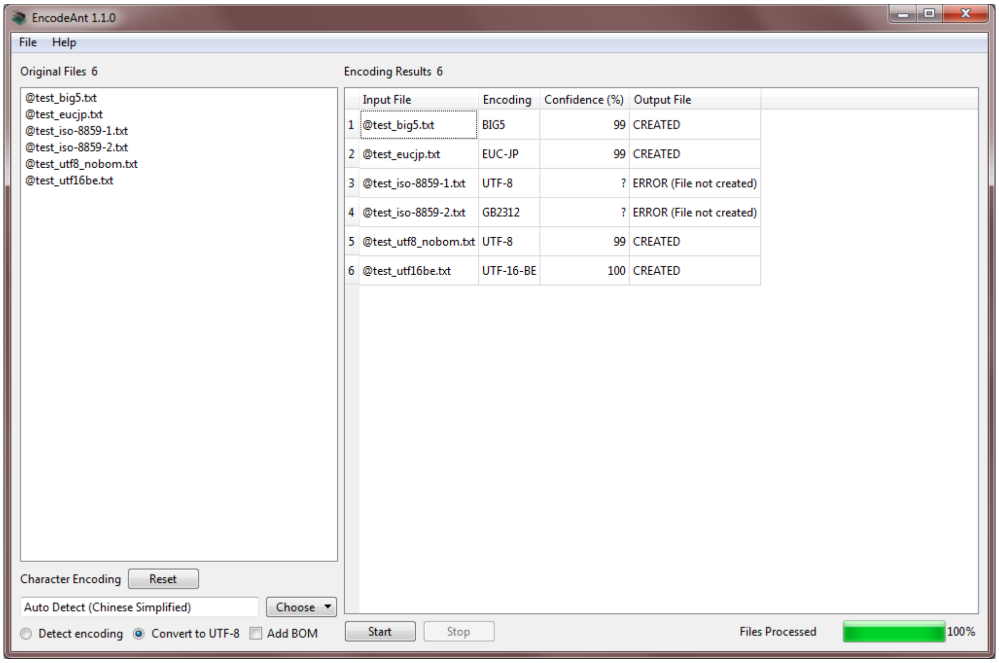
EncodeAnt (Windows)

Build 1.2.1 (Released March 23, 2017)

Laurence Anthony, Ph.D.
Center for English Language Education in Science and Engineering, School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan.

Help file version: 002.



Original Files 6

@test_big5.txt
@test_eucjp.txt
@test_iso-8859-1.txt
@test_iso-8859-2.txt
@test_utf8_nobom.txt
@test_utf16be.txt

Encoding Results 6

	Input File	Encoding	Confidence (%)	Output File
1	@test_big5.txt	BIG5	99	CREATED
2	@test_eucjp.txt	EUC-JP	99	CREATED
3	@test_iso-8859-1.txt	UTF-8	?	ERROR (File not created)
4	@test_iso-8859-2.txt	GB2312	?	ERROR (File not created)
5	@test_utf8_nobom.txt	UTF-8	99	CREATED
6	@test_utf16be.txt	UTF-16-BE	100	CREATED

Character Encoding: Auto Detect (Chinese Simplified)

○ Detect encoding ● Convert to UTF-8 ☐ Add BOM

Files Processed 100%

Introduction

EncodeAnt is a freeware character encoding detection and conversion tool. *EncodeAnt* takes an input list of text files (e.g. .txt) and attempts to auto-detect the character encoding that the files use. The character encoding can also be set manually. *EncodeAnt* also has an option to auto-convert the character encoding of the files to UTF-8, which is a standard used in most corpus research. The converted files are saved in a separate folder leaving the original files untouched.

EncodeAnt runs on any computer running Microsoft Windows (tested on Win 7), Macintosh OS X (tested on OS X 10.9 Mavericks), and Linux (tested on Linux Mint 17) computers. It is developed in Python and Qt using the *PyInstaller* compiler to generate executables for the different operating systems.

