13% or 42 misclassification/31577 samples). When the article is shorter than Min length, it will be marked as cant classify.

The default value for max length is set at 30000 characters (about 10 pages); a longer text will be split into several part which are tested individually. Default min length is set at 10000 characters.
Chapter 2

SciDetect Web Service

A web service version of SciDetect is also provided and will be presented in this chapter.
CHAPTER 2. SCIDETECT WEB SERVICE

INSTALLATION REQUIREMENTS AND USAGE

2.1.1 Web application

The Java web application implementing the web service requires Apache Tomcat 7 and Java SE 6 or higher.

The web application and all required runtime libraries are contained in the deployment package file SciDetectServerXX.war which must be deployed on a Tomcat server.

The web application caches some of its data in a temporary directory (/tmp/tomcat7_tmp) and should be clean periodically.

2.1.2 Web service client

A client for the SciDetect web service is implemented in SciDetectClient.jar and can be used as a stand-alone Java program. The client component requires Java SE 6 or higher, no additional libraries are needed; However the configuration file (configClient.txt) is required by the client.

2.1.3 Usage

The SciDetect web service client can be used in the same manner as the SciDetect local (with the same parameters), Please see section 1.2.
CHAPTER 2. SCIDETECT WEB SERVICE

CONFIGURATION

A configuration file (configClient.txt and configServer.txt) is included with both the Server and the Client.

2.2.1 Client Configuration

The configuration file for the client (configClient.txt) should be found in the same directory with the SciDetectClient.jar and it contains:

Endpoint service

```
# Endpoint service location
Endpoint_Service
  http://lexicometrie.imag.fr/SciDetectServer2.2/Checker?wsdl
```

This line is used to point the client to where the SciDetectServer is located, normally it is in the form of:

```
```

Threshold configuration & Path for log files

These configuration are the same as for Scidetect Local, Please refer to section \[1.3\]

2.2.2 Server Configuration

The configuration file for the server should be found in the following directory on tomcat server along with the data directory:

```
<path_to_tomcat>/webapps/SciDetectServer2.0/WEB-INF/
```

It contains:

- path to sample folder
- Max-Min text length

And can be configured the same in section \[1.3\]
Chapter 3

Extra information
CHAPTER 3. EXTRA INFORMATION

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 used to get a more detailed look at the results.

For example: an article returned with a distance 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.

<table>
<thead>
<tr>
<th>INDEX</th>
<th>data/samples/Mathgen/INDEX-mathgen55.txt</th>
<th>0.6821885795569994</th>
</tr>
</thead>
<tbody>
<tr>
<td>INDEX</td>
<td>data/samples/Mathgen/INDEX-mathgen63.txt</td>
<td>0.6608131367167517</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Scigen/INDEX-scigen36.txt</td>
<td>0.39296257670516693</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Mathgen/INDEX-mathgen9.txt</td>
<td>0.6679829987841077</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Scigen/INDEX-scigen0.txt</td>
<td>0.3534265841094817</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Mathgen/INDEX-mathgen47.txt</td>
<td>0.660816573503142</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Scigen/INDEX-scigen52.txt</td>
<td>0.3808927385660057</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Mathgen/INDEX-mathgen71.txt</td>
<td>0.689795647595604</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Scigen/INDEX-scigen28.txt</td>
<td>0.38955875898790254</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Scigen/INDEX-scigen60.txt</td>
<td>0.39994884474379633</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Mathgen/INDEX-mathgen39.txt</td>
<td>0.6868800914402744</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Physgen/INDEX-physgen81.txt</td>
<td>0.5303053819516341</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Propgen/INDEX-17-html.txt</td>
<td>0.7981193467108959</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Physgen/INDEX-physgen65.txt</td>
<td>0.510647010647008</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Propgen/INDEX-53-html.txt</td>
<td>0.788066966830156</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Physgen/INDEX-physgen5.txt</td>
<td>0.5160079114941755</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Physgen/INDEX-physgen73.txt</td>
<td>0.5115960731657623</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Physgen/INDEX-physgen49.txt</td>
<td>0.5055891144600811</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Propgen/INDEX-86-html.txt</td>
<td>0.7643301386966208</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Physgen/INDEX-physgen96.txt</td>
<td>0.5069873754844876</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Propgen/INDEX-45-html.txt</td>
<td>0.791353315721742</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Scigen/INDEX-scigen21.txt</td>
<td>0.3848492600335824</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Mathgen/INDEX-mathgen78.txt</td>
<td>0.6692076400040969</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Propgen/INDEX-0-html.txt</td>
<td>0.787686114719592</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Mathgen/INDEX-mathgen16.txt</td>
<td>0.68202115990133</td>
</tr>
<tr>
<td>INDEX</td>
<td>data/samples/Physgen/INDEX-physgen10.txt</td>
<td>0.5261174636174665</td>
</tr>
</tbody>
</table>
Tuning/Setting Thresholds

