Informatics for Integrating Biology and the Bedside

i2b2 User Guide

Annotator Plug-in

Document Version: *1.1.1*

I2b2 Software Release: *1.5*

Table of Contents

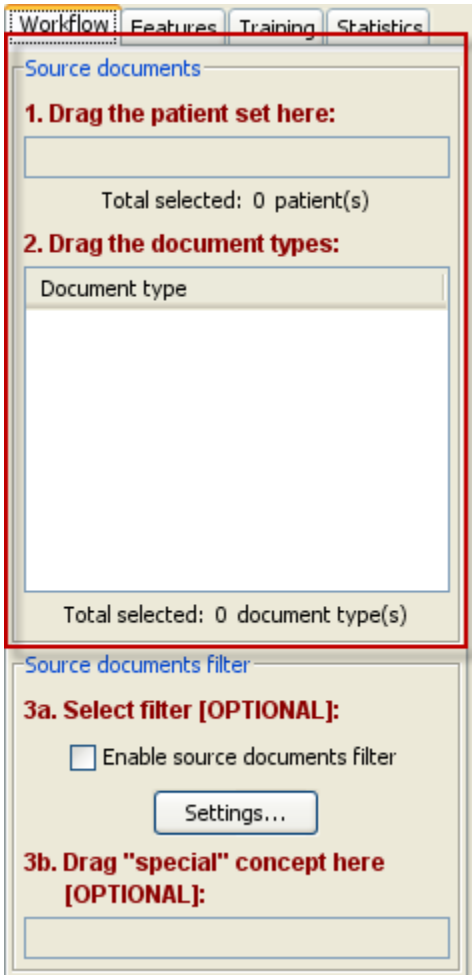
1. INTRODUCTION

This document describes a typical annotation workflow using the Annotator plug-in.

2. SETUP INFORMATION

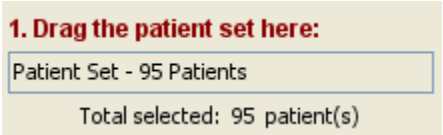
2.1 Source documents

The source documents section is where you will define both the patient set and types of documents that will be used during the retrieval process.



2.1.1 Adding Patients

The **patient set** to be used during document retrieval can be added to the **Annotator** by simply dragging the name of the patient set from either *Previous Queries* or *Workplace views* and dropping it into the Patient Set text box located on the Workflow panel.

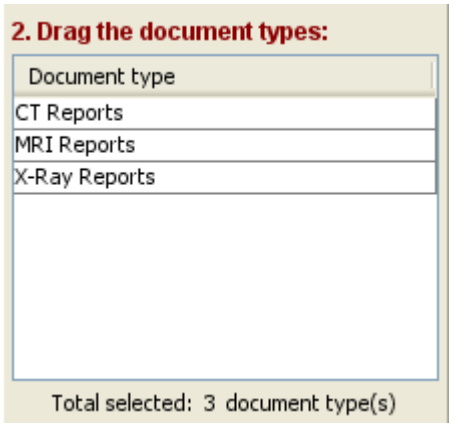


2.1.2 Adding document types

The **types of documents** to be used during document retrieval can be added to the **Annotator** by simply dragging the concept from one of the views listed below and dropping it into the Document types text box located on the Workflow panel.

1. Navigate Terms
2. Find Terms
3. Workplace

Only those concepts that are defined in Ontology as a "doc" type can be added to the Document Type text field in the Annotator.



2.2 Source documents filter

Both fields in the source documents filter are ***optional*** and are designed to assist the user in filtering the documents to be retrieved.

2.2.1 Setup source document filter

The filter defines which source documents are selected for inclusion in a batch or for classification. The source documents are only selected if they match the filter expression. The filter expression can be loaded from an external file or directly typed in the dialog window.

The filter is based on regular expressions.

