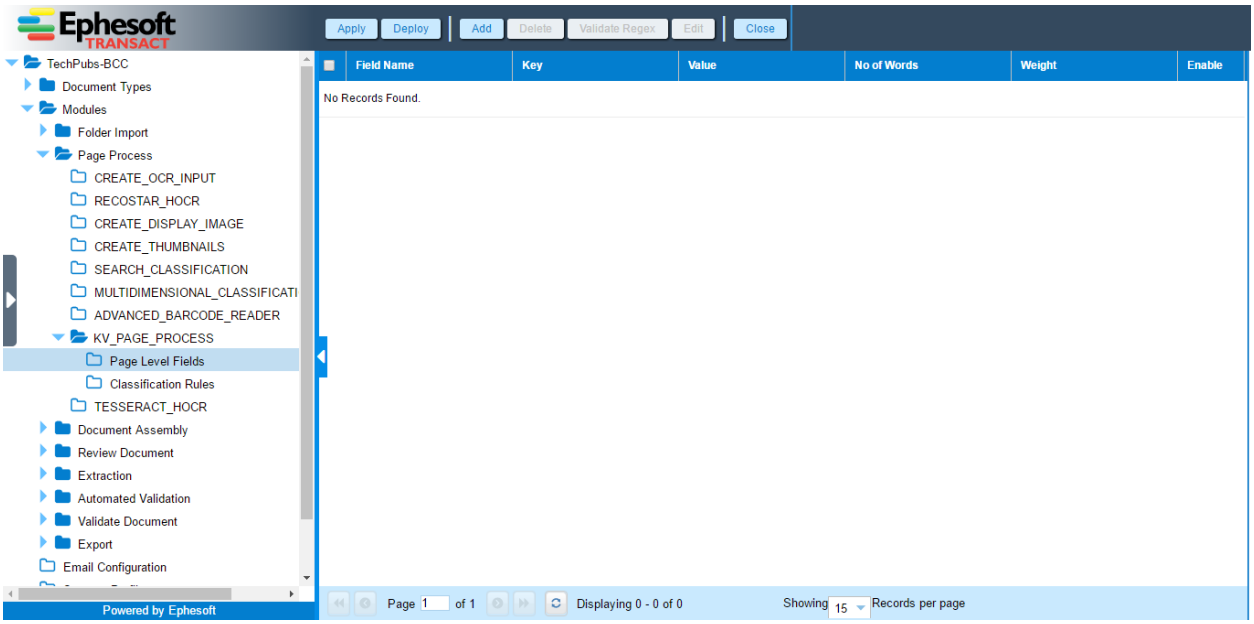
Keywords-Based Classification

Earlier, for user to be able to classify documents using keywords, user needed to specify the rules through '**ScriptPageProcessing**' in **SharedFolder**.

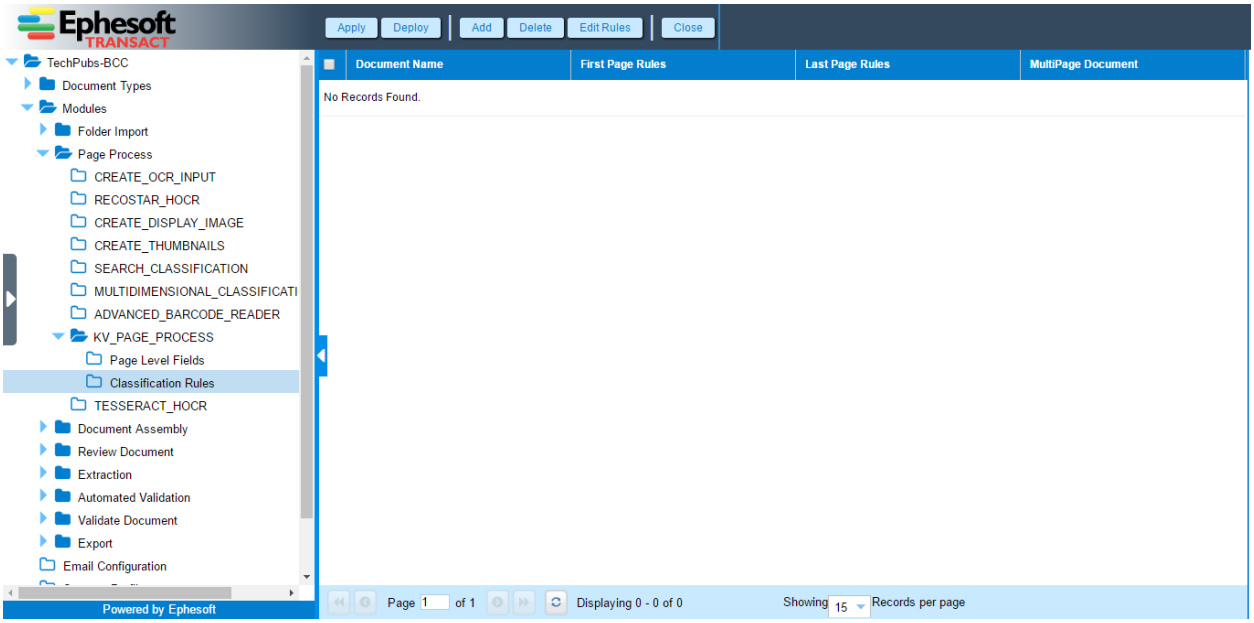
With Ephesoft v4.1.1.0 a new feature, **Keywords Based Classification**, has been implemented to improve document classification based on keywords often present/not present in the document types.

Writing the scripts to classify documents used to be a manual process. **Keywords Based Classification** intends to automate this. User can now have an interface to configure the rules enabling classification of document types based on keywords.

User can use the **Page Level Fields** node within the **KV_PAGE_PROCESS** Plugin to define page level fields for a document type.

