SparkCognition DeepNLP ™
User Guide

A SparkCognition ™ Education Document

Q2-2020
v. 2.6
04.2020

DeepNLP User Guide

Contents

About this Guide 2

Introduction to DeepNLP 3
- DNLP Installation and Access . . . 4
- DNLP GUI Interface Help . . . 4

DeepNLP Actions 4
- Actions Overview . . . 4
- Using DeepNLP Actions . . . 5

Ingesting Data 8
- Permissions . . . 8

Enriching Content 8
- Enriching Content - Overview . . . 9
- Collections . . . 10

DeepNLP User Roles 10
- DeepNLP Integrator (Superuser) . . . 10
- DeepNLP Creator (SME) . . . 11
- DeepNLP Explorer (User) . . . 12

DeepNLP Activities by Role 13
- DeepNLP Integrator/Creator Activities . . . 13
  - Creating and Viewing Tasks . . . 13
  - Creating and Viewing Pipelines . . . 16
  - Creating Collections . . . 18
  - Preprocessing Data . . . 22
  - Enriching Collections . . . 25
  - Manually Adding/Editing Extractions or Categories . . . 26
- DeepNLP User Activities . . . 27
  - Displaying Collection Results . . . 27
  - Creating New Collections . . . 29

Exploring Collections 31

Document View 31

Contact Support **33**

Reference **33**

Methods to Enhance Accuracy 33

DeepNLP Actions Reference 34

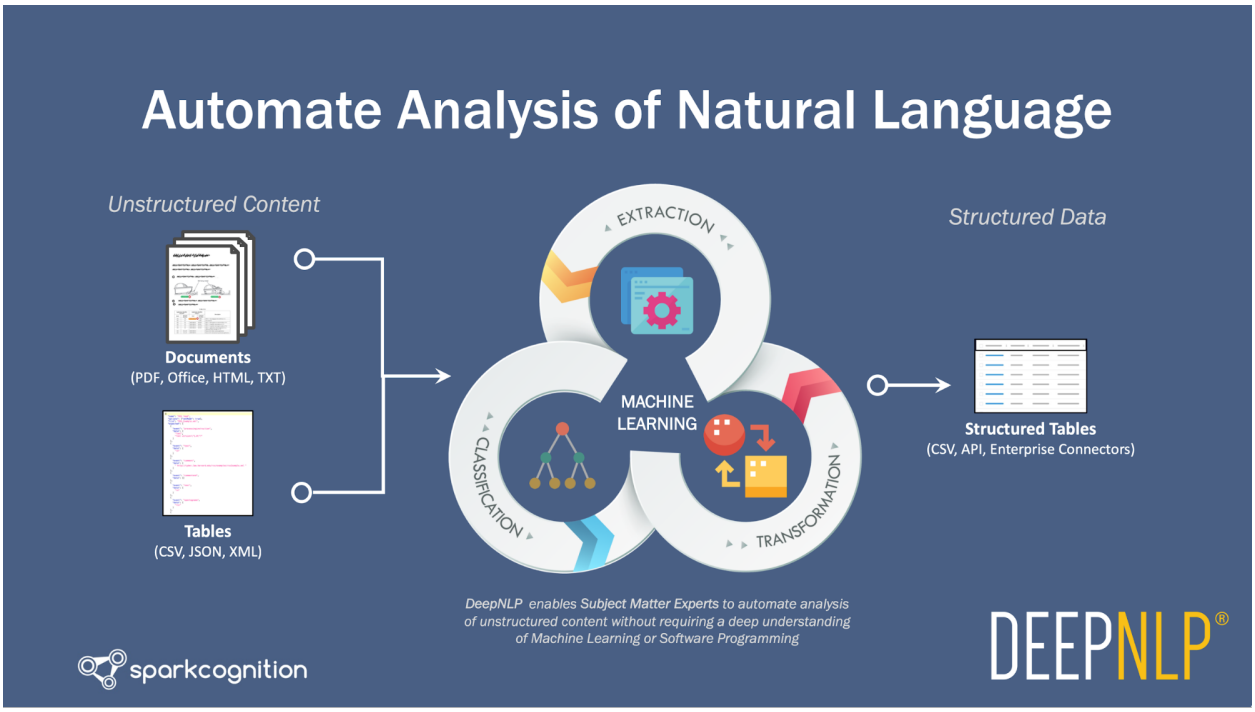
About this Guide

This guide describes the SparkCognition™ web-based GUI User Interface for the DeepNLP™ service and its use. It is aimed at the information worker who is comfortable working with spreadsheets, modern productivity and analytics tools. It is not necessary to be a data scientist or programmer to effectively use the GUI. Note that to be productive with DeepNLP, moderate training and orientation is required.

