CZSaw Tutorial
July 2011

Contents

Getting Started

When CZSaw is run it initially has an empty window. At this point, you must create a new CZSaw project or open an existing one.

Creating a New Project

To create a new project, select *New Project…* from the **File** menu. A file dialog box appears to let you choose the location to save the new project. When you have chosen a location, click the *Save Project* button in the dialog box to create the new project.

The project is saved as a folder that includes its script and the history data and images for CZSaw's history view. The script and history data are automatically saved to files throughout your use of CZSaw, which is why a save location is required before you can work with CZSaw. For this reason there is also no *Save Project* option in the **File** menu. The project is always being saved.

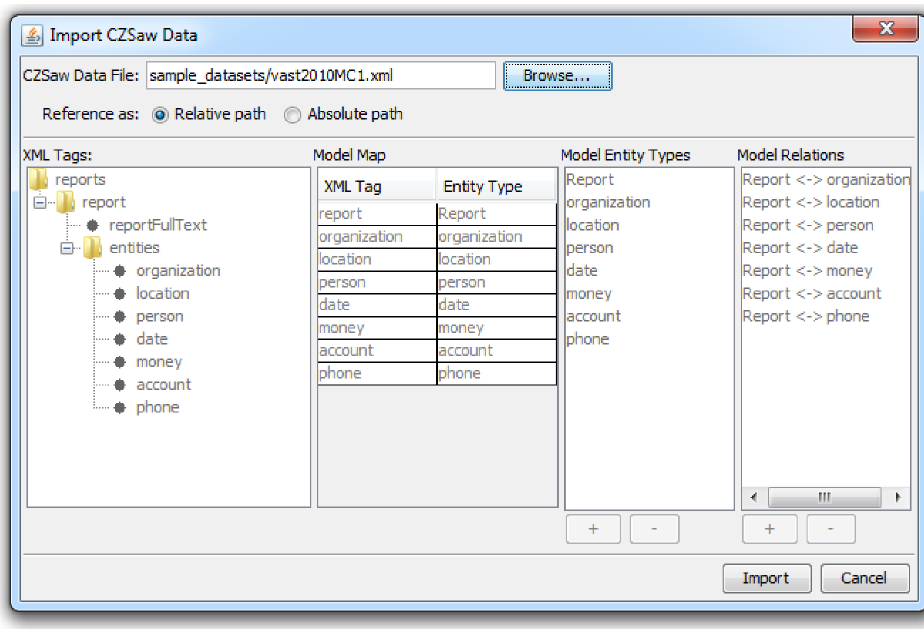
When a new project is created, the Script View, History View, and Dependency Graph View will fill the three divided areas of the window; however, none of them will have any content. The next step is to import some documents into CZSaw.

Import Documents

There are two options for importing documents into CZSaw. An XML file containing an entire document collection that is already in CZSaw's czd format can be imported or you can import unstructured text documents that are each in their own file from a directory.

Import CZSaw XML Files

The sample XML files that come with CZSaw are in CZSaw's czd XML Schema Definition (XSD). XML files that match this schema can be loaded by selecting *Import CZSaw XML File…* from the **Data** menu. A dialog box will appear for the import of an XML file. Click the *Browse…* button and select the XML file you wish to import. Click *Choose* and the dialog box should look similar to below.



The field beside the *Browse…* button displays the path to the data file, which will be used in CZSaw's script to load the file each time the script is run. Below the field is an option to choose between using a relative path or an absolute path to the script.