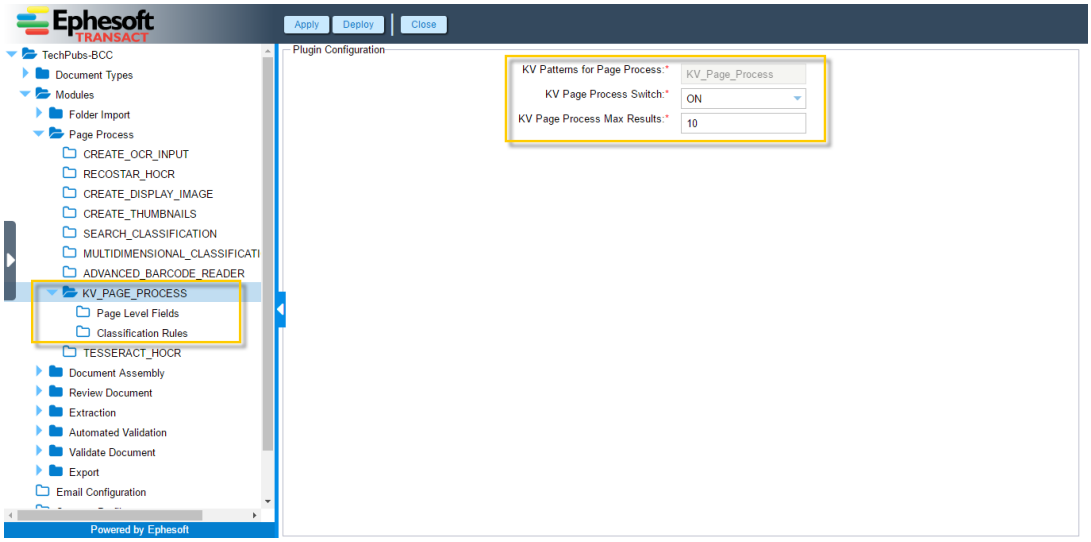


User can use the **Classification Rules** node within the **KV_PAGE_PROCESS** Plugin to set rules to be used for classification of document type based on keywords.



Configuration

KV_PAGE_PROCESS Plugin within the **Page Process** module governs the classification of documents based on keywords.



If the value of the switch is set to **ON**, user can classify documents based on keywords, else not. By default, the switch is set to **ON**.

Creating Page Level Fields

To create page level fields

1. From the DCMA Home page, click **ADMINSITRATOR** and select **BATCH CLASS MANAGEMENT.**

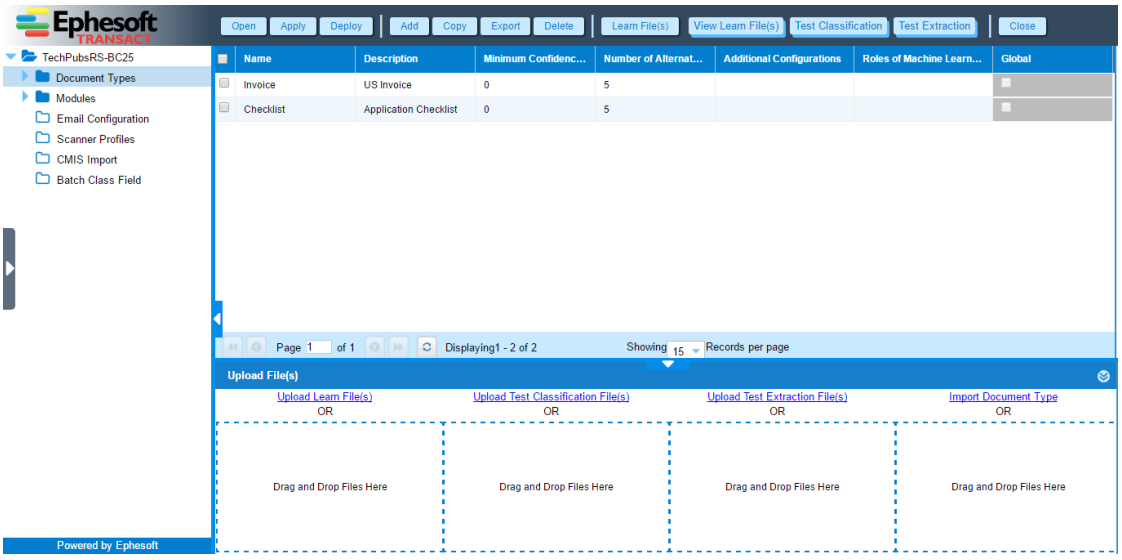
   The Ephesoft Enterprise **Login** page displays.

2. Enter valid credentials to login.

   The **Batch Class Management** screen displays.

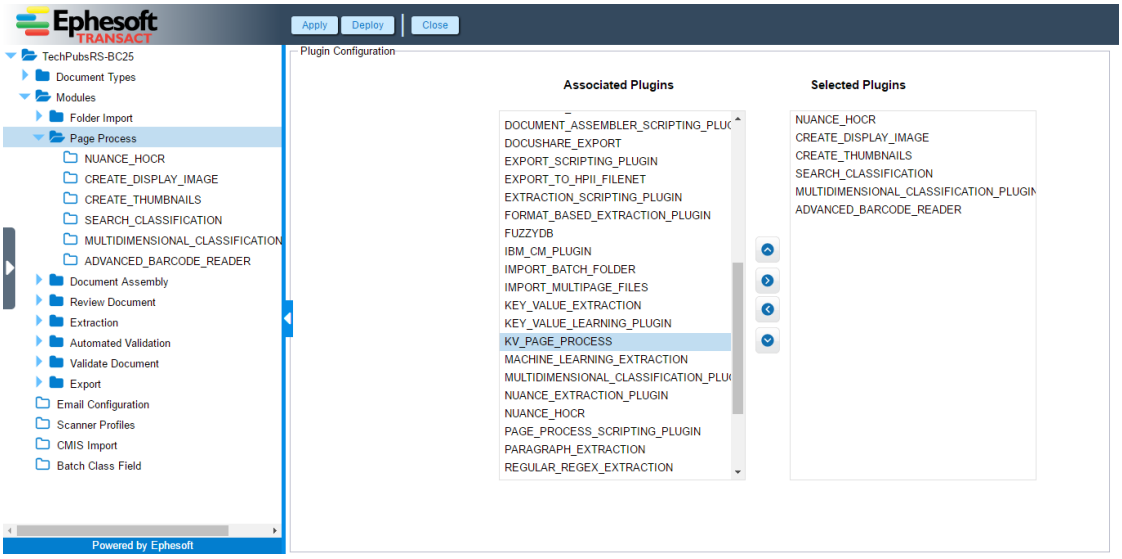
3. Select the batch class in question and click **OPEN** from the toolbar on top of the **Batch Class Management** screen.