The documentation for SparkCognition DeepNLP includes the following documents available from the SparkCognition DeepNLP Support Portal:

- This guide
- The *SparkCognition DeepNLP Installation Guide*
- The *SparkCognition DeepNLP API Guide*
- The *SparkCognition DeepNLP Integrator Guide*
- The *DeepNLP Release Notes*

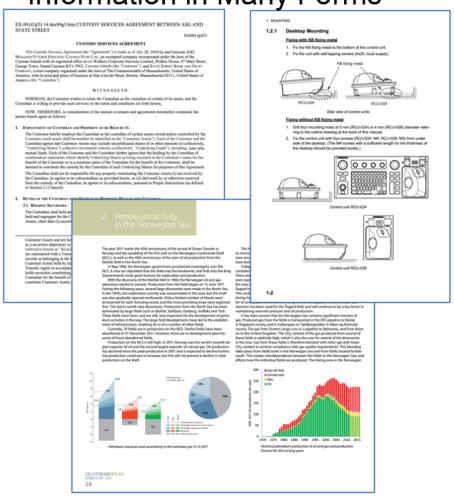
Introduction to DeepNLP



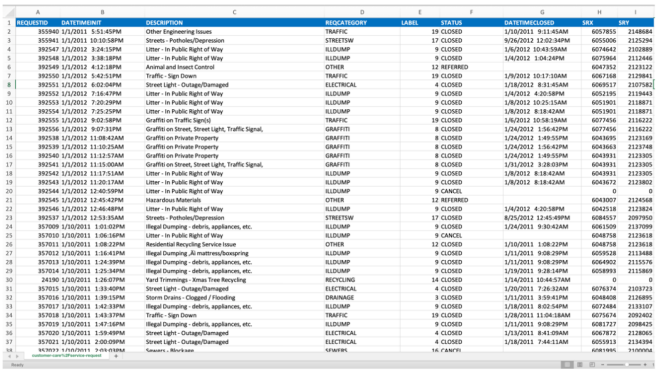
DeepNLP is a content enrichment pipeline that enables SMEs (Subject Matter Experts) to automate analysis of natural language content.

DeepNLP can read natural language content directly from documents or from tables such as customer support tickets, emails, and maintenance records exported from a database. It enables SMEs to automate extraction and/or categorization of information into a structured data set. This means SMEs can take information in various forms and use DeepNLP to generate structured data:

Information in Many Forms



Structured/Tabular Data



	A	B	C	D	E	F	G	H	I
1	REQUESTID	DATETIMEINIT	DESCRIPTION	REQCATEGORY	LABEL	STATUS	DATETIMECLOSED	SRX	SRY
2	355940	1/1/2011 5:51:45PM	Other Engineering Issues	TRAFFIC		19 CLOSED	1/10/2011 9:11:45AM	6057855	2148684
3	355941	1/1/2011 10:10:58PM	Streets - Potholes/Depression	STREETSW		17 CLOSED	9/26/2012 12:02:34PM	6055006	2125294
4	392547	1/1/2012 3:24:15PM	Litter - In Public Right of Way	ILLDUMP		9 CLOSED	1/6/2012 10:43:59AM	6074642	2102889
5	392548	1/1/2012 3:38:18PM	Litter - In Public Right of Way	ILLDUMP		9 CLOSED	1/4/2012 1:04:24PM	6075964	2112446
6	392549	1/1/2012 4:12:18PM	Animal and Insect Control	OTHER		12 REFERRED		6047352	2123122
7	392550	1/1/2012 5:42:51PM	Traffic - Sign Down	TRAFFIC		19 CLOSED	1/9/2012 10:17:10AM	6067168	2129841
8	392551	1/1/2012 6:02:04PM	Street Light - Outage/Damaged	ELECTRICAL		4 CLOSED	1/18/2012 8:31:45AM	6069517	2107582
9	392552	1/1/2012 7:16:47PM	Litter - In Public Right of Way	ILLDUMP		9 CLOSED	1/4/2012 4:20:58PM	6052195	2119443
10	392553	1/1/2012 7:20:29PM	Litter - In Public Right of Way	ILLDUMP		9 CLOSED	1/8/2012 10:25:15AM	6051901	2118871
11	392554	1/1/2012 7:25:25PM	Litter - In Public Right of Way	ILLDUMP		9 CLOSED	1/8/2012 8:18:42AM	6051901	2118871
12	392555	1/1/2012 9:02:58PM	Graffiti on Traffic Sign(s)	TRAFFIC		19 CLOSED	1/6/2012 10:58:19AM	6077456	2116222
13	392556	1/1/2012 9:07:31PM	Graffiti on Street, Street Light, Traffic Signal,	GRAFFITI		8 CLOSED	1/24/2012 1:56:42PM	6077456	2116222
14	392538	1/1/2012 11:08:42AM	Graffiti on Private Property	GRAFFITI		8 CLOSED	1/24/2012 1:49:55PM	6043695	2123169
15	392539	1/1/2012 11:10:25AM	Graffiti on Private Property	GRAFFITI		8 CLOSED	1/24/2012 1:56:42PM	6043663	2123748
16	392540	1/1/2012 11:12:57AM	Graffiti on Private Property	GRAFFITI		8 CLOSED	1/24/2012 1:49:55PM	6043931	2123305
17	392541	1/1/2012 11:15:00AM	Graffiti on Street, Street Light, Traffic Signal,	GRAFFITI		8 CLOSED	1/31/2012 3:28:03PM	6043931	2123305
18	392542	1/1/2012 11:17:51AM	Litter - In Public Right of Way	ILLDUMP		9 CLOSED	1/8/2012 8:18:42AM	6043931	2123305
19	392543	1/1/2012 11:20:17AM	Litter - In Public Right of Way	ILLDUMP		9 CLOSED	1/8/2012 8:18:42AM	6043672	2123802
20	392544	1/1/2012 12:40:59PM	Litter - In Public Right of Way	ILLDUMP		9 CANCEL		0	0
21	392545	1/1/2012 12:45:42PM	Hazardous Materials	OTHER		12 REFERRED		6043007	2124568
22	392546	1/1/2012 12:46:48PM	Litter - In Public Right of Way	ILLDUMP		9 CLOSED	1/4/2012 4:20:58PM	6042518	2123824
23	392537	1/1/2012 12:53:35AM	Streets - Potholes/Depression	STREETSW		17 CLOSED	8/25/2012 12:45:49PM	6084557	2097950
24	357009	1/10/2011 1:01:02PM	Illegal Dumping - debris, appliances, etc.	ILLDUMP		9 CLOSED	1/24/2011 9:30:42AM	6061509	2137099
25	357010	1/10/2011 1:06:16PM	Litter - In Public Right of Way	ILLDUMP		9 CANCEL		6048758	2123618
26	357011	1/10/2011 1:08:22PM	Residential Recycling Service Issue	OTHER		12 CLOSED	1/10/2011 1:08:22PM	6048758	2123618
27	357012	1/10/2011 1:16:41PM	Illegal Dumping - Ãi mattress/boxspring	ILLDUMP		9 CLOSED	1/11/2011 9:08:29PM	6059528	2113488
28	357013	1/10/2011 1:24:39PM	Illegal Dumping - debris, appliances, etc.	ILLDUMP		9 CLOSED	1/11/2011 9:08:29PM	6064902	2115576
29	357014	1/10/2011 1:25:34PM	Illegal Dumping - debris, appliances, etc.	ILLDUMP		9 CLOSED	1/19/2011 9:28:14PM	6058993	2115869
30	24190	1/10/2011 1:26:07PM	Yard Trimmings - Xmas Tree Recycling	RECYCLING		14 CLOSED	1/14/2011 10:44:57AM	0	0
31	357015	1/10/2011 1:33:40PM	Street Light - Outage/Damaged	ELECTRICAL		4 CLOSED	1/20/2011 7:26:32AM	6076374	2103723
32	357016	1/10/2011 1:39:15PM	Storm Drains - Clogged / Flooding	DRAINAGE		3 CLOSED	1/11/2011 3:59:41PM	6048408	2126895
33	357017	1/10/2011 1:42:33PM	Illegal Dumping - debris, appliances, etc.	ILLDUMP		9 CLOSED	1/18/2011 8:02:54PM	6072484	2133107
34	357018	1/10/2011 1:43:37PM	Traffic - Sign Down	TRAFFIC		19 CLOSED	1/28/2011 11:04:18AM	6075674	2092402
35	357019	1/10/2011 1:47:18PM	Illegal Dumping - debris, appliances, etc.	ILLDUMP		9 CLOSED	1/11/2011 9:08:29PM	6081727	2098425
36	357020	1/10/2011 1:59:49PM	Street Light - Outage/Damaged	ELECTRICAL		4 CLOSED	1/13/2011 8:41:09AM	6067872	2128065
37	357021	1/10/2011 2:00:09PM	Street Light - Outage/Damaged	ELECTRICAL		4 CLOSED	1/18/2011 7:44:11AM	6055913	2134394
38	353939	1/10/2011 2:03:03PM	Sewers - Blockage	SEWER		16 CANCEL		6081995	2100994