2.2.2 Special label

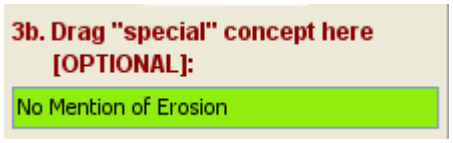
If the document filter is specified, this label will be assigned to the documents that do not pass the source document filter during automatic classification.

Example:

If the source document filter is configured to allow only documents with a mention of the word “erosion” the special label might be “No mention of erosion”.

The **“**special” label can be added to the **Annotator** view by simply dragging the concept from one of the views listed below and dropping it into the Special concept text box located on the Workflow panel.

1. Navigate Terms
2. Find Terms
3. Workplace

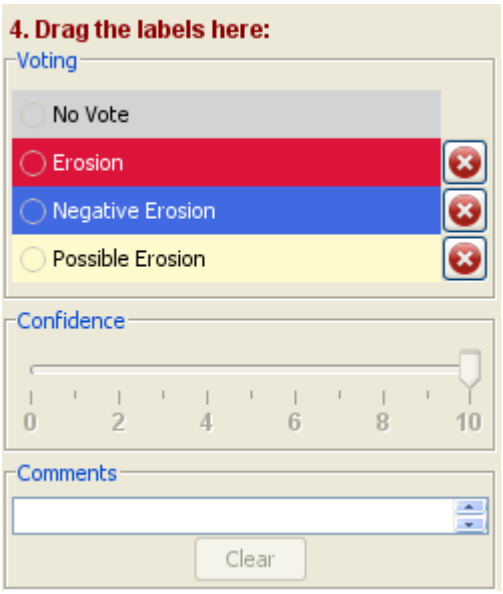


2.3 Voting Labels

The next step in the setup process is to define the labels that are to be used when voting on the documents. These labels are concepts that can be assigned to the documents.

The **Voting labels** can be added to the **Annotator view** by simply dragging the concept from one of the views listed below and dropping it into the Voting panel located on the left side of the view.

1. Navigate Terms
2. Find Terms
3. Workplace



The minimum number of labels required is two. Although there is no maximum number, the usual number of labels does not exceed three.

- ***Duplicates are not allowed.***

3. FEATURES

Features are *regular expressions* for which the Annotator looks in each document when running the active learning batch acquisition or automatic classification. They are required to build a classification model. The power of regular expression syntax allows user to create very dense features that cover large sets of related phrases/keywords. It is not required for each feature to uniquely identify one label. Although tremendously helpful, in practice this can rarely be achieved. Usually a collection of features is tightly associated with a label, similar to multiple symptoms that can be associated with a disease, but doesn't uniquely identify the label. Each feature may be associated with multiple labels, as each symptom can be associated with multiple diseases. It is the job of classifier to decide which features to associate with which label.

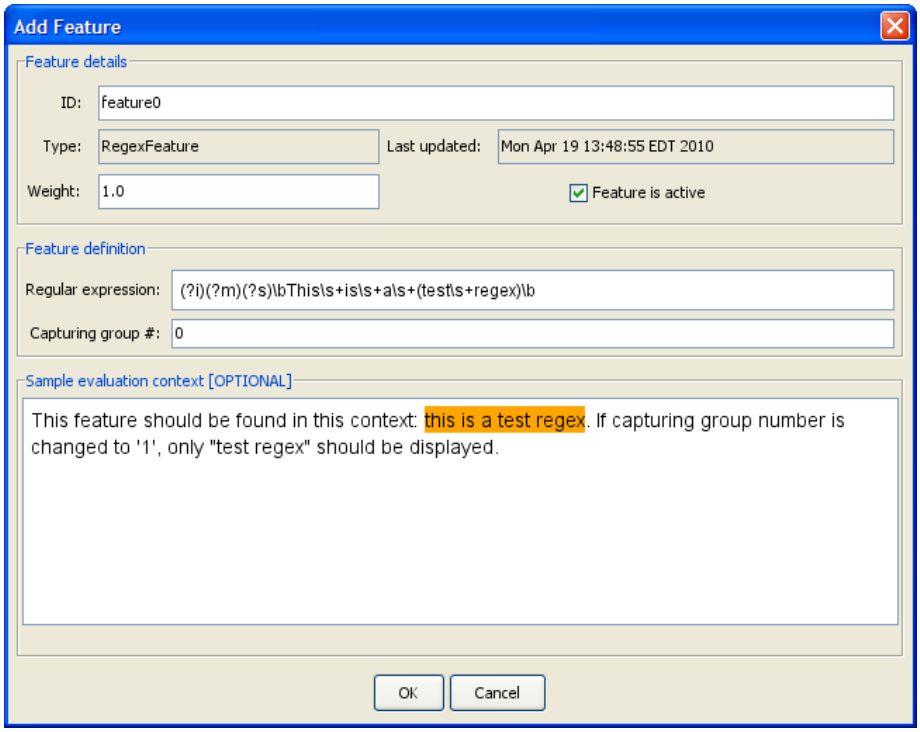
Because features may contain fragments of sensitive information from actual medical records, the content of features is encrypted. User will need a decryption key to load features from an external XML file. The key entry dialog will pop up if the key has not yet specified.

