Web2Eventus v.2.0
User's guide

Osorio, Javier
javier.osoriozago@gmail.com

Reyes, Alejandro
alejandroe4@hotmail.com

May, 2014

Copyright

Terms of Use

How To Cite This Program

Please cite program as:

Javier Osorio and Alejandro Reyes. 2014. Web2Eventus. Version 2.0.

BibTex citation:

```
@misc{Osorio2014,
address = {Puebla, Puebla},
author = {Osorio, Javier and Reyes, Alejandro},
title = {{Web2Eventus v.2.0}},
url = {http://www.javierosorio.net/#!software/cqbi},
year = {2014}
}
```

Report bugs

Please report any bugs to: javier.osoriozago@gmail.com

Updated copies of the Web2Eventus program and manual can be found at http://www.javierosorio.net/#!software/cqbi

1 Introduction

This manual shows how to use the program Web2Eventus. The program is used for processing previously downloaded news reports and inserting the content into a single document to be used as *text corpus* in Eventus ID (Osorio and Reyes, 2014*a*). In particular, Web2Eventus processes each news report by extracting the content, breaking it down by paragraphs, attaching a paragraph counter, formatting the information and storing all the text into a file readable in Eventus ID.

Prior to using this program, users must have a set of individual files containing news reports. Users are suggested to use the software Web Text Downloader and follow the procedure indicated in its manual for extracting news reports from the web.

2 Required Software

1. Web2Eventus runs on Windows command line interface.

2. Web2Eventus also requires having Perl 5, or any later version, already installed.

   (a) We recommend using Strawberry Perl, a canned Perl environment already containing the set of tools and libraries necessary for using Web2Eventus. Strawberry Perl is available for download at `http://strawberryperl.com/`.

   (b) Users can find more Perl resources and documentation at `http://www.perl.org/`.

3. Users need a text editor for programming languages. We recommend using Notepad++, a source code editor that supports Perl scripts. Notepad++ can be downloaded at `http://notepad-plus-plus.org/`.

3 Required Files

To process web reports into Eventus ID readable format, it is necessary to have the following files in the same directory:

1. The program file web2eventus.pl

2. A folder containing the individual files of news reports previously extracted from the web. The example presented in this manual considers news reports stored in Hyper Text Markup Language (*.html*) format. But the software is also capable of processing plain text files (*.txt*).

4 News Report File Naming Convention

In order to processing news report files using Web2Eventus, individual file names must follow a specific set or naming rules. Complying with this nomenclature is key for adequately managing files and formatting the corpus of text used in EVENTUS ID. The nomenclature consists of four main elements:

- Date
- Counter
- Source
- Extension

4.1 File names

File names should be structured in the following way: yyyymmddccc_SRC.ext, where:

- yyyy is a four digit number representing the year (e.g. 2009).
- mm is a two digit number representing the month (e.g. 02 for February).
- dd is a two digit number representing the day (e.g. 17).
- ccc is a three digit number counting the number of reports issued by the same source in a given day. The three digits allow the counter to range from 001 up to 100.
- SRC is a short acronym indicating the name of the information source.
- ext is the extension indicating the format of the file. WEB2EVENTUS can process files in *.html* or *.txt* format.

There should be no duplicate file names. The counter in the nomenclature (ccc) is an effective way of assigning unique file names even in the context of a large number of news reports. For convenience, the counter considered in this example is set to three digits, but users can assign as many digits to the counter as they consider necessary.

4.2 File Name Example

For example, consider a set of three on-line press releases issued by the Mexican Army (Secretaría de la Defensa Nacional, SEDENA). The first two news reports were issued on August 23, 2009 and the third report on October 17, 2010. According to the nomenclature, the files should have the following names:

```
20090823001_SEDENA.html

20090823002_SEDENA.html

20101017001_SEDENA.html
```