   The batch class opens with a list of document types.



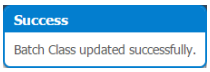
4. From the left navigation pane, go to **Modules** > **Page Process**.

   The **Plugin Configuration** screen displays.



5. Move the **KV_PAGE_PROCESS** Plugin from the **Associated Plugins** column to **Selected Plugins** column and click **APPLY** and **DEPLOY** from the toolbar on top of the screen.

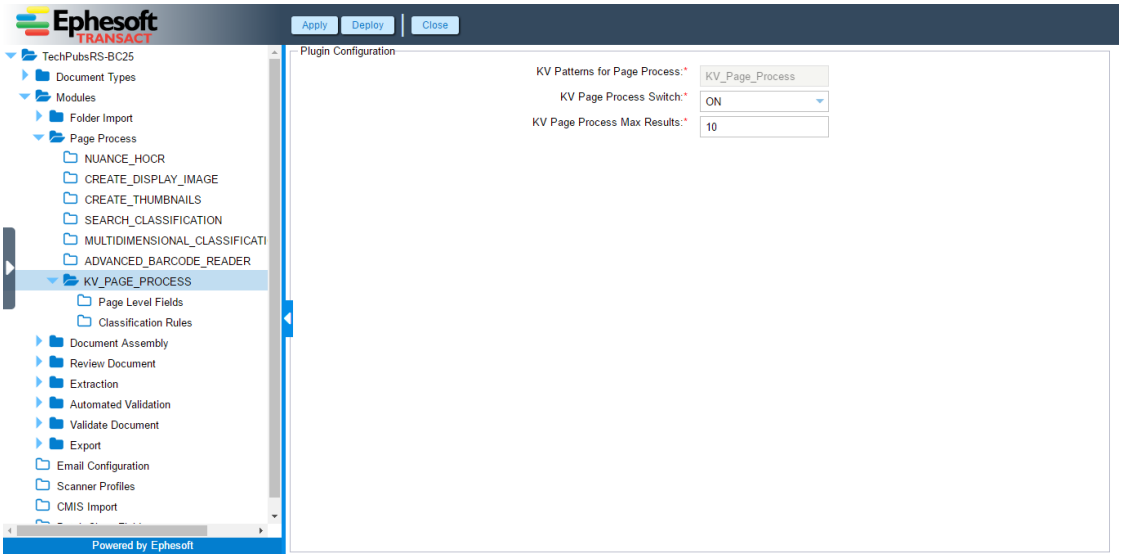
   The following message appears notifying that the plugin has been added to the batch class.



Some plugins have dependencies on certain other plugins. In this case, user may get the following message.

Click **YES** to add selected plugin along with the dependencies. Click **NO** to add the selected plugin without the dependencies. Click **CANCEL** to cancel the operation.

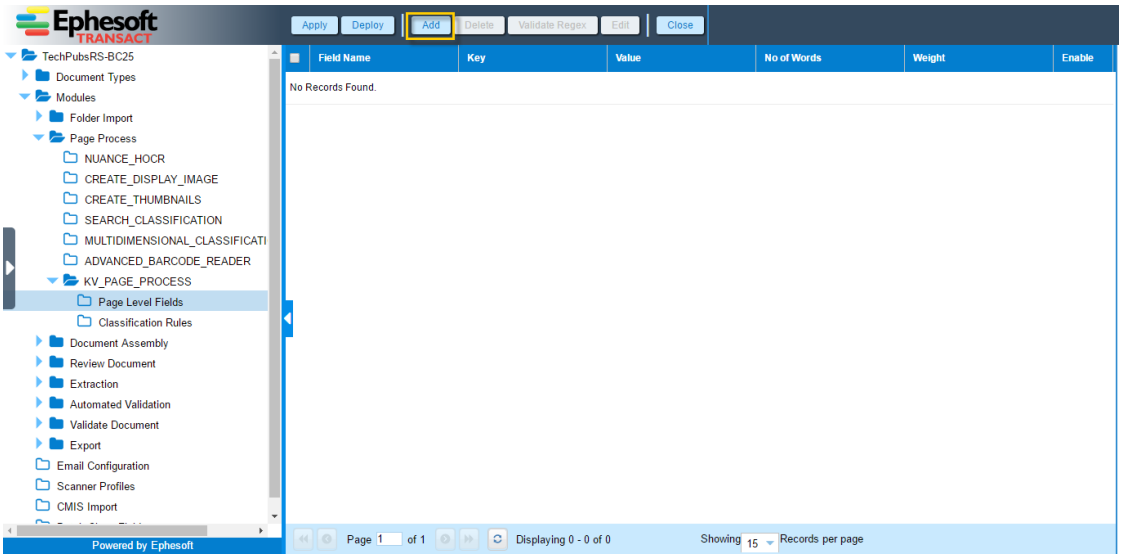
The **KV_PAGE_PROCESS** Plugin starts displaying in the **Page Process** module.



If the value of the switch is set to **ON**, user can classify documents based on keywords, else not. By default, the switch is set to **ON**.