When content is structured, it can be extracted as a table (CSV) or through APIs to support

various tasks, for example:

- Downstream automation
- Decision support or analytics
- Predictive modeling

DNLP Installation and Access

DeepNLP can be installed on any physical or virtual machine that meets the hardware and software requirements. DeepNLP is platform agnostic and can be deployed in public or private clouds or on-premise infrastructure. See the *SparkCognition DeepNLP Installation Guide* for additional information.

The DeepNLP User Interface can be accessed through any modern browser running on a desktop, laptop or mobile device.

DNLP GUI Interface Help

The DeepNLP browser-based GUI interface includes extensive mouse-over and pop-up help screens. The help includes information specific to the item selected as well as examples, where appropriate.

DeepNLP Actions

Actions Overview

DeepNLP has various useful and powerful actions available to the *DeepNLP Integrator*. These filter and enhance data from collections.
Records collections, in .csv format (CSV) in particular, can contain multiple columns with natural language content (*unstructured text*). Actions that offer *on Column* options enable running operations granularly on each column separately. See the Actions Reference section for additional information.

Actions in DeepNLP can be broadly categorized into four categories:

- **Automate Extraction**
  - *Extract Significant Terms* - Automatically extracts highly mentioned terms
  - *Extract Specified Terms* - Extract terms from a user specified list
  - *Train New Custom Extractor* - Train custom Extractor model on user generated examples of information extracted from Context
  - *Run Custom Extractor* - Run a previously trained custom Extractor model on new/un-enriched content
  - *Retrain Custom Extractor* - Add training on a previously trained custom Extractor with additional examples
- **Automate Categorization**
  - Suitable for Records Collections

    - *Train Custom Categorizer* on **Column** - Train Categorizer model on user supplied Categories for specified text column
    - *Run Custom Categorizer* on **Column** - Run a previously trained custom Categorizer model on new/uncategorized content
  - Suitable for Document Collections
    - *Train Custom Categorizer* on **Context** - Train Categorizer model on user supplied Categories for Context column
    - *Run Custom Categorizer* on **Context** - Run a previously trained custom Categorizer model on new/uncategorized content in Context column
- **Content Analysis**
  - Suitable for Records Collections -
    Extract Sentiment from **Column** - Automatically assign sentiment category for content and extract sentiment carrying terms from specified text column
  - Suitable for Document Collections -
    Extract Sentiment from **Context** - Automatically assign sentiment category for content and extract sentiment carrying terms from Context column
- **Advanced**
  - DeepNLP enables Data Scientists to measure performance of Extractor and Categorizer models
    - *Split Train-Test* -
      Uses text content to randomize and split the total content into train and test sets in user specified proportions
    - Compute Scores -
      Computes Accuracy, Precision, Recall and F-scores for user specified Label and Prediction columns
  - **Extract Text Patterns** -
    Extracts phrases matching user-specified *SpaCy* patterns
    Note: For patterns with Boolean arguments, specify the boolean values as strings: `"True"` or `"False"`
    For example:
    `[{'LIKE_URL': True, 'OP': '?'}]`
    becomes:

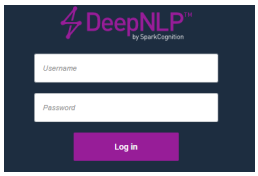
```
[{"LIKE_URL":"True", "OP":"?"}]
```

Using DeepNLP Actions

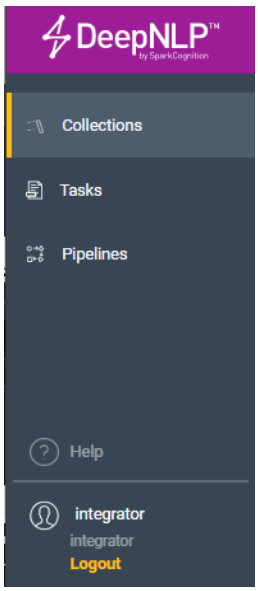
DeepNLP actions are available to the *DeepNLP Integrator* only.
In addition to manual options for adding/editing cells, DeepNLP includes actions that can operate on data at scale.

