Pagealyzer

Installation and Configuration Manual

Andrés Sanoja, 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 pageanalyzer.rb, change_detection.rb and capture.rb 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 debien 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
```

Installing Dependencies

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

```
sudo gem install --version '= 0.8.6' hpricot
sudo apt-get install libxslt-dev libxml2-dev
sudo gem install --version '= 1.5.5' nokogiri
sudo gem install --version '= 2.0.3' sanitize
sudo gem install --version  '= 2.25.0' selenium-webdriver
sudo apt-get install openjdk-6-jdk
sudo apt-get install imagemagick
```

Note1: Installing the selenium-webdriver can 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 optional, only needed for thumbnailing and cropping web page viewport. This thumbs area usefull for integrating with other tools and for future optimization of change detection process . If you don't want to use it ignoring parameter “--thumb” should do the trick

Instalation of Pagealyzer 0.9

Pagealyzer 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:

https://github.com/openplanets/pagelyzer/downloads

It is enough to unzip the compressed file into the desired destination and after all dependencies are met we are ready to go.

The folder structure is the following:

- capture.rb
- change_detection.rb
- js
  - js/dump.js
  - js/jquery.min.js
  - js/jquery.unique-element-id.js
- lib
  - lib/block.rb
  - lib/convex_hull.rb
  - lib/dimension.rb
  - lib/heuristic.rb
  - lib/point.rb
  - lib/separator.rb
  - lib/url_utils.rb
  - lib/util.rb
- marcalizer
  - marcalizer/in
  - marcalizer/marcalizer.jar
  - marcalizer/out
- out
- pageanalyzer.rb
- vidiff
  - vidiff/DIFF.jar
  - vidiff/lib

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

Command-line Parameters

ATTENTION: The general script is Change Detection. The other scripts are called from this script. If you don't want to use other scripts independently, it is enough to call change_detection

1 - Capture:

```
USAGE: ruby capture.rb --url=URL [--js-files-url=BASE_URL] [--output-
folder=FOLDER] [--browser=BROWSER_CODE]
```

This tool aims to produce an HTML document with the visual cues integrated, called Decorated HTML. This allows to save the state of a browser at the moment of capture

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

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

The JS Files must be available for the chosen browser (accesible via http), i.e.: http://myserver/path/myfolder

Inside 'myfolder' should be all the js files provided. Do not include the last slash '/'

2 - PageAnalyzer:

```
USAGE: ruby pageanalyzer.rb --decorated-file=FILE [--output-file=FILE] [--
pdoc=(0..10)] [--version] [--help]
```

For commandline parameters is better to escape them, e.g:

```
pageanalyzer.rb --input-file=/my/path with/spaces -- only processes /my/path !
pageanalyzer.rb --input-file=/my/path\ with/spaces -- results in correct behaviour
```

3 - Change Detection:

ATTENTION: This is the general script which uses the previous tools.

```
USAGE: ruby change_detection.rb --url1=URL --url2=URL [--js-files-
url=BASE_URL] [doc=(1..10)] [--output-folder=FOLDER] [--
browser=BROWSER_CODE] [--verbose]
```

Either the browser code and the js-files-url have the same restriction as the previous tools, that is normal because the change detection tool uses them.

If no Degre of Coherence is given, a default of doc=6 will be chosen.

[1] http://answers.oreilly.com/topic/2845-installing-ruby-1-9-on-a-debian-or-ubuntu-system/