



TerMine plugin for Protégé 4



1. TerMine plugin for Protégé 4 documentation

The Protégé TerMine plugin uses the TerMine tool to extract candidate terms from text and provides an interface for rapidly bringing these terms into an OWL ontology. The plugin accesses TerMine via a Web Service over the Internet. Use is currently free to academic researchers. In the UK, access is automatic for academic researchers connecting from a '.ac.uk' address. Non-UK based academic researchers must first register before being able to use the service. Other types of user may request access, although commercial rates will apply depending on the type and level of use. To request access to the TerMine service or for more information about TerMine please see <http://www.nactem.ac.uk/software/terminer/>.

TerMine is a term extraction tool developed at the UK National Centre for Text Mining (NaCTeM). It uses a domain-independent method (based on the C-value measure) to extract candidate terms from English text for consideration by the ontologist or terminologist. It is particularly oriented towards extraction of candidate multiword compound terms. References to relevant publications on the C-value and TerMine are available at the above URL.

Gathering terms is an important part of the ontology development process. Once gathered, these terms are entered into the ontology as labels for classes, which can then be used to build a knowledge model covering a particular domain. Most domain experts will have access to large corpora of data, be it text books, journals or scientific papers; these documents will contain many of the concepts that the ontologist would want to model in their ontology. The TerMine plugin can be used to get a snapshot of the important terms used in their domain and provides a simple interface for bringing these terms into their ontology. Such a tool can supplement a domain specialist's own knowledge and provide a relatively rapid means of harvesting many terms and generating classes in an ontology en masse—without much tedious clicking. Term recognition will gather a broad range of terms from a corpus, irrespective of the scope of the developing ontology. The TerMine plugin comes equipped with a means of filtering out unwanted terms and importing selected terms to specific locations within an ontology.

The plugin is built to work with Protégé 4; there is no support for protégé 3.

2. Installation

You will need the latest version of Protégé 4 available from the protégé web site <http://protege.stanford.edu>.

The TerMine plugin is installed via the protégé 4 plugins mechanism. Download the `uk.ac.nactem.owl.terminer<latest_version>.zip` file via the CO-ODE web site (<http://www.co-ode.org/downloads/protege-x/plugins.php>). Unzip this folder and copy the Java Archive (Jar) file called `uk.ac.nactem.owl.terminer.jar` to your Protégé 4 plugins directory. You can find the Protégé 4 plugins folder in your Protégé 4 installation directory. Next time you start Protégé 4 the plugin will be installed.

The TerMine service is accessed via a Web Service. You will need an Internet connection to access this service. In addition to this, free use of the TerMine service is restricted to certain types of user, as explained above. For more information about TerMine see <http://www.nactem.ac.uk/software/termine/>

3. Using the TerMine Plugin

You can access the TerMine plugin from the Tools menu in Protégé. Select 'Import from TerMine' and the TerMine wizard should appear: see Figure 1.