To access DeepNLP Actions:

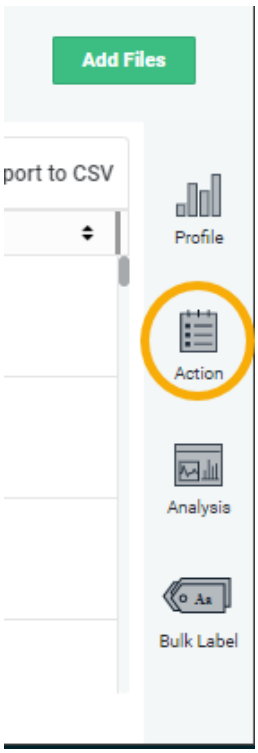
1. Log into DeepNLP as a user with *integrator* permissions



2. Select **Collections** from the left menu

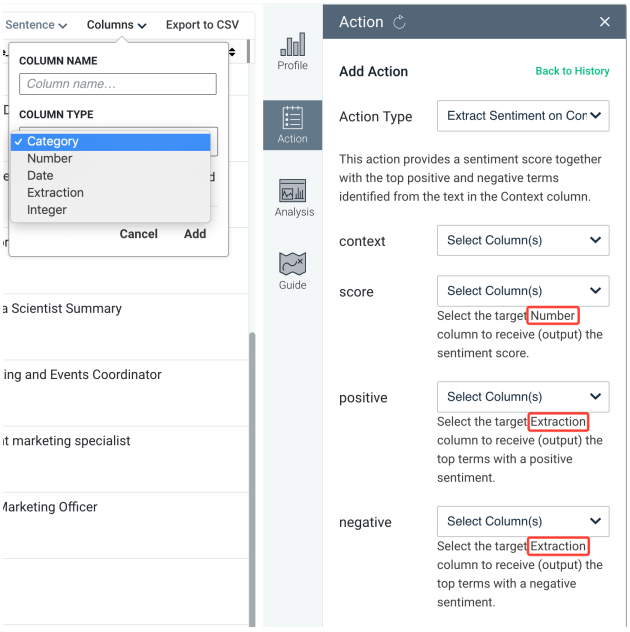


3. Click the **Action** icon in the far right pane

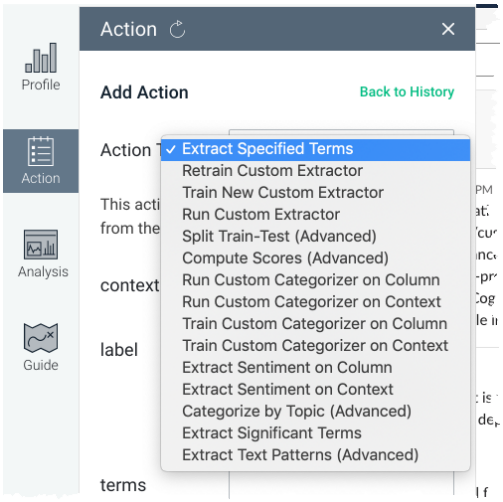


4. Use the drop-down selections menu for additional action options

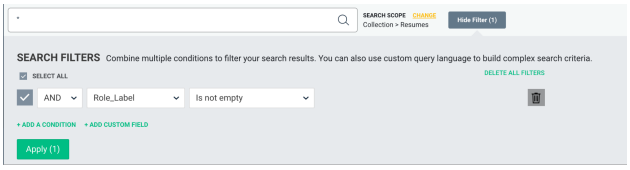
Each selectable **Action** in the drop-down describes its operation, its inputs and outputs. Use the in-line documentation to create and configure the right columns as inputs and outputs.



Note: The Actions in the drop-down menu are loaded using a plug-in framework. This framework can be extended by data scientists using Python with a moderate amount of training. The list of available actions within the drop-down list can be customized for any DeepNLP deployment.



Note: Actions triggered from the *Action* tab always operate on the filtered Collection. Filters can be configured from the *Hide/Show Filter* control.



The number of table rows matching the current filter configuration can be inferred from the *rows*

indicator at the bottom of the *Table* view:

For example, the results might indicate:

```
Results shown: 1-500 of 3233 rows
```

Note that any triggered action will operate on all *3233* rows and the number of rows currently loaded into the *Table* is immaterial for **Actions**.

Ingesting Data

Ingesting files is the process that enables DeepNLP to access and analyze data within the files.