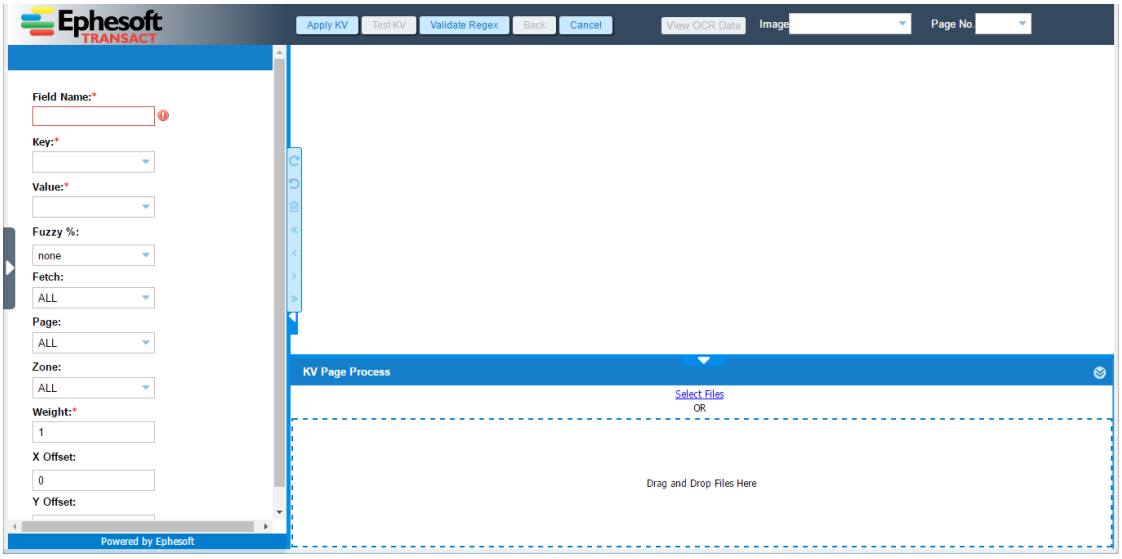
6. From the left navigation pane, select **Page Level Fileds**.

   The **Page Level Fields** screen displays.



7. Clcik **ADD** from the toolbar on top of the page.

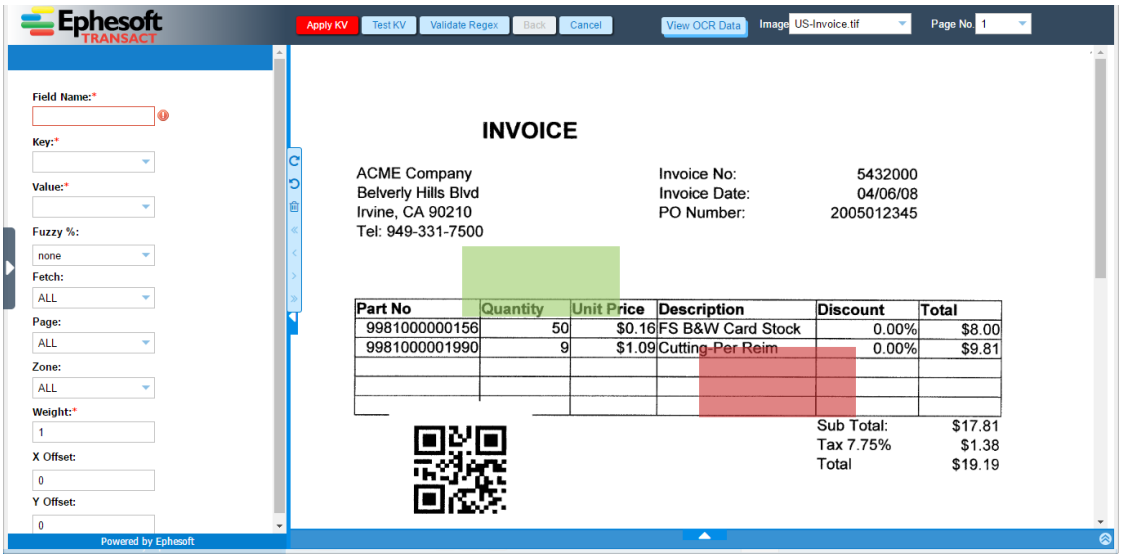
   The following screen displays.



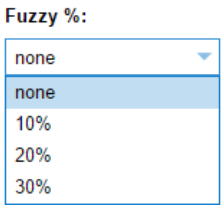
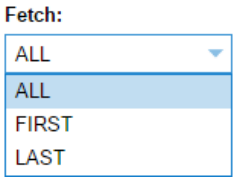
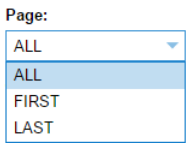
8. Click **Select Files** link from **KV Page Process** section and upload a image file.

   The uploaded image is displayed in the image view pane.

User can also drag and drop the image file to the **Drag and Drop Files Here** area below the **Select Files** link.



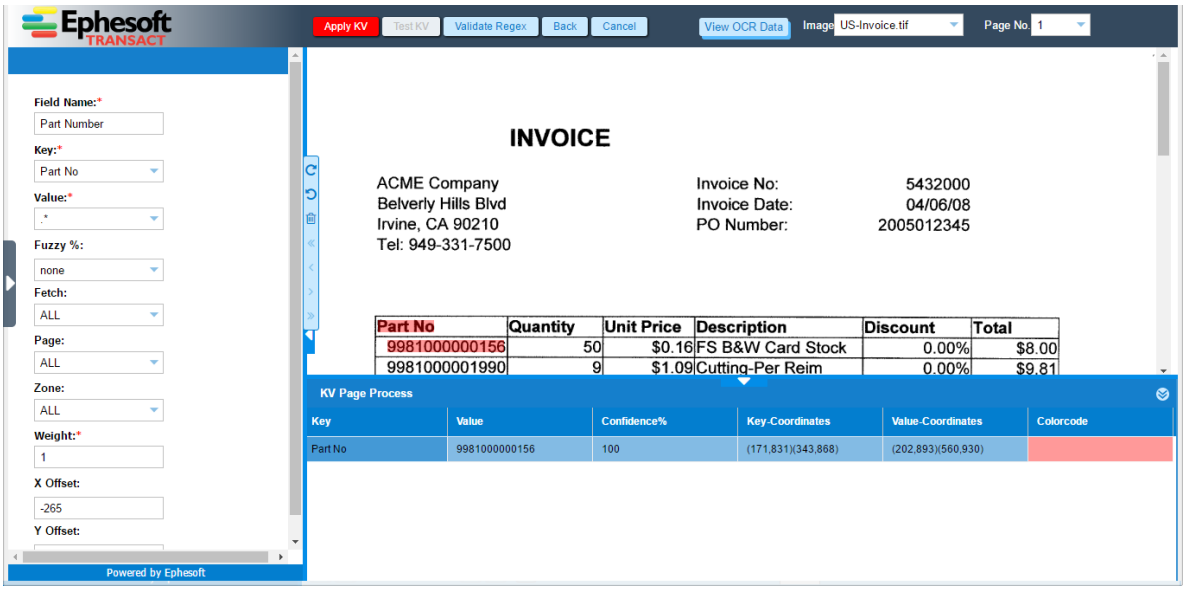
9. Enter the relevant configuration details as described in the table below:

<table>
<tr><th>Component</th><th>Description</th></tr>
<tr><td>Field Name</td><td>Enter a unique name for the page level field. This name is used in classification rules to create rules using this Page Level Field.</td></tr>
<tr><td>Key</td><td>Enter a regex pattern or predefined field corresponding to which extraction value is to be located.</td></tr>
<tr><td>Value</td><td>User can enter a regex pattern or use Regex Builder/Regex Pool options to enter a search pattern for the value to be extracted.</td></tr>
<tr><td>Fuzzy %</td><td>User can use this parameter to do a fuzzy search while searching for the value.
 
</td></tr>
<tr><td rowspan="4">Fetch</td><td colspan="2">User have three options available to choose from for this parameter: ALL, FIRST, and LAST.
 
</td></tr>
<tr><td>ALL</td><td>To extract all data from the value zone matching the value pattern specified.</td></tr>
<tr><td>FIRST</td><td>To extract only first data from the value zone matching the value pattern specified.</td></tr>
<tr><td>LAST</td><td>To extract only last data from the value zone matching the value pattern specified.</td></tr>
<tr><td>Page</td><td>User have three options available to choose from for this parameter: ALL, FIRST, and LAST. Depending on the selected value, the extraction algorithm runs on ALL/FIRST/LAST Page of the document.
 
</td></tr>
<tr><td>Zone</td><td>Every page in divided into 5 zones: TOP, MIDDLE, BOTTOM, LEFT and RIGHT along with the default option of ALL.</td></tr>
</table>

Component	Description
	Zone: ALL ALL TOP LEFT RIGHT MIDDLE BOTTOM User can use this parameter to specify the portion of the page where the algorithm searches for value. For example, if user configures this parameter value as BOTTOM, the value is searched only in the BOTTOM zone.
Weight	User can use this parameter to implement weighted confidence values. **Weight:*** 1 This is used to give weightage to a particular extraction rule.
X Offset	Coordinates of the value on X axis.
Y Offset	Coordinates of the value on Y axis.

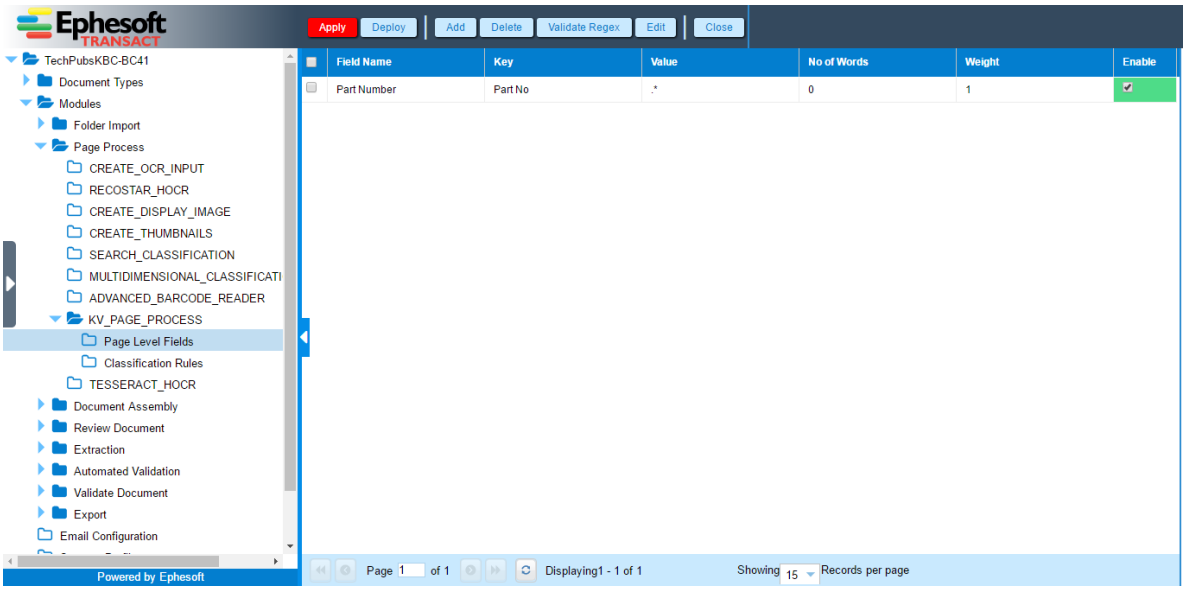
10. Click **TEST KV** from the toolbar on top of the page.

    The extraction result is highlighted on the image as an overlay and are also displayed in the **KV Page Process** grid as shown in the image below.



11. Click **APPLY KV** to apply the rule to the page level field.

The updated **Page Level Fields** screen displays with the following information: **Field Name**, **Key**, **Value**, **No of Words**, **Weight**, and **Enable**.



12. Click **APPLY** and **DEPLOY**.

Follow the process described above to add multiple page level fields.