There are three ways to add features:

1. Manually add a new feature via the Feature tab.
2. Load an existing feature xml file.
3. Add when reviewing the document.

3.1 Manually Add Features

Users who are familiar with regular expressions can setup their own features by clicking on the Add button located at the bottom of the features panel. Once you click on the Add button the Add feature dialog box will open with an example.



Once you have modified the sample click on the OK button and the feature will be added to your list of features.

Once you have added your features you can save them to a features XML file. This will enable you to use the same features in a future session.

3.2 Load Features XML File

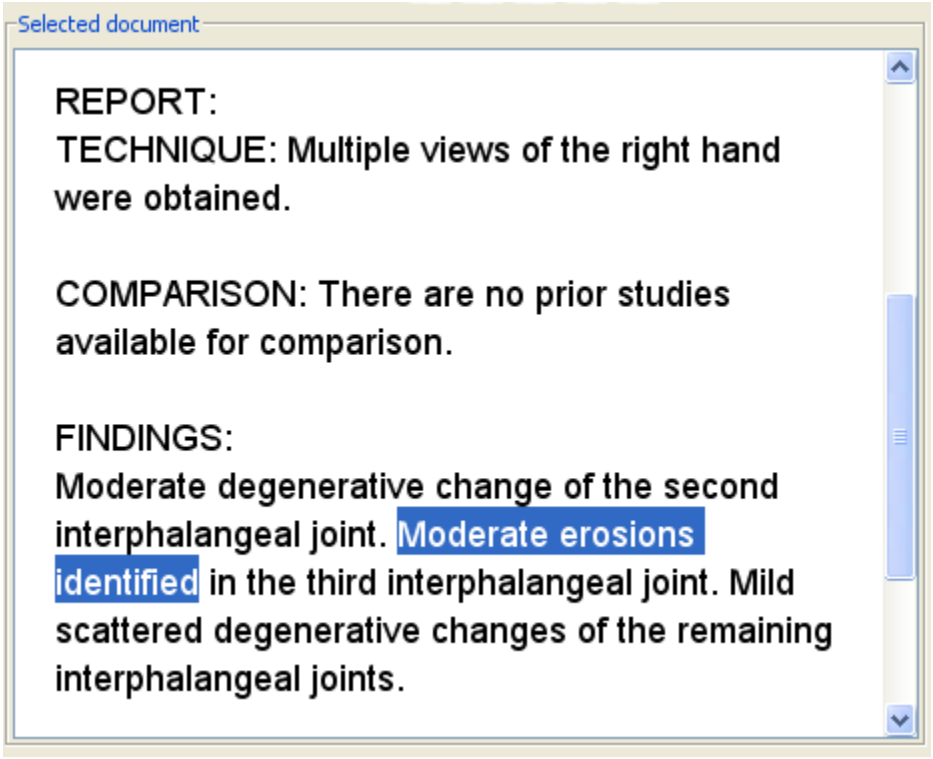
If you have a features XML file you can add the features in it to your list by clicking on the Load button located at the bottom of the Features panel.