Ingesting data includes some caveats that are important in the DeepNLP preparation process:

- During DeepNLP installation, a *user* is specified to which DeepNLP is installed:
  - `deepnlp_user` (or `(${DEEPNLP_USER}`) is the variable that represents the user
  - `USERNAME` is a variable that points the same user

- The DeepNLP installation directory is: `~deepnlp_user/deepnlp-docker`

- The datasets folder is: `~deepnlp_user/deepnlp-docker/datasets`

- Documents within the datasets folder are referred to using the `file:///datasets` prefix

Permissions

DeepNLP users and directories require appropriate permissions to function correctly:

- Ensure permissions on the directories leading up to and including the `/datasets` directory can be run as a regular user instead of requiring *root* or *sudo* authority
- It is recommended that you enable the `/datasets` folder as readable by more than a single user, for example, to enable a *group* to read the directory, run the following as *root*:

```
# chmod g+r ~deepnlp_user/deepnlp-docker/datasets
```

Enriching Content

The process of *content enrichment* is a task exclusive to the *DNLP Integrator*. Natural language content (or broadly unstructured data) is generally unsuitable for automated analysis or software based business process automation. Even though unstructured data constitutes ~80% of data generated in enterprises, human effort is required to extract and distill information from it.

Enriching content is the process of deriving analyzable information from *Natural Language Content* to make it suitable for traditional process automation of analytics.

Content Enrichment broadly involves two types of actions:

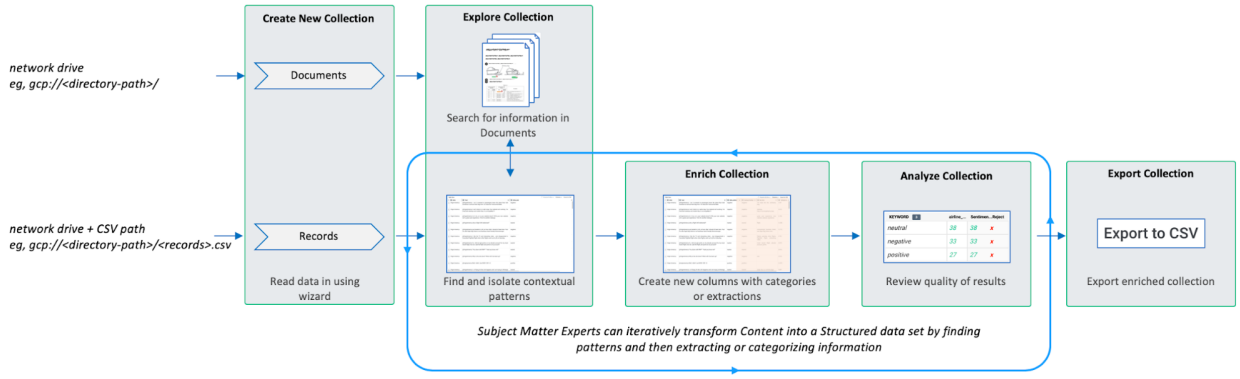
- Extraction of information (for example: contract parties or invoiced items) from content
- Categorization of information in the content (for example: language and/or sentiment)

Arbitrary human decisions about content can also be framed as categorization problems if all the decision inputs are available in the content.

Automating the content enrichment process reduces human effort required to understand, analyze and act on natural language content. Predefined specifications called *Content Enrichment Pipelines* can be used to help enable the specificity and accuracy of the process. In this document the word **pipelines** refers to *pipelines used to enrich content*.

Enriching Content - Overview

The following graphic shows the DeepNLP *Pipeline to enrich content*:



Note: In this example, `gcp://` refers to a *Google Cloud Platform* location. Other path specifications are supported, such as `gs://`, that denotes *Google Storage* locations.

DeepNLP provides wizards to read in natural language content from documents or tables into a *Collection*.

Tables containing one or more columns of natural language content can be meaningfully enriched in DeepNLP. A single row from such tables is generically referred to as a *record*.

DeepNLP generalizes both *Document* and *Record Collections* into a table with short text segments for the purposes of enrichment.

The process of *Content Enrichment* generally includes the following steps:

1. Group common content patterns that contain information of interest
2. Iteratively extract or categorize the information using patterns or *Machine Learning Models*
3. Create and populate new columns in the table using *Extracted* or *Categorized* information.
   Note: Extracted information can be additionally transformed into structured data, like *numbers* or *dates*.

4. Export the data. When the enriched content meets the standards of the downstream application, the structured data can be exported either as a CSV file or through an API interaction

Collections

Collections are a central component of the DeepNLP workflow pipeline.