Creating Classification Rules for Page Level Fields

To create classification rules for page level fields

1. From the DCMA Home page, click **ADMINSITRATOR** and select **BATCH CLASS MANAGEMENT.**

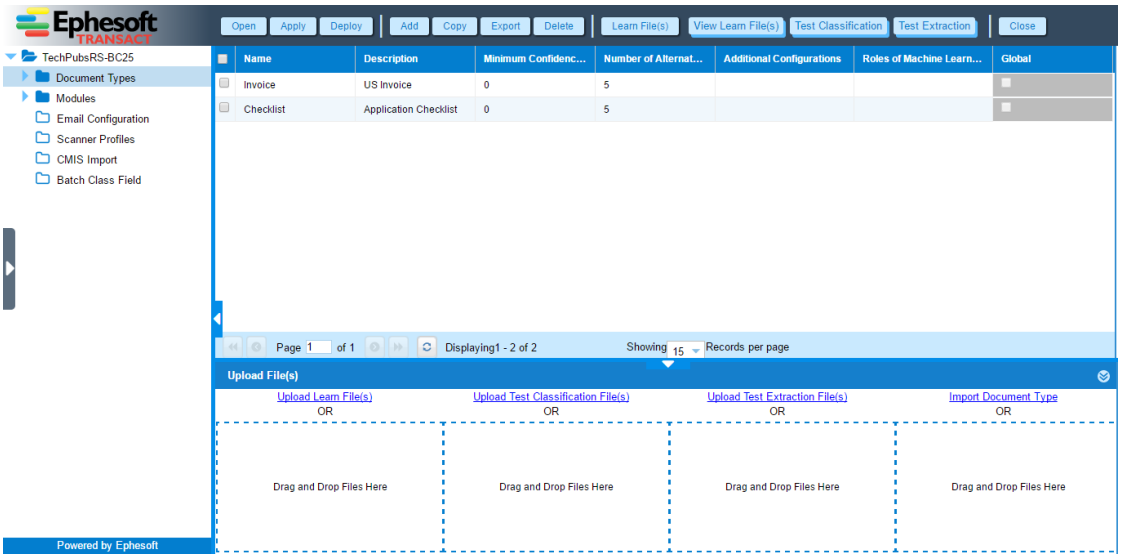
   The Ephesoft Enterprise **Login** page displays.

2. Enter valid credentials to login.

   The **Batch Class Management** screen displays.

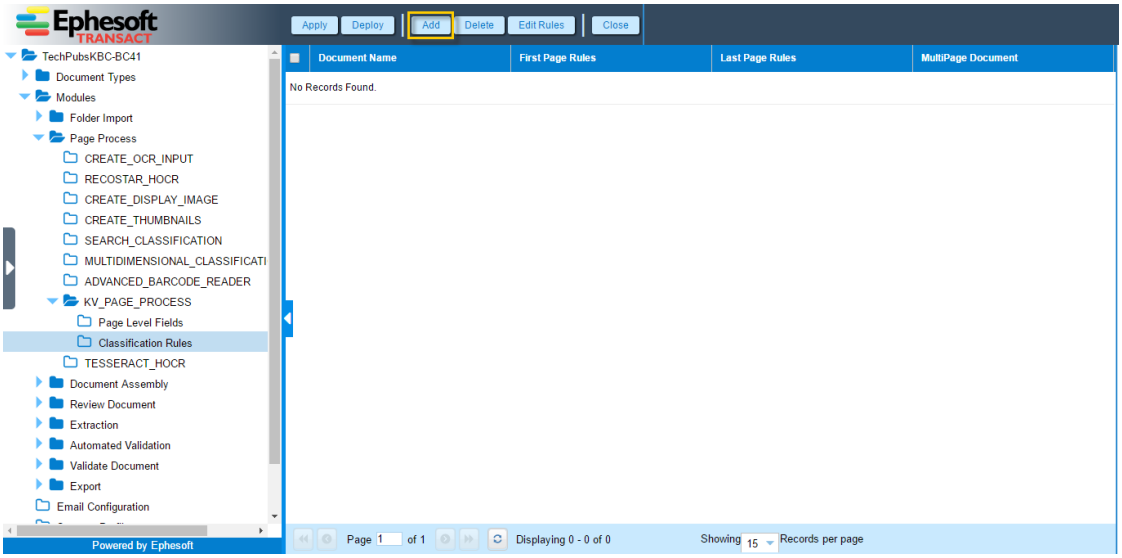
3. Select the batch class in question and click **OPEN** from the toolbar on top of the **Batch Class Management** screen.

   The batch class opens with a list of document types.



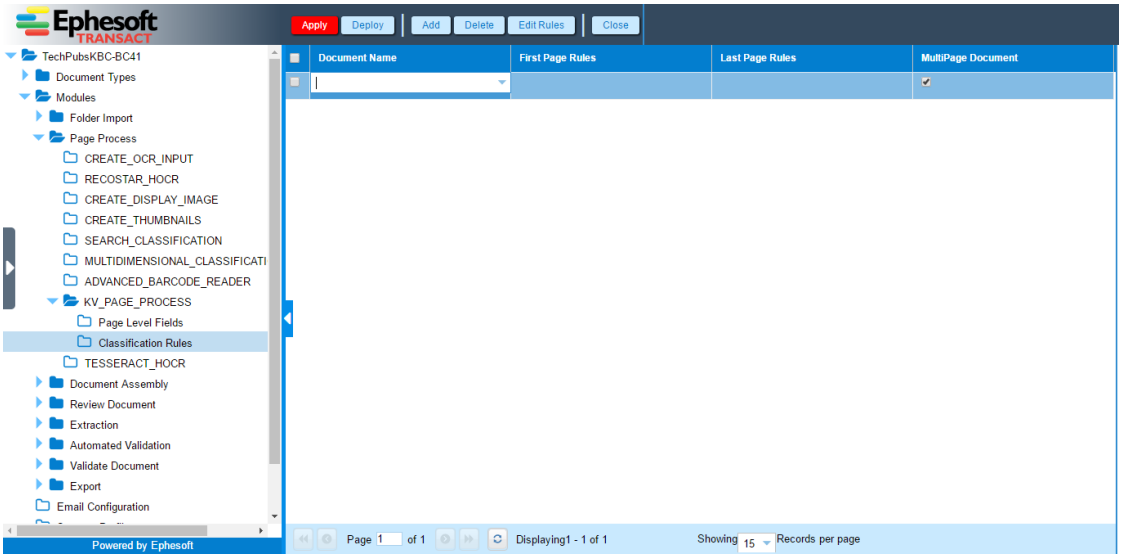
4. From the left navigation pane, go to **Modules** > **Page Process** > **KV_PAGE_PROCESS** > **Classification Rules**.