Thresholds for the current known generators have been empirically set according to tests presented in this section. These tests involves the computation of the intertextual distance presented in [1].

For each generator (Scigen, scigen-physics, Mathgen and propgen) a set of 400 texts is used (i.e: 1600 texts for the whole). For each text the distance to its nearest neighbour in the sample set is computed. The sample is composed of an extra 100 texts per generator (i.e: 400 additional texts). The nearest neighbour is always of the same nature than the tested text and columns 1-2-3-4 of Table 3.1 show statistical information about the observed distances.

A set of 8200 genuine papers is also used. For each genuine text the distance to its nearest fake in the sample set is computed. The sample still being composed of the same 400 texts (100 per generator). For each of the 8200 genuine papers, the nearest fake neighbour is in one of the generated sample group.

The first 2 rows of Table 3.1 show that, for a genuine paper, the minimal distance to the nearest fake is always greater than the maximal distance of the nearest neighbour of a fake.

Table 3.1 Mean, min-max distances between papers and theirs nearest neighbour, along with standard deviation and median.

<table>
<thead>
<tr>
<th></th>
<th>Scigen</th>
<th>scigen-physics</th>
<th>Mathgen</th>
<th>Propgen</th>
<th>Genuine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min distance to NN</td>
<td>0.30</td>
<td>0.31</td>
<td>0.19</td>
<td>0.11</td>
<td>0.52</td>
</tr>
<tr>
<td>Max distance to NN</td>
<td>0.40</td>
<td>0.39</td>
<td>0.28</td>
<td>0.22</td>
<td>0.99</td>
</tr>
<tr>
<td>Mean distance to NN</td>
<td>0.35</td>
<td>0.35</td>
<td>0.22</td>
<td>0.14</td>
<td>0.69</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.014</td>
<td>0.012</td>
<td>0.014</td>
<td>0.015</td>
<td>0.117</td>
</tr>
<tr>
<td>Median</td>
<td>0.35</td>
<td>0.35</td>
<td>0.22</td>
<td>0.14</td>
<td>0.64</td>
</tr>
</tbody>
</table>


scigen-physics ([https://bitbucket.org/birkenfeld/scigen-physics](https://bitbucket.org/birkenfeld/scigen-physics)) (dir data/samples/Physgen) The graph 3.2 shows the observed distribution for texts having a scigen-physics text as nearest fake neighbour.

Mathgen ([http://thatsmathematics.com/mathgen/](http://thatsmathematics.com/mathgen/)) (dir data/samples/Mathgen) The graph 3.3 shows the observed distribution for texts having a mathgen text as nearest fake neighbour.

propgen ([http://www.nadovich.com/chris/randprop/](http://www.nadovich.com/chris/randprop/)) (dir data/samples/Propgen) The graph 3.4 shows the observed distribution for texts having a randprop text as nearest fake neighbour.
Figure 3.1 Distribution of distances to the Scigen nearest neighbour. In blue for a set of non-scigen paper. In red for a set of scigen papers.
Figure 3.2 Distribution of distances to the scigen-physics nearest neighbour. In blue for a set of non-scigen-physics paper. In red for a set of scigen-physics papers.
Figure 3.3 Distribution of distances to the $mathgen$ nearest neighbour. In blue for a set of non-$mathgen$ paper. In red for a set of $mathgen$ papers.
Figure 3.4 Distribution of distances to the *randprop* nearest neighbour. In blue for a set of *non-randprop* paper. In red for a set of *randprop* papers.