A Collection is a group of logically similar documents or records, where:

- **Documents** - richly formatted natural language content that can contain such things as *Sections*, *Titles*, *Paragraphs*, *Notes*, *Images*, and *Tables*. Typically, these are saved as PDFs, Microsoft Office documents (Word, Excel, PowerPoint), TXT files, or RTF files. For example, a collection of invoices, a collection of resumes, or a collection of contracts.
- **Tables** - clusters of related information saved as a set of key-value pairs. These are typically saved as database records and imported into DeepNLP as a CSV table. For example, a CSV file exported from a database of customer support tickets, where each row is one customer interaction.

DeepNLP User Roles

DeepNLP provides different roles through its web-based GUI:

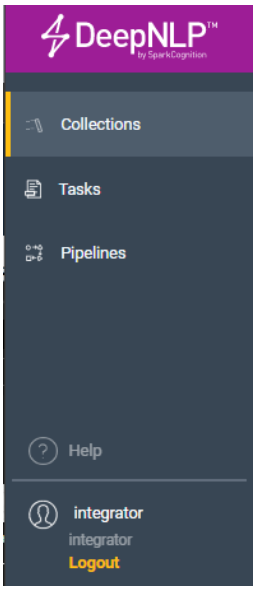
- DeepNLP *Integrator* - superuser
- DeepNLP *Creator* - subject matter expert (SME)
- DeepNLP *Explorer* - standard user

The roles are accessed via specific login to the DeepNLP GUI. Each role provides a different view into the product and defines a different set of capabilities. The menu options and displayed information vary accordingly.

DeepNLP Integrator (Superuser)

The *DeepNLP Integrator* (or *superuser*) role enables all administrative functions, including management for the content enrichment process.

The navigation menu for the *DeepNLP Integrator* includes various links:



Capabilities include:

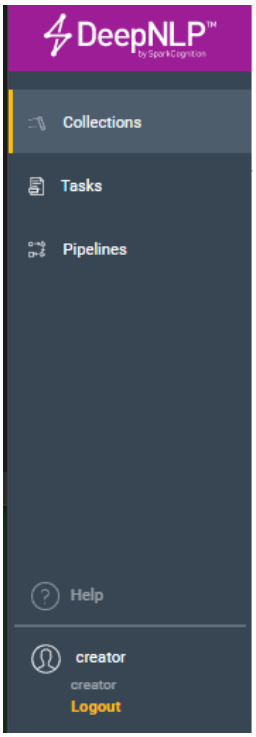
- Creating predefined collections - either from pipeline output or not
- Creating Tasks
- Creating and viewing pipelines
- Managing how content is enriched, including application of bulk changes to labels

This enables the *DeepNLP Integrator* to manipulate and enhance data analysis for collections. In general, the collections or pipelines are designed for consumption by the *DeepNLP Explorer*.

DeepNLP Creator (SME)

The *DeepNLP Creator* (or *SME*) role enables a subset of administrative functions, excluding manging how content is enriched.

The navigation menu for the *DeepNLP Creator* includes the following tabs:



The *DeepNLP Creator* (or *SME*) role enables:

- Creating and viewing pipelines
- Creating Tasks
- Creating predefined collections - either from pipeline output or not

This enables the *DeepNLP Creator* to manipulate and enhance data analysis for collections. In general, the collections or pipelines are designed for consumption by the *DeepNLP Explorer*.

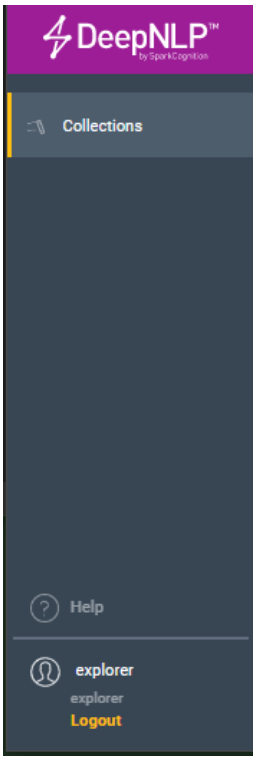
The *DeepNLP SME* general work flow includes:

1. Creating a New Collection
2. Providing Basic Set Up for the collection - including selecting appropriate collection *data type(s)*, specifying a *name* for the collection and providing *descriptive text*
3. Adding Files to the collection - upload local files or specify a Google location
4. Preprocessing Data - specify column types to enhance data extraction and accuracy

DeepNLP Explorer (User)

The *DeepNLP Explorer* is permitted a subset of the abilities of the *DeepNLP SME*. The DeepNLP Explorer employs predefined pipelines or collections to create new collections, analyze and enhance data extraction, train models, and browse results of their queries.

