ICE Web Manual (v. 2.0)

1 Introduction

ICE Web is a tool designed to facilitate finding and downloading web pages for creating new ICE corpora, or updating existing ones.

It allows you to specify different search engines to run queries in your default browser using seed terms for a specified region (generally a continent), country, and category, and then collect and download these URLs.

All existing ICE categories are already pre-defined within the tool and can easily be selected, but new categories, as well as regions, can also be created if necessary.

Once web pages have been downloaded to folders that will automatically be created based on your selections, they can also automatically be converted to text format, or a dedicated form of XML, post-edited to remove unwanted materials, tagged, as well as analysed using the built-in concordancer or n-gram analysis tool.

Almost all of the actions are carried out via the relevant buttons on the different tabs for handling 'URLs', 'Retrieval' of web pages, 'Data' viewing/editing/conversion, the 'Concordancer', or 'N-gram(s)' analysis.

2 Handling Locations

2.1 Adding New Regions or Categories

Although the pre-defined regions (continents) or categories should be sufficient for most users, it is possible to create a new 'region' definition, perhaps for grouping data for different countries together, rather than keeping them associated with a particular continent.

To do so, you only need to type its name in the dropdown box for 'Region', select 'Locations→Add→Region' from the menu, and it will automatically be added to the dropdown list, as well as saved to the configuration file for the tool upon exiting the program, so that the option will be available again the next time ICE Web is run.

The same mechanism can be used for creating a new category, selecting 'Locations→Add→Category'.

3 Editing/Creating URLs

3.1 Defining URLs for Retrieval

To create or edit a file containing relevant URLs for retrieval, you first need to make a selection from the dropdown lists for 'Region', 'Country', and 'Category'. Choosing 'Locations→ Edit/add URLs' will then create a new file named 'url.txt' in the relevant folder for the category, creating folders for regions and countries as per selection automatically if they don't already exist.

Once the file has been created, it will automatically be opened in the editor window on the 'URLs' tab where all the editing operations are performed.

To fill the file with URLs, you can then select a search engine and provide potential seed terms. Clicking the 'Run query' button will then open the default browser with the search engine URL and seed terms already filled in in the search box. In addition, search restrictions setting the language to English and to the relevant country's domain will also be filled in, so that you can simply run the search and identify relevant pages.

Once suitable web pages have been found, you only have to copy the URL from the browser's address bar and paste it into the URL editor window, one URL per line.

Existing URL files can be opened, saved, and closed directly from the 'URLs' tab.

When adding files to be downloaded if download data already exists for a category, previously downloaded URLs should normally be commented out by prefixing them with a hash mark (#), unless you deliberately want to download a new (perhaps updated) version.

4 Downloading Web Pages

Downloading pages listed in a URL file is a simple process. Once the relevant location and category information has been selected and URL file created, all you need to do is to select the 'Retrieval' tab and click on the 'Get web pages' button to start the download process.

ICE Web will then attempt to download the HTML pages listed in the URL file one by one, posting messages in the editor window on the tab indicating the beginning of each download and, if successful, a relevant message. If the URL cannot be retrieved for some reason, a relevant message to that effect will also be displayed.

The HTML code for each successfully retrieved page will be stored in a sub-folder named 'html' of the category folder. If possible, a modified version of the original title will be used as the file name, and, to be able to distinguish between files with the same title, a running number is prepended to each file.

In addition, ICE Web will also create, or append to, a csv file in the category folder, which records information about the 'Title' (if present) of the original file, as well as the URL, the local file name, and the download time, in order to keep a record of all downloads. This also makes it possible to easily access the original URL again, for instance in order to later check and see if the page has changed or been removed.

The csv file can be viewed using a text editor or a spreadsheet application, such as MS Excel or Open Office Calc.

5 Working with Downloaded Data

Once data has been downloaded to a category folder, there are different ways of editing or processing it from the 'Data' tab.

The downloaded HTML files can be converted to raw text, a dedicated form of XML that allows further annotation in other tools for more detailed text analysis, or text files, once created from HTML, can be tagged using the built-in tagger.

To create a derived file type from HTML or raw/text, you simply need to click one of the three buttons 'HTML→txt', 'HTML→XML', or 'Tag text'.

Viewing/editing a specific type of file, if it already exists, can be done by clicking on the relevant button marked by a folder icon at the top and a file type description at the bottom.

5.1 Converting to Different File Types/Formats

The most basic conversion operation most you will want to carry out in order to create a corpus is to convert the HTML to raw text. Here, ICE Web not only extracts the text from the HTML tags, but also makes an effort to clean it up and remove irrelevant material, such as purely navigational elements, i.e. structures that represent menus, or other lists of links or images, for instance those used for advertising, but which represent no real content.

Conversion to XML will do something similar, but also try to preserve structural information, such as the distinction between headings, paragraphs, and lists, by enclosing the relevant textual material into different types of unit tags. This can provide a basis for various

types of filtering in later analysis stages that would not be possible to carry out on raw text, e.g. analysing headings only, etc.

Raw text may also be tagged using the built-in tagger, which is based on a probabilistic tagging module written by other authors, so that I would definitely recommend post-processing all tagging output manually in order to remove potential errors such probabilistic systems invariably introduce!

