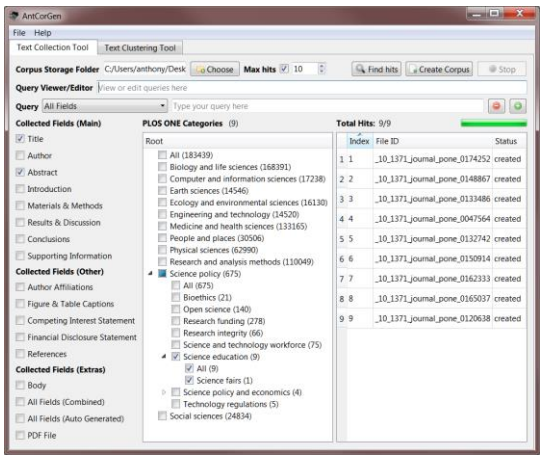
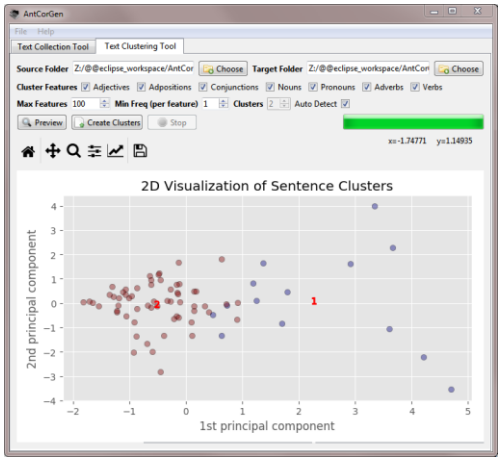
AntCorGen

Build 1.1.1 (Released February 28, 2018)

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: 001.





Introduction

AntCorGen is a freeware corpus generation tool. AntCorGen lets you search for documents in the PLOS ONE research database via search queries and/or subject category browsing and decide which sections (e.g. title, abstract, introduction) of these documents should be stored. *AntCorGen* then accesses the database, downloads the sections, and saves each one as a text file in an appropriate folder. *AntCorGen* can also analyze the different parts of speech (e.g. adjectives, verbs) of words in the files and cluster similar sentences into sub-groups. These sub-groups will show similar patterns of language use.

AntCorGen 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 *AntCorGen* icon to launch the program.

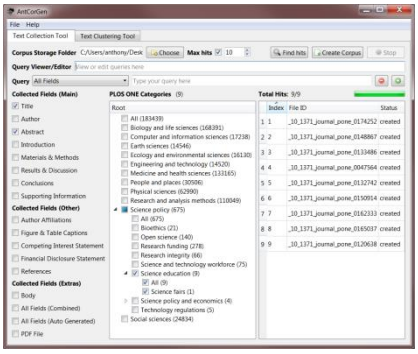
Macintosh OS X

On Macintosh systems, simply double click the *AntCorGen* zip file. The zip file will unzip the *AntCorGen* application. Then, you can drag the *AntCorGen* 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 *AntCorGen* icon to launch the program.

Text Collection - Quick Guide



Step 1: Select a corpus storage folder into which the corpus files will be saved using the "Choose" button.

Step 2: Choose documents to be included in the corpus collection.
Option A: Search for relevant documents using the "Query Viewer/Editor" and/or "Query" settings:

1) The "Query Viewer/Editor" will show a complete query in the Solr search query language used by PLOS ONE. More information about the query language can be found at the following links:
    - [Tutorial] http://www.solrtutorial.com/solr-query-syntax.html
    - [Examples] http://api.plos.org/solr/examples/
    - [Main Solr site] http://lucene.apache.org/solr/
2) The "Query" tool will allow you build a query using field names, queries items, and AND/OR/NOT operations. To add/delete parts to the query, using the -/+ buttons. All changes made in the "Query" tool will be reflected in the "Query Viewer/Editor".

Option B: Browse for relevant documents using the "PLOS ONE Categories" browser tree:

1) Click on category branches in the browser tree to expand the branches and show sub-categories. The number of documents in each category is shown in parentheses.
2) Select categories to be included in the collection. The total number of documents within the selected categories is shown in the browser tree header.

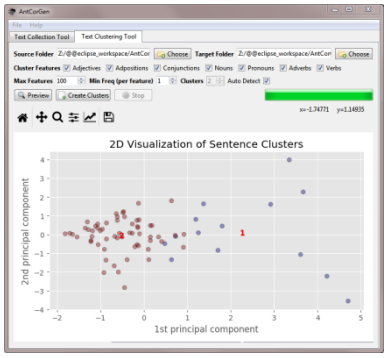
Step 3: Decide whether or not to set a maximum number of corpus files to collect using the "Max hits" checkbox option and spinbox values widgets.

Step 4: Click the "Find Hits" button to show an estimate of the total number of documents ("Total Hits") that will be collected. The result is shown in the status window.

Step 5: Click the "Create Corpus" button to collect the documents and store them in the corpus storage folder. The total number of documents (hits) will be updated to show how many have been collected. The id and status of the collection for each document is shown in the status window.

Step 6: Click the "Stop" button to stop the collection process at any time.

Text Clustering - Quick Guide



Step 1: Select a "source folder" of corpus files that you want to cluster using the "Choose" button.

Step 2: Select a "target folder" into which the clustered files will be saved using the "Choose" button.

Step 3: Choose features that you want to include as part of the clustering algorithm.

Step 4: Choose parameters (max number of features, min frequency of features, number of clusters) that you want to use in the clustering algorithm. To use all the possible features, set the "Max features" option to -1 (the default). If you are not sure how many clusters to pick, use the "Auto Detect" option.

Step 5: Click the "Preview" button to show a scatter-plot visualization of the clusters. If the clusters are not separated in the visualization, adjust the features and parameters as necessary. The scatter-plot can be resized, zoomed, label-adjusted, and saved using the icons above the plot image.
Step 6: Click the "Create Clusters" button to cluster the document sentences and store them in the target folder.

Step 7: Click the "Stop" button to stop the clustering process at any time.

NOTES

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 *AntCorGen*

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

Anthony, L. (YEAR OF RELEASE). *AntCorGen* (Version VERSION NUMBER) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

For example if you download *AntCorGen 0.0.1* which was released in 2017, you would cite/reference it as follows:
Anthony, L. (2017). *AntCorGen* (Version 1.0.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

LICENSE for *AntCorGen*

AntCorGen 0.0.1 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 *AntCorGen* 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. *AntCorGen* 1.0').

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 at present.

REVISION HISTORY

1.1.1 Minor update

Bug fixes

1) Fixed an error causing the program to create files for empty sections of an article with the contents of the previous non-empty section.

2) Although not strictly a bug, the program now generates two folders for the "Abstract" section of a research article, "abstract" and "abstract_primary_display". The data returned by the PLOS ONE API for the "abstract" field seems noise. The "abstract_primary_display" field, on the other hand, seems to match closer to what users need.

1.1.0 Minor update

New feature

1) Added an option to set a maximum number of articles to collect.

1.0.1 Minor update

Bug fixes

1) Fixed error causing the cluster tool to not generate results on some systems.

1.0.0

This is the first version of the program