   The following screen displays.



5. Clcik **ADD** from the toolbar on top of the page.

   The following screen displays.



6. Select an exisiting document type from the **Document Name** column drop-down list.

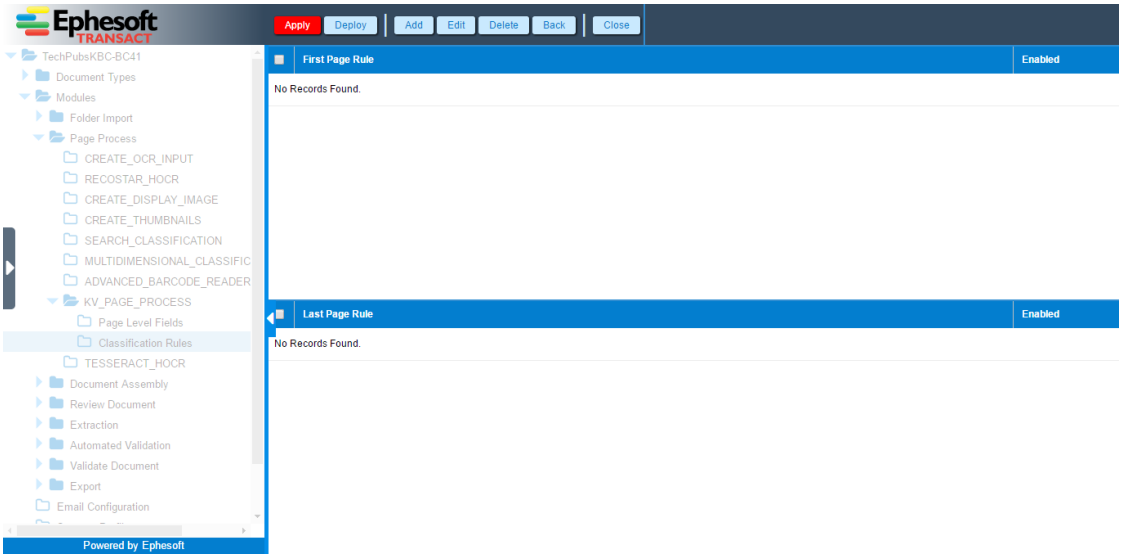
The **Multipage Document** checkbox is selected by default. Deselect it for single page document types.

The following columns are available in the grid on the **Classification Rules** page of **KV_PAGE_PROCESS** Plugin.

Column Name	Description
Document Name	Document type for which rule expressions are configured.
First Page Rules	Rules set for identification of first page of the document. Multiple rules for first page would be combined using OR operation
Last Page Rules	Rules set for identification of last page of the document. Multiple rules for first page would be combined using OR operation
MultiPage Document	This options enables user to specify whether the document classified under this document type is to be assembled as single page or multipage document. User can select/deselect the **MultiPageDocument** option from the grid on the **Classification Rules** page of **KV_PAGE_PROCESS** Plugin, but not modify the document type.

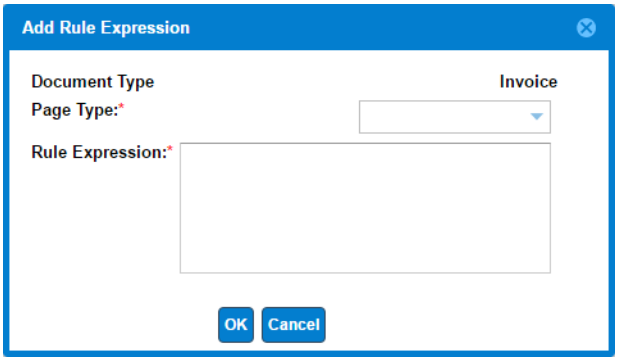
7. Click **EDIT RULES** from the toolbar on the top of the page.

   The following screen displays.



8. Click **ADD** from the toolbar on the top of the page.

   The **Add Rule Expression** dialog box displays.



9. Select page type from the **Page Type** drop-down list. The avaialbe options are **First Page** and **Last Page**.
10. Enter the desired rule expression using the auto-suggestions. The following options are available.

Operation Name	Operator	Example	Description
EQUALS	==	'Invoice No' == 1234	Value can be of type Integer or String
NOT_EQUALS	!=	'Invoice No' != 'abc'	Value can be of type Integer or String
GREATER_THAN	>	'Invoice No' > '2011-12-31'	Value can be of type Integer or Date in format yyyy-MM-dd
LESSER_THAN	<	'Invoice No' < 1234	Value can be of type Integer or Date

Operation Name	Operator	Example	Description
AND	&&	('Invoice No' > 12345) && ('Invoice No' < 23456)	Used for logical combination of expressions.
OR	||	('Company' == 'ABC') || ('Company' == 'DEF')	Used for logical combination of expressions
GREATER_THAN_OR_EQUALS	>=	'Invoice No' >= '2011-01-01'	Value can be of type Integer or Date
LESSER_THAN_OR_EQUALS	<=	'Invoice No' <= 1234	Value can be of type Integer or Date
STARTS_WITH	=^	'Invoice No' =^ 'INV'	Value can be of type String
ENDS_WITH	=$	'Company' =$ 'Ltd.'	Value can be of type String
IS_EXISTS	is exists	'Invoice No' is exists	True, if value is found for the corresponding page level field
IS_MISSING	is missing	'Invoice No' is missing	True, if value is not found for the corresponding page level field
IS_UNIQUE	Is unique	'Invoice No' is unique	True for the first occurrence of page level field in the uploaded batch, then false for all occurrences