3.3 Add Feature from Document Review

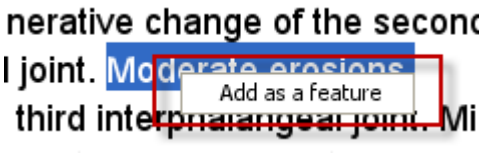
Features can also be added during annotation by highlighting fragments of the text in the active document. The following are the steps to add a feature from a document.

User should identify important short phrases in the document, highlight them and add as features.

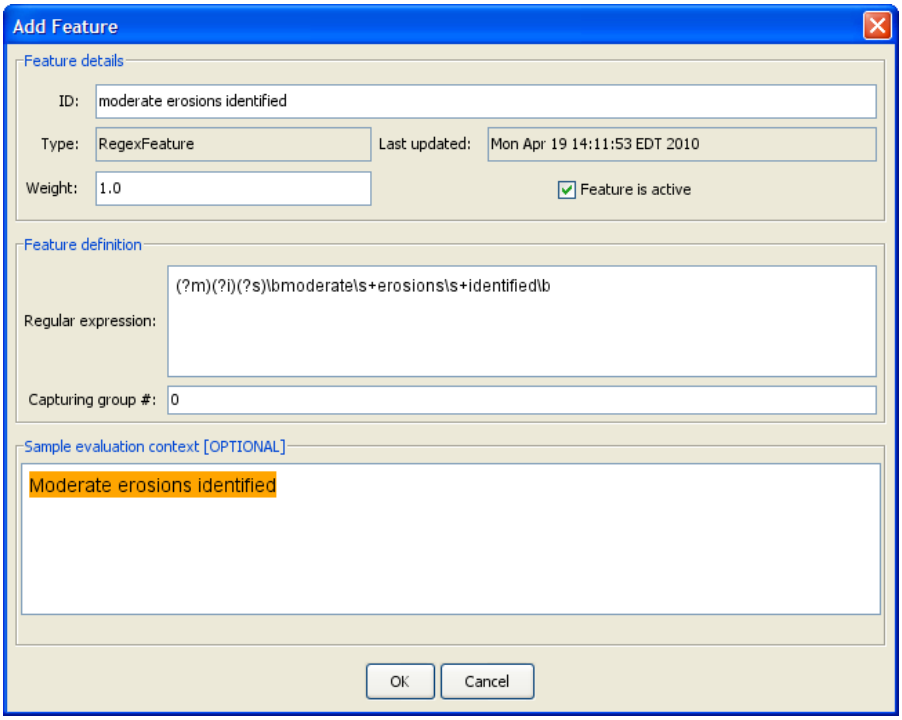
1. Highlight the text to be used in the feature.



2. Using the *right mouse button* click on the highlighted text.
3. Select **Add as feature** from the pop-up menu.



4. The **Add feature dialog box** will open and the highlighted text will automatically be formatted in *regular expression*.



5. Click on the **OK button** and the feature will be added to your list of features.

4. REQUESTING A BATCH

Request the batch of non-annotated notes using the “Request the next batch” button on the Training tab of the Annotator view.

You will be prompted to enter a decryption key if you have not already entered one during your current session.

There are two modes of batch acquisition; (1) Active learning and (2) Random. The following section describes each mode.

4.1 Active Learning Mode

The **active learning mode** allows users to use their previously annotated notes to guide the program to select the most relevant notes in the batch. The most relevant notes are those that are the most challenging for the classifier.

During this step the program builds an **SVM classifier** using the annotated (labeled) notes and evaluates its performance using 10-fold cross validation. Then a large random pool of *non-annotated notes* is selected from available notes. The size of this pool is controlled by the user. It can grow up to the number of all available documents for the patient set. However, large pool will considerably slow down batch acquisition and may results in out-of-memory errors.