For future versions, I will try to design some automatic post-processing templates to fix the most common errors.

6 Viewing/Editing Files

When a folder containing one of the file types (HTML, text, XML, or tagged text) has been chosen, the files of this type will be listed in the window on the 'Data' tab where they can be selected individually.

Once a file has been selected, it can either be edited in the internal editor by clicking on the 'Edit file' button, or the corresponding HTML file viewed in the browser associated with it on your system by clicking on 'View in browser'. The latter is particularly useful in post-editing the output of the conversion processes to text or XML to establish whether these processes may either inadvertently have removed or misclassified specific page elements, or to clean the results even further. For editing tagged output, the viewing option should be largely irrelevant.

7 Analysis Options

This section describes the different analysis options that ICE Web provides for analysing raw text, XML files, or tagged text. The tabs for these analysis options are 'Concordancer' and 'N-grams'.

Both of the analysis actions are triggered via the 'Run' button or pressing 'Ctrl+r'. The action performed depends on which analysis tab has been selected, as well as the 'Target folder' option selected from the dropdown list next to the 'Run' button.

7.1 Concordancing

Concordancing in ICE Web is relatively similar to working with other concordancers, but with a number of important differences.

For one thing, the ICE Web concordancer is line-based, and it is also possible to search for two separate terms, either on the same line, or with the second term preceding or following a specified distance away from the first.

In addition, unlike in other concordancers, each search term actually represents a full Perl regular expression that is validated before the search is started, also providing feedback on any potential errors. In the simplest case, this is only a case-sensitive word form.

ICE Web expects at least one search term to be present in the box next to 'Term 1'.

The first input box for search terms also has an autocomplete feature, based on customisable options specified in the ICE Web configuration file, but of course auto-completed content can still be modified if necessary.

The relative position for the 2nd term is by default set to 0, and needs to be adjusted in order to specify whether the search term should occur prior to (negative number) or following the 1st term (positive number). When looking for items on the same line after choosing a relative position before, it is important to remember to reset this!

The display context for the 1st term can be adjusted to show a number of preceding and/or following lines. When a search is run, these numbers will automatically be adjusted if they don't include the relative position of the 2nd term.

Once a search is complete, the number of 'Hits' is displayed, together with its 'Document frequency', i.e. dispersion across the corpus files.

The concordance output itself contains the hits and their respective context, but is also hyperlinked via the filename & line number, so that any hit can immediately be viewed in its full context in the built-in editor. In contrast to other concordancers, though, the file that contains the hit can also be edited in order to correct annotation errors or to add markup in case one wants to categorise specific hits further.

ICE Web provides no direct way of saving the list of concordance results, but these can be selected by pressing 'Ctrl+a', and then copied ('Ctrl+c') and pasted into an external editor in order to save them.

7.2 N-gram Analysis

N-gram analysis in ICE Web can be carried out by selecting the 'N-grams' tab and adjusting the parameters before running the analysis. A setting of 1 essentially produces a basic word list,

while numbers larger than 1 produce proper n-grams. Prior to creating the lists, all texts are cleaned of any XML elements, if present, too, as well as punctuation, so that the n-gram sequences are not distorted, and only report proper sequences of words.

As is customary for such lists, they can also be sorted by various options, including reverse-sorting, but the default setting is by descending frequency.

In addition, lists can also be filtered by using regular expressions before the output is created, which, however, does not affect the frequency counts. In this way, it e.g. becomes possible to display only proper names or n-grams containing particular strings.

The output consists of the word/n-gram itself, the raw frequency (Freq.), the relative frequency (RFreq), and the document frequency (DFreq.).

The n-gram string is hyperlinked to prime and run the concordancer with the n-gram string for further inspection of the larger context of the n-gram in context. When the concordance is triggered from the n-gram analyser, punctuation is also automatically interpolated.

As with the concordance, it is possible to copy the n-gram window contents via ‘Ctrl+a’ & ‘Ctrl+c’ and then either pasting them into an editor or importing them into a spreadsheet application for further analysis.

A total count of all n-gram strings identified is also reported.

8 Editing the Configuration File

It is possible to configure a number of different options using the ICE Web configuration file. The file can be opened for editing by selecting ‘Edit→Configuration’ on the menu bar. When editing the options, it’s generally best to comment out the original ones, so that it will be possible to revert to them in case you have inadvertently specified an invalid option, such as an incorrect file path or a non-existent option, which may, in the worst case, cause ICE Web to fail loading even though I have made every effort to prevent this from happening.

Currently, it is possible to change/set options and/or defaults for regions, categories, countries, the preferred search engine, the folder containing your data (pre-set to the data folder in the program folder), which type of data should be used for the analysis options, which tab to activate at startup, as well as a number of pre-defined search terms for the Term1 box.

9 Keyboard Shortcuts

There are number of shortcuts defined for different purposes, but predominantly for the built-in editor.

9.1 General Shortcuts

Ctrl+a	select all
Ctrl+c	copy
Ctrl+x	cut
Ctrl+v	paste

9.2 Editor Shortcuts

Ctrl+s	save
Ctrl+w	close editor window/file
Ctrl+f	find
Ctrl+h	replace
Shift+F5	insert date and time

9.3 Analysis Shortcuts

Ctrl+r	run analysis (based on selected analysis tab)