Pagelyzer

Installation and Configuration Manual Standalone Version

Andrés Sanoja,
Jordi Creus
LIP6 / Université Pierre et Marie Curie

Responsables WP :
Matthieu CORD/UPMC
Stéphane GANÇARSKI/UPMC

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Enviroment Verification and Configuration

The tools *pagelyzer: analyzer*, *changedetection* and *capture* are written in Ruby 1.9.1. In the other hand for the change detection process others tools are used that are written in Java, therefore this should be taken into account in the enviroment verification process. The development enviroment was Linux Ubuntu 11.40, the package description is done following its repositories, but in theory should be compatible with Debian repos.

Ruby Installation

We need to be carefull with this step because the software won't work on the 1.8.x versions of Ruby.

```
sudo apt-get install ruby1.9.1-full
```

After that we should check that both, ruby and *rubygems*, are been properly installed.

```
$ ruby -v
1.9.2p290 (2011-07-09 revision 32533) [i686-linux]
```

It is enough to match the version number. Any doubts there are several tutorials to do this [1]. Now we check the *rubygems* package manager:

```
$ gem -v
1.3.7
```

Instalation of Pagelyzer 0.9

Pagelyzer is a set of components that can be used (most of them) independently, but in the case of change detection they are all used as a chain for simplicity of integration.

The software can be downloaded from:

```
http://www-poleia.lip6.fr/~sanojaa/pagelyzer_0.9-standalone.zip
```

Now we un-compress the zip file in the desired destination

The folder structure should be like the following:

- pagelyzer
- Gemfile
- bin/
  - pagelyzer_analyzer
  - pagelyzer_capture
  - pagelyzer_changedetection
- data/

    - js/
        - compress_js.rb
        - decorate.js
        - decorate_mini.js
- doc/
    - rdoc/
- ext/
    - marcalizer
        - marcalizer.jar
        - clean.sh
        - in/
        - out/
- lib/
    - DIFF.jar
    - pagelyzer_block.rb
    - pagelyzer_convex_hull.rb
    - pagelyzer_dimension.rb
    - pagelyzer_heuristic.rb
    - pagelyzer_point.rb
    - pagelyzer_separator.rb
    - pagelyzer_url_utils.rb
    - pagelyzer_util.rb
- out/

Note: *out* folder is intended to be an output folder, but it is optional. Can be overridden with parameters.

Installing Dependencies

After the language and the package manager are properly configured and installed, we may proceed to install the dependencies:

```
$ sudo apt-get install libxslt-dev libxml2-dev
$ sudo apt-get install openjdk-7-jdk
$ sudo apt-get install imagemagick
```

Note1: Installing the selenium-webdriver may cause some warnings in text encoding that should be fine, in almost all the cases.

Note 2: The java installation is a reference to remember that it should be present.

Note 3: ImageMagick 6 is mandatory, it is needed for thumb-nailing, cropping web page visible

area and get a homogenous image format (RGB 8-bit color) that can be processed by Marcalizer tool. This thumbs area is useful for integrating with other tools and for future optimization of change detection process .

We need to install also some ruby libraries needed by the software. This step can be done simple using *Bundler* gem. To install it:

```
$ sudo gem install bundler
```

Get into the project folder and type:

```
$ bundle
```

When finished we will have all dependencies installed.

Command-line Parameters

pagelyzer:

```
USAGE: ./pagelyzer [--help|--version] [<command> <command_options>]
```

The available commands are:

- capture
- analyzer
- changedetection

Capture:

```
USAGE: ./pagelyzer capture --url=URL [--output-folder=FOLDER] [--
browser=BROWSER_CODE] [--thumbnail] [--help]
```

Analyzer:

```
USAGE: ./pagelyzer analyzer --decorated-file=FILE [--output-file=FILE] [--
pdoc=(0..10)] [--version] [--help]
```

Changedetection:

```
USAGE: ./pagelyzer changedetection --url1=URL --url2=URL [--doc=(1..10)]
[--output-folder=FOLDER] [--browser=BROWSER_CODE | --browser1=BROWSER_CODE
--browser2=BROWSER_CODE] [--verbose] --type=[hybrid|visual]
```

Browsers code are the same as defined in selenium. For instance:

- firefox (default)
- chrome
- iexploreproxy
- safariproxy
- opera

Remarks:

- Firefox driver is the default to selenium. For installing other browsers can reference to [2], e.g. to run pagelyzer on your chrome/chromium instance, you should install the ChromeDriver before:
    - Download the appropriate version from http://code.google.com/p/chromedriver/downloads/list
    - Unzip it and copy it to a visible folder, e.g:

```
$ sudo cp chromedriver /usr/bin/
```

- For command-line parameters is better to escape them, e.g:

```
pagelyzer analyzer --decorated-file=/my/path with/spaces -- only processes /my/path !
Pagelyzer analyzer --decorated-file=/my/path\ with/spaces -- results in correct behaviour
```

- If no Degree of Coherence is given, a default of doc=6 will be chosen.
- The URL's should include the http schema

```
--url=http://www.host.com ---it is ok!
--url=host.com ---won't work!
```

External References:

[1] http://answers.oreilly.com/topic/2845-installing-ruby-1-9-on-a-debian-or-ubuntu-system/
[2] http://code.google.com/p/selenium/wiki/FrequentlyAskedQuestions#Q:_Which_browsers_does_WebDriver_support?