Getting Started (No installation necessary)

Windows

On Windows systems, simply double click the *EncodeAnt* icon to launch the program.

Macintosh OS X

On Macintosh systems, simply double click the *EncodeAnt* zip file. The zip file will unzip the *EncodeAnt* application. Then, you can drag the *EncodeAnt* application to your application folder, your desktop, or anywhere else you like. Throw away the zip file when you are finished.

Linux

On Linux systems, set the permissions to run the executable, then double click the *EncodeAnt* icon to launch the program.

Step 1: Select the files you want to analyze. You can do this in three ways:
a) Click on the File->Open File(s) menu option and select the files you want to segment;
b) Click on the File->Open Dir menu option and select a directory of files you want to segment;
c) Drag and drop files directly onto the *EncodeAnt* application.
Note 1: The number of selected files is shown next to the "Original Files" label.
Note 2: If you click on the File->Close Files menu option, the files will removed from the list.

Step 2: Choose one of the character encoding options. If you choose one of the "auto detect" options, *EncodeAnt* will attempt to guess the encodings of the files. If you choose one of the "common" or "all" options, AntConc will attempt to use this encoding choice.

Step 3: Choose the "Detect encoding" radiobox to detect the encodings of your files or the "Convert to UTF-8" radiobox to convert the encodings of your files (based on the character encoding option you chose in step 2) to the standard UTF-8 encoding.

Step 4: Click the "Add BOM" checkbox if you want to add a UTF-8 BOM (Byte Order Mark) to the beginning of your converted files from step 3. This is generally not recommended but it is useful on Microsoft Windows systems that use the BOM to open the file correctly in programs like Notepad.

Step 5: Click "Start" to begin the detection/conversion process.

Additional Features

The output display can be selected, copied, and pasted as is standard on the operating system:

Windows:	CTRL-A ⇨ Select All	CTRL-C ⇨ Copy	CTRL-V ⇨ Paste
Macintosh:	CMD-A ⇨ Select All	CMD -C ⇨ Copy	CMD -V ⇨ Paste

NOTES

General points

- If a "common" or "all" encoding option is chosen, the results will report a confidence of the encoding as 100% (even in cases when the actual encoding of the file is different). If an "auto-detect" option is chosen, the confidence will be reported on a scale of 0% to 100% or given a value of "?" when the confidence cannot be determined. (See the notes on detection methods below).
- If the "Convert to UTF-8" option is selected, the converted files will be saved where the original file was stored under a sub-folder called "utf8".
- The detection/conversion process can be stopped at any time by clicking the "Stop" button.

Detection methods used by *EncodeAnt*

- The first method attempts to read BOMs at the beginning of files. These BOM unambiguously decide the encoding.
- The second method uses Chardet (https://pypi.python.org/pypi/chardet) to guess the encoding. This tool produces confidence measures for the guess, which are reported in the results table.
- The third method uses Chared (https://code.google.com/p/chared/). This relies on knowing the target (human) language of text in the file. If the first method is unsuccessful or the confidence level of the second method is too low, this third approach is adopted.

COMMENTS/SUGGESTIONS/BUG FIXES

All new editions and bug fixes are listed in the revision history below. However, if you find a bug in the program, or have any suggestions for improving the program, please let me know and I will try to address the issues in a future version.

This software is available as 'freeware' (see Legal Matter below), but it is important for my funding to hear about any successes that people have with the software. Therefore, if you find the software useful, please send me an e-mail briefly describing how it is being used.