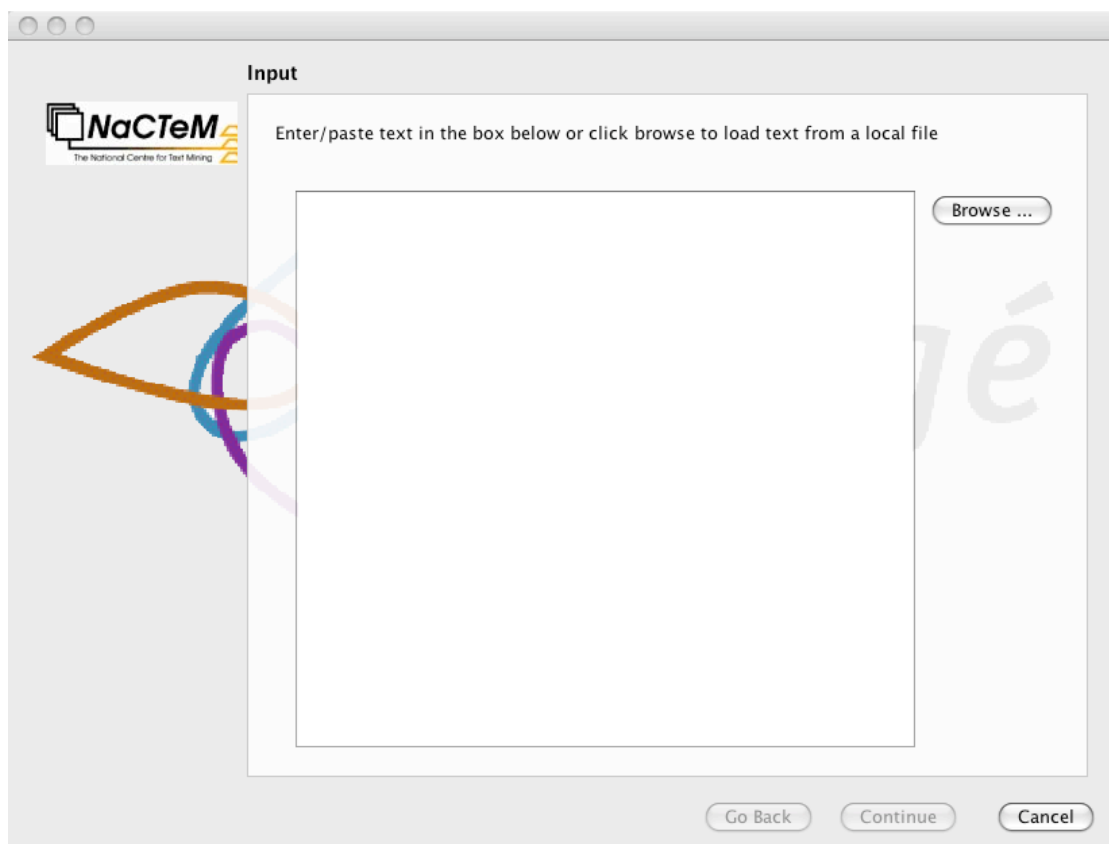


Figure 1. Step 1 view of TerMine wizard.

You can simply paste some text into the box or upload a file containing some text. Select continue to execute the TerMine service. (NB Ensure you have an active Internet connection and access rights to the TerMine service.)