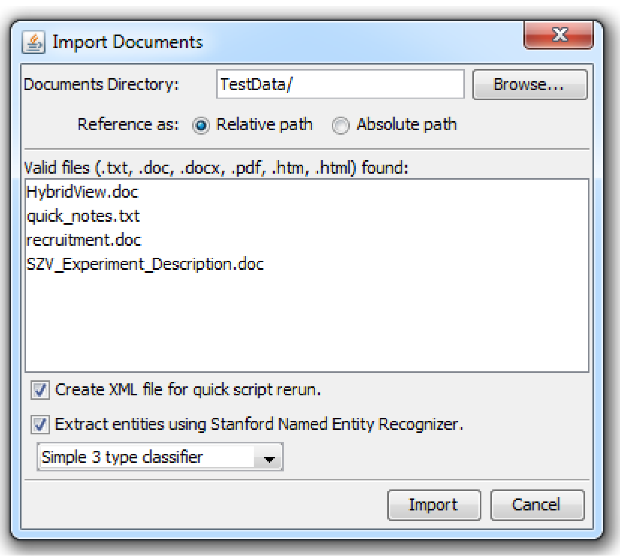
An absolute path does not depend upon the directory that you run CZSaw from. This means you could move CZSaw to another program and rerun the script of your project successfully as long as you do not move the data file you choose here. A relative path depends upon where you are running CZSaw from. This is useful if you place the data file in a folder within your CZSaw directory. Then you may move the CZSaw directory and the data file will go with it. Thus, you can transfer CZSaw to another computer or give your script to another analyst who has their copy of the data file in the same relative location.

The panels below the choice of reference are not currently editable in CZSaw. They display the hierarchy of the XML file chosen as well as the entity types (eg. person) and relations from the file that will be used in CZSaw. Click the *Import* button to import all the documents in the XML file into CZSaw.

A loading bar will show the progress of loading all the reports and their entities into CZSaw. The terms "report" and "document" are used interchangeably in CZSaw. After this finishes, a second loading bar will show the progress of creating the Lucene search index. We use Apache Lucene for providing quick search capabilities in CZSaw. These loading bars may go by very quickly if the data set is around 100 documents or less. Once the documents have been imported, the Search Panel will appear on the left and you may begin your analysis of the document collection.

Import Text Documents

You can also import a collection of unstructured text documents into CZSaw. Each document must be in its own file and the current valid file extensions are txt, doc, docx, pdf, htm, and html. Put all of the documents you wish to import into the same folder. To import the documents, choose *Import Documents…* from the **Data** menu. A dialog box will appear for the import of a collection of text documents. Click the *Browse…* button and select the folder containing the documents you wish to import. Click *Open* and the dialog box should look similar to below.



The field beside the *Browse*… button displays the path to the data file, which may (see next paragraph) be used in CZSaw's script to load the file each time the script is run. Below the field is an option to choose between using a relative path or an absolute path to the script. The difference between these two options is described in the previous section. The panel lists the valid files that were found in the folder chosen. Below this list are some options for how to handle the incoming documents.

Unless you uncheck the first box, the import process will create a CZSaw XML file from the documents and the script will refer to that file instead of the original directory.

Upon rerunning the script the data is reloaded and it is much faster to load from an XML file than to import all the individual files (and redo entity extraction). The downside of using an XML file is that any changes you make to the original document files will not automatically be used next time you rerun the script. Instead after making changes or adding more files, you must redo the import process to create a new XML file. The XML file that is created will have the same name as the folder chosen and will be saved in your CZSaw project folder. Thus, for the example image above, a file TestData.xml will be created and referred to by the script.

Unless you uncheck the second box, the import process will also extract entities from all the text documents being imported. CZSaw uses the Stanford Named Entity Recognizer to extract entities from the text. You can choose between 3 different classifiers that extract a different number of entity types:

- Simple 3 type classifier: location, person, organization
- Intermediate 4 type classifier: location, person, organization, misc.
- Advanced 7 type classifier: location, person, organization, date, time, money, percent

The more advanced classifiers do not take that much longer to run and identify many more entities so it is usually best to use the 7 type classifier. If you choose not to extract entities as part of the import process then the only entities types in CZSaw after the import will be reports (documents) and notes, although entities can later be manually extracted from the documents' text.

Click the *Import* button to import all the documents into CZSaw. Loading bars will show the progress of creating the XML file (if chosen), loading the entities and creating the Apache Lucene index for providing quick search capabilities in CZSaw. The first loading bar may take a while as it includes the extraction of entities (if chosen). Once the documents have been imported, the Search Panel will appear on the left and you may begin your analysis of the document collection.

Opening a Project

To open an existing CZSaw project, select *Open Project…* from the **File** menu. A file dialog box appears to let you choose the project. When the project opens, the Script View, History View, and Dependency Graph View will fill the three divided areas of the window; however, no other views will be present until some of the script is run. You can then use the Script View controls or click a thumbnail in the History View to execute the script to any point in your analysis process.