RULE EXPRESSIONS

- Rule Expression can be any valid logical expression (that should be either true or false)
- Any rule expression can be in the form **<Page Level Field> <Operator> <Value>**
  - Page Level Field is the Field Name, defined in 'Page Level Fields'. It will be auto-suggested to users.
  - Page Level Fields name should be unique for rules to be created. For duplicate field name, an error is displayed.
  - Operator can be any valid operation from suggestions drop down
  - Value can be any combination of characters enclosed within single quotes

Valid Rules Expressions:

('Invoice No' == '123456') || ('Company' is unique)
('Invoice No' > 12345) && ('Invoice No' < 23456) && ('Company' is unique)
'Invoice No' is exists
(('Invoice No' is unique) && ('Company' is exists)) || ('Invoice No' == '123456')

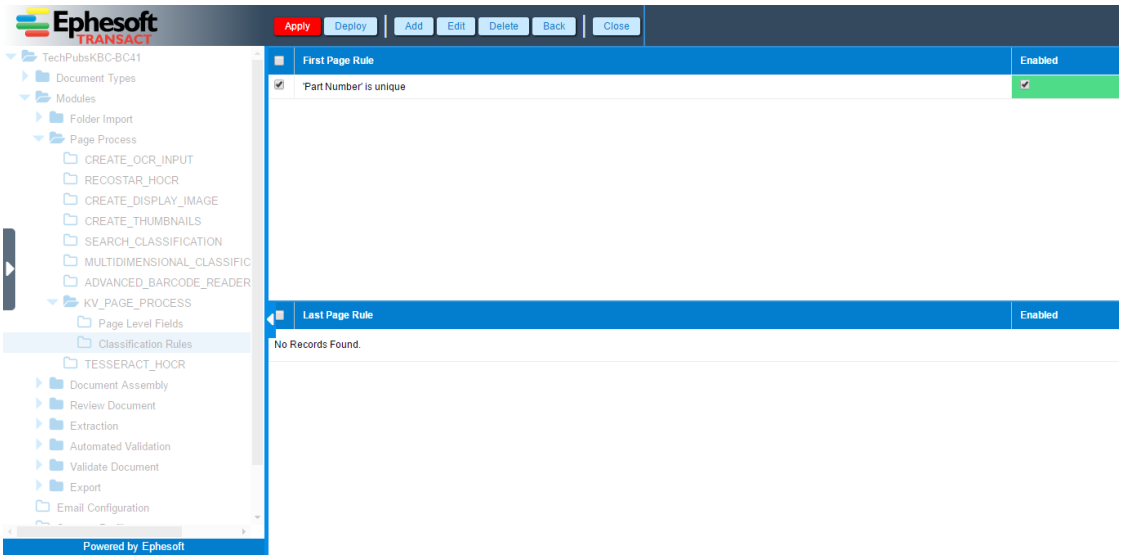
Invalid Rule Expressions:

'Invoice No' == '123456' || 'Company' is unique (Incorrect: Every expression needed to be enclosed in parentheses)
'Invoice No' =^ 123 (Incorrect: Numerical value is only applicable for operations: ==, !=, >, >=, <, <=)

('Invoice No' > 12345) && ('Invoice No' < 23456) != '123' (Incorrect: This is an invalid logical expression. Expressions can be combined using &&, || operators only)

11. Click **OK** on the **Add Rule Expression** dialog box.

    The following screen displays with the rule expression added.



12. Click **APPLY** and **DEPLOY** from the toolbar on top of the screen.

User can delete the existing document types by clicking **DELETE** from the **Classification Rules** page of **KV_PAGE_PROCESS** Plugin. On deleting document types, all rule expressions created for the document type are also deleted.

Document Assembly using Keyword Classification

Documents classified through Keyword Classification workflow can be assembled through either of the following strategies:

- **MultiPageDocument** – Document boundary start from first page to last page of same or different document type. All Unknown pages between First and Last pages of multipage document are treated as Middle Pages of the document.
- **SinglePageDocument** – Each page is classified as a separate document. If a page is classified as first or last page of any document type, it is converted into document of that type. Every Unknown page is classified into a separate **Unknown** document type.

Example 1: If Keyword Classification algorithm has classified the pages into the following order for a single-page document A and multipage document B:

- PG0: UNKNOWN
- PG1: UNKNOWN
- PG2: A_FIRST_PAGE
- PG3: B_FIRST_PAGE
- PG4: UNKNOWN
- PG5: UNKNOWN
- PG6: B_LAST_PAGE
- PG7: UNKNOWN

Result: Five individual documents will be created as

DOC1: [PG0] **(Unknown)**
DOC2: [PG1] **(Unknown)**
DOC3: [PG2] **(Document A)**
DOC4: [PG3, PG4, PG5, PG6] **(Document B)**
DOC5: [PG7] **(Unknown)**

Example 2: If Keyword Classification algorithm has classified the pages into the following order for a single-page document A and multipage document B:

- PG0: A_FIRST_PAGE
- PG1: B_FIRST_PAGE
- PG2: UNKNOWN
- PG3: B_FIRST_PAGE
- PG4: B_LAST_PAGE
- PG5: A_FIRST_PAGE
- PG6: B_LAST_PAGE

Result: Five individual documents will be created as

DOC1: [PG0] **(Document A)**
DOC2: [PG1, PG2] **(Document B)**
DOC3: [PG3, PG4] **(Document B)**
DOC4: [PG5] **(Document A)**
DOC5: [PG6] **(Document B)**

Page Confidence Calculation for Keyword Classification

Page confidence is calculated based on the following rule:

If 'x' is the number of rules matched for any page type (First/Last page), and 'y' is the total number of all rules matched configured in the batch class, then confidence can be calculated as:

$$Confidence\ for\ page\ type = \frac{Number\ of\ matched\ rules\ for\ the\ page\ type\ (x)}{Total\ number\ of\ all\ matched\ rules\ (y)}$$

If a page confidence conflicts i.e. it is same for two (or more) page types, then the confidence is reduced by number of conflicting matched page types, for example, if page confidence is 50% for two page types, then confidence will be reduced to (50/2) % = 25% and assigned to one of those page types.

Middle pages of the document is assigned the same confidence as that of first page of that document.