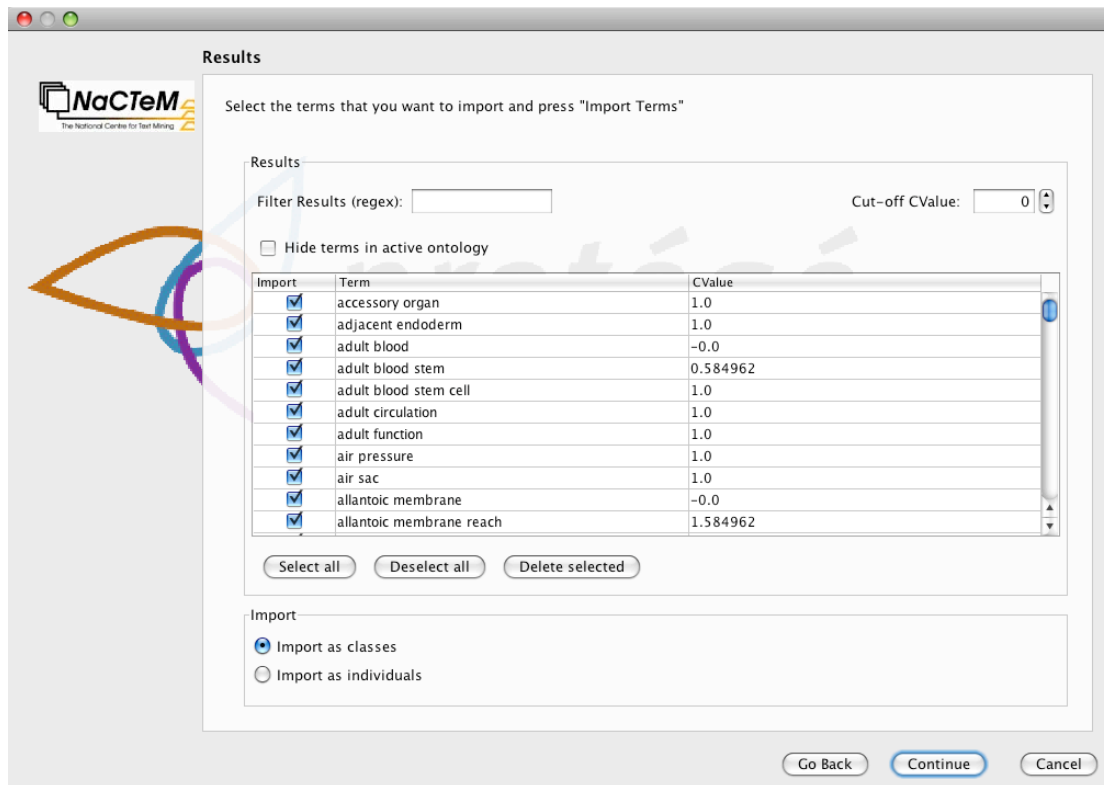


Figure 2. TerMine wizard results panel.

Figure 2 shows the TerMine wizard results panel. You can now select the terms you wish to import into your ontology. The extracted candidate terms are sorted alphabetically by default, however you can sort by the c-value score for each term by selecting the top of the c-value column. The c-value is a measure of the significance of the candidate term within your corpus, calculated by the TerMine service.

You can filter your results based on the c-value score by using the Cut-off CValue counter. You can also filter your results using regular expressions. For example, if you wanted to just see the terms that mention 'blood cell', you could type 'blood cell' into the filter results text box. For more information about using Java regular expressions see <http://java.sun.com/j2se/1.4.2/docs/api/java/util/regex/Pattern.html>.

A final filter available is to hide terms that already appear in your ontology. If you are using the wizard multiple times to bring terms into your ontology, you may want terms in your active ontology filtered from the results set, just select the tick box and they will be removed. You can get them back by deselecting the check box.

If you want to exclude a term from being imported into your ontology you can deselect using the tick box in the results table. Selected terms can be removed from the results set entirely by using the 'Delete Selected' button.

When importing terms into your ontology, you can choose to import these as classes or as individuals. By default, terms are imported as classes, however you can change this by selecting 'Import as individuals' at the bottom of the panel. By selecting continue you will move to the final panel of the TerMine wizard.

