Genesys Knowledge Management User Guide

Notes on Language

5/7/2025

Contents

Notes on Language

This topic describes part of the functionality of Genesys Content Analyzer.

Selecting a Language

The first step in creating a training object is to select a tenant and language. The language that you choose has special relevance in the following two cases:

- If you want to build a model that classifies according to language, you must use a tree whose language is specified as `unknown`. First you must add this language attribute in `Configuration Manager > Business Attributes > Languages`.
- Selecting English activates a lexical analyzer that is specific to English. If you are operating in a language other than English, you should not select English because the English lexical analyzer will hinder the training.

Selecting any language other than English activates a default lexical analyzer. You can also create a lexical analyzer that is specific to any language you choose.

Lexical Analyzer

The function of a lexical analyzer is to convert input text (such as the text of an e-mail) to an array, either of words or of stems.

Content Analyzer includes the following:

- Language-specific lexical analyzers for English and Japanese. The latter is available with the Content Analyzer - Japanese edition.
- A default lexical analyzer that is much simpler than the English analyzer.
- Sample code that you can use to create your own lexical analyzer for a language of your choice.

Language-Specific Analyzers

The language-specific analyzers convert words to *stems,* delete digits and special characters, and segment the result into an array, omitting *stop words.* To explain two key terms:

- A *stem* is a basic string that is shared by a family of words; for example, *see, seen,* and *seeing* all have the same stem, as do `allow`, `allows`, and `allowance`.
- *Stop words* are words that are so common that there is little to be gained in searching for them or listing their occurrences. Examples for English are *the, a, an, of, to, is,* and so on. The system does not consider stop words when performing classification, and stop words do not appear on the Indexing Tab of the TO Data Analyzer. The installation packages for Genesys Content Analyzer install a file of stop

words for English, called `English.stop`, in the home directories of Knowledge Manager and Training Server. This is a simple text file containing a list of words separated by carriage return. You can create other files of stop words for other languages.

Default Analyzer

The default analyzer is less robust than the language specific ones: it does not distinguish stems, and it does not ignore stop words. It considers any sequences of alphabetic characters (a–z, A–Z) to be words, and all other characters to be word separators. For this reason, this analyzer requires significantly more training to work satisfactorily with languages such as Arabic, Chinese, or Korean, which do not indicate separations between words.

> Here is an example contrasting the English lexical analyzer with the default lexical analyzer: Take the phrases *I find your service disappointing* and *I find your service to be a disappointment*. The English lexical analyzer is smart enough to know that *disappointing* and *disappointment* share the stem *disappoint*. It also ignores the word *a*. So if it is trained to classify the first phrase as negative, it does not have to conduct a separate analysis of the second phrase to know that it is also negative. The default lexical analyzer is not as smart at picking out similarities in words and phrases, so it would conduct separate classifications of *disappointing* and *disappointment*. It would also waste time considering that *a* might be a marker of negative sentiment. That is why the amount of training needed to obtain the same level of precision is greater with the default lexical analyzer than with the English one.

Analyzer Code Sample

This sample is a class that implements the LexicalAnalyzer and Serializable interfaces. The LexicalAnalyzer interface includes two methods:

```
public interface LexicalAnalyzer {
  public String getLanguage();
  public String[] convert(String text);
}
```

- `public String getLanguage()` returns the name of the language that this lexical analyzer applies to.
- `public String[] convert(String text)` converts text to words or stems.

You can add one or more lexical analyzers for languages of your choice. To do this, you must prepare a Java class that implements the LexicalAnalyzer interface with the two methods just described.

The lexical analyzer example is located in `<KnowledgeManagerHome>\LexicalAnalyzerExample`. `<KnowledgeManagerHome>` is normally something like `C:\Program Files\GCTI\eServices 8.1.4\Knowledge Manager`. The source code is in `LexAnalyzerTest.java`. To adapt it to a language of your choice, use the following procedure.

1. Select a name for the language that your analyzer will apply to.
2. Adapt the `LexAnalyzerTest` class to the target language, changing the name of the class and

substituting the language name that you selected for the name used in the example (`English09`). For the purposes of this description, suppose you rename the class `MyLexAnalyzer.`

3. Compile the `MyLexAnalyzer` class using the following command:

   ```
   javac -classpath "gcengine.jar" MyLexAnalyzer.java
   ```

   The `gcengine.jar` file is located in the `LexicalAnalyzerExample` directory.

4. Copy the resulting `MyLexAnalyzer.class` file to the home directories of Knowledge Manager, Training Server, and Classification Server.

> **Important**
>
> The stop word file must be in the UTF-8 format (prior to release 7.6, stop word files required the ANSI format).

Content Analyzer Japanese

Genesys Content Analyzer – Japanese is a lexical analyzer for Japanese, available as an extra option. To use it, contact your Genesys representative to purchase a license, then proceed as follows:

- Locate the `license.dat` file and copy it to `<KnowledgeManagerHome>\LexicalAnalyzerGLA\lang.` Overwrite the dummy `license.dat` file that is already there.
- Add a language called `Japanese_GLA` to `Configuration Manager > Business Attributes > Languages.`