CITING/REFERENCING *EncodeAnt*

Use the following method to cite/reference *EncodeAnt* according to the APA style guide:

Anthony, L. (YEAR OF RELEASE). *EncodeAnt* (Version VERSION NUMBER) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

For example if you download *EncodeAnt 1.1.0*, which was released in 2014, you would cite/reference it as follows:
Anthony, L. (2014). *EncodeAnt* (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http:// www.laurenceanthony.net/

Note that the APA instructions are not entirely clear about citing software, and it is debatable whether or not the "Available from ..." statement is needed. See here for more details:
http://owl.english.purdue.edu/owl/resource/560/10/

LICENSE for EncodeAnt

EncodeAnt 1.0 and any minor updates issued by AntLab Solutions (collectively 'the Software')

TERMS GOVERNING THE USE OF THE SOFTWARE
The Software is protected by copyright and must not be used, displayed, modified, adapted, distributed, transmitted, transferred or published or otherwise reproduced in any form by any means other than strictly in accordance with the terms set out below. By installing the Software, you agree to be bound by the terms of the license. This EncodeAnt License (“License”) is made between AntLab Solutions, Tokyo, Japan as licensor, and you, as licensee, as of the date of your use of the Software. The Software is in use on a computer when it is loaded into the RAM or installed into the permanent memory of that computer, e.g., a hard disk or other storage device.

1. License Material
These terms govern your use of the Software but not including subsequent versions (e.g. EncodeAnt 2.0').

2. License Grant
AntLab Solutions grants to you a personal non-exclusive non-transferable license ('the License') to use the Software in the following specific contexts.

a) Non-Commercial (Freeware) Use:
You may use the software for non-profit purposes on more than one computer or on a network so long as you are the sole user of the Software. (A “network” is any combination of two or more computers that are electronically linked and capable of sharing the use of a single software program.) You are not permitted to sell, lease, distribute, transfer, sublicense, or otherwise dispose of the Software, in whole or in part, for any form of actual or potential commercial gain or consideration.

b) Commercial Evaluation (Trial) Use:
You may evaluate (trial) the software for commercial purposes for a period of no more than fourteen (14) days from the date of download on more than one computer or on a network so long as you are the sole user of the Software.

c) Commercial Use
When you pay the commercial license fee established by AntLab Solutions, you may use the software for non-profit or commercial purposes on more than one computer or on a network so long as you are the sole user of the Software. (A “network” is any combination of two or more computers that are electronically linked and capable of sharing the use of a single software program.) You will obtain a separate license for each additional user of the Software (whether or not such users are connected on a network). You are not permitted to sell, lease, distribute, transfer, sublicense, or otherwise dispose of the Software, in whole or in part, for any form of actual or potential commercial gain or consideration.

3. Termination
You may terminate this License at any time by uninstalling the Software and deleting it. The License will also terminate if you breach any of the terms of the License.

4. Proprietary Rights
The Software is licensed, not sold, to you. AntLab Solutions reserves all rights not expressly granted to you. Ownership of the Software and its associated proprietary rights, including but not limited to patent and patent applications, are retained by AntLab Solutions. The Software is protected by the copyright laws of Japan and by international treaties. Therefore, you must comply with such laws and treaties in your use of the Software. You agree not to remove any of AntLab Solutions' copyright, trademarks, and other proprietary notices from the Software.

5. Distribution
Except as may be expressly allowed in Section 2, or as otherwise agreed to in a written agreement signed by both you and AntLab Solutions, you will not distribute the Software, either in whole or in part, in any form or medium.

6. Transfer and Use Restrictions
You may not sell, license, sub-license, lend, lease, rent, share, assign, transmit, telecommunicate, export, distribute or otherwise transfer the Software to others, except as expressly permitted in this License Agreement or in another agreement with AntLab Solutions. You may not modify, reverse engineer, decompile, decrypt, extract, or otherwise disassemble the Software.