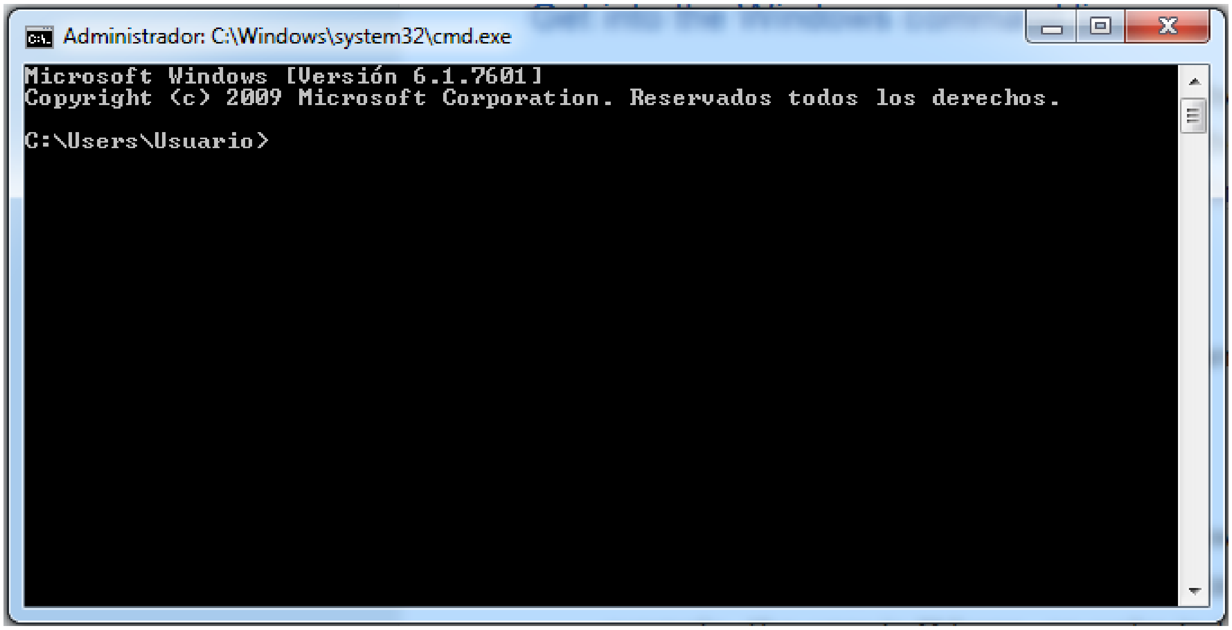
The demonstration file of WEB2EVENTUS contains a set of *.html* files containing the web content of real press releases issued by the Mexican Army. These files were extracted using the program WEB TEXT DOWNLOADER (Osorio and Reyes, 2014*b*). The folder SEDENA_permission in the demonstration .zip file contains the official letter authorizing the use of these press releases and the required citations of the documents used in the corpus of text.

5 Using Web2Eventus

To create a corpus of text for Eventus ID using previously downloaded news reports, users must follow the following steps:

Step 1: Open Windows command line interface

To launch the command line interface in Windows, click on the Start button in the lower left of the screen, then type the command `cmd` in the search bar and press enter. This will open the command line interface as presented in Figure 1.



```
Microsoft Windows [Versión 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Reservados todos los derechos.

C:\Users\Usuario>
```

Figure 1: Command line interface.

Step 2: Navigate the command line

Navigate the command line interface in order to find the directory containing the program file web2eventus.pl and the folder containing the individual news reports. Assume that the collection of news reports is hosted in a folder named `News_folder`.

- Use the command `cd` for moving downwards the directory line and `cd..` for moving upwards.
- Typing `cd /?` will present a quick tutorial on how to use the `cd` command.

Step 3: Run Web2Eventus

In order to run Web2Eventus follow the next steps:

Step 3.1: In the command line, type: `perl web2eventus.pl` and then press the Enter key. The command `perl` calls the PERL environment whereas command `web2eventus.pl` launches the program WEB2EVENTUS.

Step 3.2: Then, the program will ask the user to enter the route to the folder containing the files to be processed and then press Enter. The directory route can be entered in two ways:

- Absolute route: `C:/directory/web_DEMO/*.html`
- Relative route: `web_DEMO/*.html`

These command lines equally indicate WEB2EVENTUS to process all *.html* files from folder `web_DEMO`. The absolute route indicates the specific location of the directory in the computer station used for this task. In contrast, the relative route indicates the software to process the files contained in a sub-folder already hosted in the working directory being used.

WEB2EVENTUS can also process news report files in plain text format. To do that, users just need to replace the extension *.html* with *.txt* in the command line.

If users want to process only one particular file, they can do it by indicating the specific file name as `web_DEMO/file_name.html`.