Each document from the pool is classified by the SVM classifier. The output from the classifier is used to assign a difficulty score for each classified note. The notes with the highest difficulty score represent the biggest challenge to the classifier. The good news is that the application can put these difficult notes into a batch and let the user assign the correct vote to them. The human votes for these documents will help the classifier built during the subsequent batch acquisition to improve its classification quality.

The requirements to use the active learning are:

1. Need at least 10 human votes. This is because 10-fold-cross-validation is used internally to evaluate classifier performance during each batch acquisition.
2. Need human votes for each class label. Class labels not represented in human votes cannot be used in active learning, because the algorithm can’t “learn” anything about these class labels without human guidance.
3. Need at least one active feature.
4. Need a patient set and at least one document type selected – the batch will be selected from documents for patients in the patient set.

It is also recommended to use the source document filter to filter out irrelevant notes. Typically the note is relevant when it has some important keywords that the classifier also uses as part of features.

Example:

For smoking classification these keywords would be tobacco, smoking or cigarettes. These keywords should be presented as a single filtering regular expression such as the following:

\b(tobac\w+|cigar\w*|smok\w+)\b

4.2 Random Mode

In the case where no previous votes or features are available, random batch selection can be used. In this mode, the unlabeled notes are simply selected in the batch from the pool of all available notes. This mode also has a speed advantage over the active learning mode.

5. ANNOTATE DOCUMENTS

Annotate the documents in the batch using the labels and confidence slider on the right part of the annotator view. Confidence is optional and defaults to the maximum confidence of 10 in the scale of 0 to 10. There is also a free-text comment field on the right where the user may type any notes bout a particular document, such as why their confidence is low. This information is not used for anything.

5.1 Voting

The documents returned in the batch will be listed in the Train data panel located on the Training tab.

Voting on a document is done by simply clicking on the appropriate voting label when the document is active in the document view panel.

5.2 Loading votes

Load the existing votes from an external tab-separated file using the “Load” button on the training tab of the Annotator view. Loading votes allows users t resume annotation without losing their previous annotation results. The annotator will load all annotated notes, all notes from the last selected batch and all banned notes (if they exist).

5.3 Banning Documents

Some documents that pass through the source filter but are irrelevant can be “banned” using the “Ban selected” button. Such document will not be selected again as a part of any subsequent batches.

6. CLASSIFICATION STATISTICS

Repeat the batch acquisition / annotation steps as many times as needed. The decision about when to stop can be based on the quality of the classification model built using all the annotated notes. User can check this quality at any time during annotation by going to the statistics tab and requesting the updated statistic.

For example, users may decide to stop when the positive predictive value of the model reaches 96%. The minimum required confidence slider regulates which annotated notes are used to build the model. When it is set to 0, all notes are used. When it is set to 10, only notes about which the user had the highest confidence are used.

7. AUTOMATIC CLASSIFICATION

After the desired model quality is reached, the user may want to apply the annotation results and features to automatically classify all remaining documents in the patient set. For this purpose there is an "Automatic classification" button on the statistics tab of the annotator. It brings up the automatic classification dialog. The users can specify the type of classifier they want to use (SVM or Logistic Regression). If source document filter is set, it will be applied to filter out irrelevant documents, so that they do not reach the classifier. The user must specify the output file where to store the classification results locally. If the patient set is large and the user only wants to classify a part of it, there is an option to specify the limit. The results of classification will be uploaded to the file repository cell and committed to server. There is also an option that allows user to commit results of manual annotation before running automatic classification. The progress indicator messages logged to the console keep the user updated about classification progress.

8. SAVING WORK IN PROGRESS

The annotator allows users to save and restore their work. The following pieces of information can be used to restore the user session completely:

1. The active patient set can be dragged to the workplace.
2. The document types can also e dragged to the workplace.
3. The source document filter can be exported as an XML file.
4. All features can be exported as an XML file.
5. User votes, banned documents and the current batch of notes can be exported as a tab-separated file.
6. Labels can be dragged to the workplace.