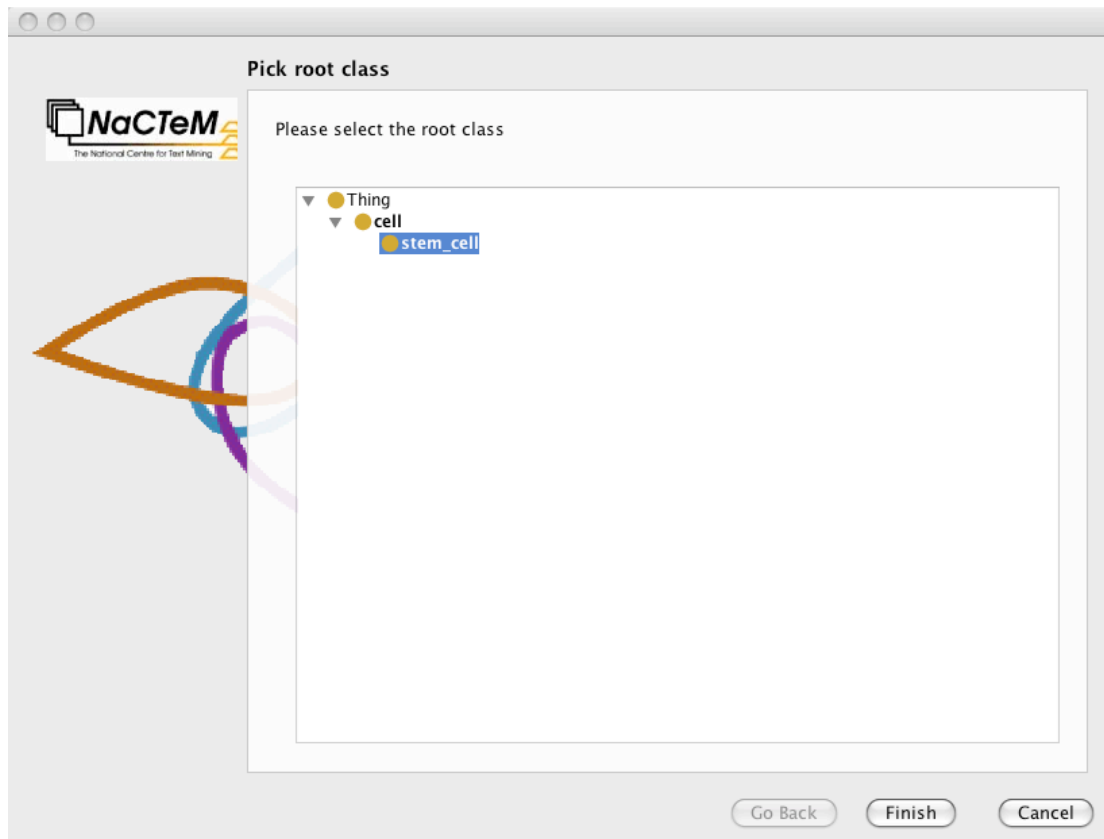


Figure 3. Select a root class for your imported terms.

With your terms selected, you can now choose where to bring them into your ontology (See Figure 3). By default, the terms will be imported as sub classes of Thing. In our example, we have selected ‘stem_cell’ as our root class. If you have chosen to import your terms as individuals, you can make them instances of the selected class. When you select the Finish button, the wizard will close and your selected terms will appear in your active ontology.

4. Getting the most out of TerMine

A small amount of data preparation will help you to get the most out of TerMine.

Input should be plain ASCII (for file upload, you can concatenate many small text files into one, for example).

You may wish to take special action over titles and section headings. TerMine has no knowledge of text structure and layout. Titles and headings often lack sentence-final punctuation and contain words that all have initial capitals. Certain capitalised words in headings may thus be wrongly interpreted as nouns or adjectives, giving rise to spurious candidate terms. Moreover, headings lacking sentence-final punctuation will be run together with the following sentence, causing the possible extraction of spurious candidates over the boundary. A simple solution to this latter issue is to ensure each heading is terminated with a full stop. It may however prove advantageous to remove all headings and titles before processing (often, the words in headings and titles will be repeated in the body of the text, in any case). It should be

stressed that any such spurious term candidates will be few in terms of number of individual occurrences: we mention these issues simply to allow you to understand why you may at times see them in the output.

Due to the statistical basis of the C-value measure, the longer the input text, the better the output will be, although TerMine gives useful results for even short texts.

In the interests of simplicity of use, the plugin does not offer the full functionality of the TerMine Web Service. If you wish to have greater control over the output, e.g. by using a stop list, then you may prefer to access the Web Service directly.

5. Contact

For help with installation or to report any bugs please contact Simon Jupp:
Simon.Jupp [at] manchester.ac.uk

Alternatively, please see the TerMine web page:
<http://www.nactem.ac.uk/software/termine/>

6. Credits

Simon Jupp and Matthew Horridge (School of Computer Science, University of Manchester) wrote the plugin for NaCTeM.

Dr Naoaki Okazaki (NaCTeM and University of Tokyo) developed the TerMine tool and associated Web Service.

7. About NaCTeM

The UK National Centre for Text Mining (www.nactem.ac.uk) is the first publicly-funded text mining centre in the world. It is operated by the University of Manchester in close collaboration with the University of Tokyo. It is funded by the UK JISC to provide text mining services to the academic community. The Director of NaCTeM is Dr Sophia Ananiadou (Sophia.Ananiadou [at] manchester.ac.uk). TerMine is only one of the services and tools available, please visit our web site to learn more.

Should you find this plugin of use in your research, we would welcome a small acknowledgement to NaCTeM when you publish.

8. Missing features

Here is a list of features that are currently missing from the current release but may be included in the future:

- Import text from multiple documents
- Import text from the web via URL

- Import text from PDF
- Submit stop list to TerMine service
- Load/Save extracted term lists