Step 3.3: Finally, the program asks the user to enter the name of the output file. For example, type the name corpus_DEMO.txt to create a file containing the corpus of text to be used by EVENTUS ID, and then press enter. Do not forget to indicate the *.txt* extension in the file name.

6 Output File: Text Corpus For Eventus ID

WEB2EVENTUS generates an output file (corpus.txt) ready to be used in EVENTUS ID as text corpus. Each line in the output contains the information of each individual paragraph from each news report. The corpus of text presents the information according to the following structure: `date FileName_P1_P2 | Text`, where:

- `date`: indicates the date of the news report by extracting this information from the file name entered by the user (see section 4.1.)
- `FileName`: indicates the file name of each news report (see section 4.1.).
- `P1`: WEB2EEVENTUS breaks each news report into paragraphs. `P1` is a progressive number indicating the local paragraph counter for each news report. This is useful for identifying a specific paragraph within each news report.
- `P2`: indicates the global paragraph counter for all news reports comprised in the corpus of text. This counter is useful for quickly locating a specific paragraph in the corpus.

- `|`: the vertical bar symbol (`|`) is a marker indicating the beginning of the text extracted from each paragraph.

- `Text`: the content of each paragraph from each news report is stored in a single line. The length of each line depends on the number of characters of each paragraph.

The corpus of text generated by Web2Eventus should look like this:

```
20130808 20130808001_SRC1_P0_P1 | Lorem ipsum dolor sit amet...
20130808 20130808001_SRC1_P1_P2 | Praesent at sem ac enim interdum...
20130808 20130808001_SRC1_P2_P3 | Donec sed mattis orci.  Praesent...
20130808 20130808001_SRC2_P0_P4 | Donec velit justo, varius non ...
20130808 20130808001_SRC2_P1_P5 | Praesent quis felis ipsum...
20130808 20130808001_SRC2_P2_P6 | Nunc blandit vitae purus vitae...
20130808 20130808001_SRC2_P3_P7 | Quisque quis lorem sed nunc egestas...
20130921 20130921001_SRC1_P0_P8 | Sed ornare, nisi vitae lacinia...
20130921 20130921001_SRC1_P1_P9 | Nulla vel condimentum sem, nec...
20130921 20130921002_SRC1_P0_P10 | Phasellus porta ipsum eu leo...
20130921 20130921002_SRC1_P1_P11 | Etiam porttitor vitae odio ut...
20130921 20130921002_SRC1_P3_P12 | Donec cursus metus vel neque...
```

Use Notepad++ to view the content of the corpus. In this example, the first three lines represent the paragraphs extracted from a news report issued on October 8, 2013 by source `SRC1`. Lines four to seven show the content from a report issued the same day by a different source, `SRC2`. Notice how the local paragraph counter indicates the number of paragraphs in each news report while the global paragraph counter indicates the total number of paragraphs in the corpus of text. Lines eight to twelve present the information from two news reports issued by the same source (`SRC1`) on the same day (September 21, 2013). Notice that the file nomenclature (see section 4.1) helps to assign unique names for each file (`20130921001` and `20130921002`, respectively), which also helps to have unique paragraph identifiers when adding the local and global paragraph counters.

Besides reformatting the text by paragraph, Web2Eeventus identifies phonetic diacritic and emphatic marks on vowels, also known as acute accents (e.g. á, é, í, ó and ú) and substitutes them from their corresponding vowels without accents. Although accent marks constitute an important grammatical rule in Spanish, this facilitates the coding process of Eventus ID. The reformatting process also cleans up the text by eliminating some punctuation signs (e.g. “ ” : ; ? ! - _). It is important to notice that Web2Eeventus also creates blank spaces to the sides of the dot “.” and comma “,” punctuation signs, thus reformatting them as “·.·” and “·,·”, where the symbol · represents a blank space. This characteristic of the corpus of text is crucial for Eventus ID to identify the geographic location of coded events (for details about the event location process see Osorio and Reyes (2014*a*).).

References

Osorio, Javier and Alejandro Reyes. 2014*a*. "Eventus ID. Supervised Event Coding From Text Written in Spanish.".
URL: *http://www.javierosorio.net/#!software/cqbi*

Osorio, Javier and Alejandro Reyes. 2014*b*. "Web Text Downloader.".
URL: *http://www.javierosorio.net/#!software/cqbi*