7. Warranties
ANTLAB SOLUTIONS MAKES NO WARRANTIES WHATSOEVER REGARDING THE SOFTWARE AND IN PARTICULAR, DOES NOT WARRANT THAT THE SOFTWARE WILL FUNCTION IN ACCORDANCE WITH THE ACCOMPANYING DOCUMENTATION IN EVERY COMBINATION OF HARDWARE PLATFORM OR SOFTWARE ENVIRONMENT OR CONFIGURATION, OR BE COMPATIBLE WITH EVERY COMPUTER SYSTEM. IF THE SOFTWARE IS DEFECTIVE FOR ANY REASON, YOU WILL ASSUME THE ENTIRE COST OF ALL NECESSARY REPAIRS OR REPLACEMENTS.

8. Disclaimer
ANTLAB SOLUTIONS DOES NOT WARRANT THAT THE SOFTWARE OR SERVICE IS FREE FROM BUGS, DEFECTS, ERRORS OR OMISSIONS. THE SOFTWARE OR SERVICE IS PROVIDED ON AN “AS IS” BASIS AND ANTLAB SOLUTIONS MAKES NO OTHER WARRANTIES OR CONDITIONS, EXPRESS OR IMPLIED, WITH RESPECT TO THE

SOFTWARE INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OR CONDITIONS OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

9. Limitation of Liability
ANTLAB SOLUTIONS WILL HAVE NO LIABILITY OR OBLIGATION FOR ANY DAMAGES OR REMEDIES, INCLUDING, WITHOUT LIMITATION, THE COST OF SUBSTITUTE GOODS, LOST DATA, LOST PROFITS, LOST REVENUES OR ANY OTHER DIRECT, INDIRECT, INCIDENTAL, SPECIAL, GENERAL, PUNITIVE OR CONSEQUENTIAL DAMAGES, ARISING OUT OF THIS LICENSE OR THE USE OR INABILITY TO USE THE SOFTWARE OR SERVICE. IN NO EVENT WILL ANTLAB SOLUTIONS'S TOTAL AGGREGATE LIABILITY (WHETHER IN CONTRACT (INCLUDING FUNDAMENTAL BREACH), WARRANTY, TORT (INCLUDING NEGLIGENCE), PRODUCT LIABILITY, INTELLECTUAL PROPERTY INFRINGEMENT OR OTHER LEGAL THEORY) WITH REGARD TO THE SOFTWARE AND/OR THIS LICENSE EXCEED THE LICENSE FEE PAID BY YOU TO ANTLAB SOLUTIONS. FURTHER, ANTLAB SOLUTIONS WILL NOT BE LIABLE FOR ANY DELAY OR FAILURE TO PERFORM ITS OBLIGATIONS UNDER THIS LICENSE AS A RESULT OF ANY CAUSES OR CONDITIONS BEYOND ANTLAB SOLUTIONS' REASONABLE CONTROL

10. Jurisdiction
These terms will be governed by Japanese law and the Japanese courts shall have jurisdiction.

KNOWN ISSUES

None

REVISION HISTORY

1.2.1
A minor upgrade correcting a single bug.
Bug fixes

1. A bug that caused files to be saved as ASCII instead of UTF-8 under certain settings has now been fixed.

1.2.0
A minor upgrade improving performance and sensible handling of malformed characters in the input files.
New features

1. The program runs much quicker by checking only the first 300 characters of a file instead of reading the entire file into memory.
2. The program now replaces malformed characters in both the input and output files with standard replacement characters.

Bug fixes

1. Dragging and dropping a file on to the program now correctly updates the input file counter.
2. Dropping a folder (instead of files) on to the program now correctly ignores the drop.

1.1.0
A minor upgrade featuring new features.
New features

3. The program can now convert files based on a manually selected character encoding or one of three auto-detection approaches.
4. The program now includes three different auto-detect systems (see the notes above)
5. The program now reports when files could not be converted correctly either because they contained no text data, the manually set character encoding produced an error when applied, or the auto-detected character encoding produced an error when applied.
6. Files loaded into the program can now be closed individually (by selection) or all at once.
7. The file selection window now has more text selection options.

8. Encodings are now always displayed in ALL-CAPS.

1.0.0
This is the first version of the program