The navigation menu for the *DeepNLP Explorer* includes the *Collections* tab:



The *DeepNLP Explorer* general work flow includes:

1. Creating a New Collection
2. Choosing a predefined pipeline to apply to the collection
3. Providing a *Collection Name* and reviewing the summary about how the pipeline works
4. Adding Files for the collection - upload local files or specify a Google location
5. Performing the analysis
6. Reviewing/Browsing the results
7. Viewing Collection Analysis Summary

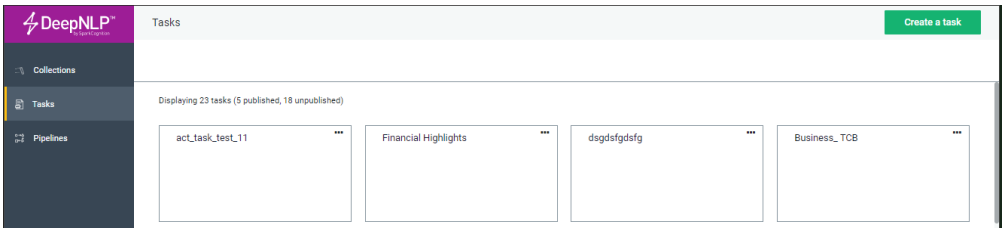
DeepNLP Activities by Role

DeepNLP Integrator/Creator Activities

The following section describes the various activities of the DeepNLP *Integrator* and *Creator*. Activities for the DeepNLP User role are described in *DeepNLP User Activities*.
Note: The DeepNLP Integrator role and DeepNLP Creator role share all abilities and activities except *managing enrichment of content*. Managing content enrichment is solely a DeepNLP Integrator task.

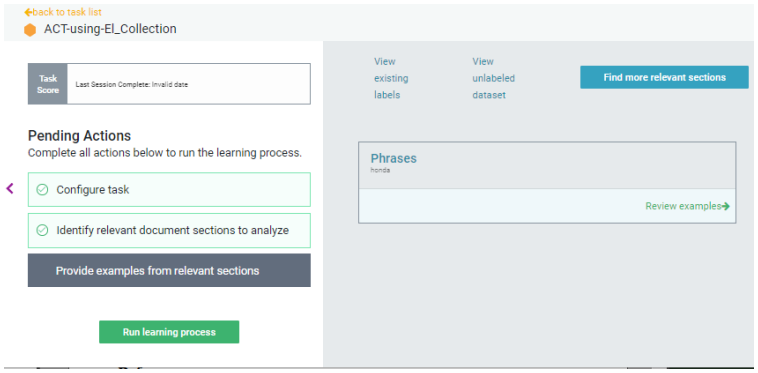
Creating and Viewing Tasks

To view the *Tasks* pane, select **Tasks** from the left menu to display the existing task list in the right pane



View Task Information To display information about an existing (pre-defined) task:

1. Click **Tasks** in the left pane to display the *Tasks Pane*. This pane shows all defined tasks and provides the **Create a Task** button
2. Select a task to view
3. Click the *task name* to display information for that task, for example:



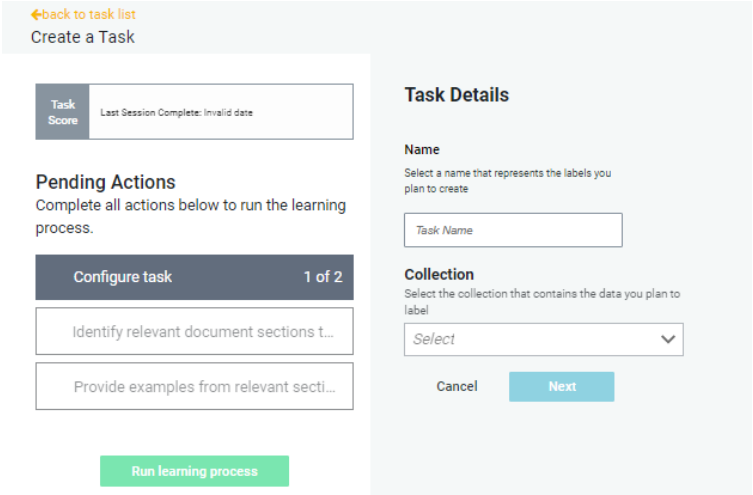
The information includes previous scores and a list of pending actions where a check mark indicated=s completion. Available options include:

- View current settings
- **Run Learning Process**
- **Find more relevant sections**
- **Review examples**

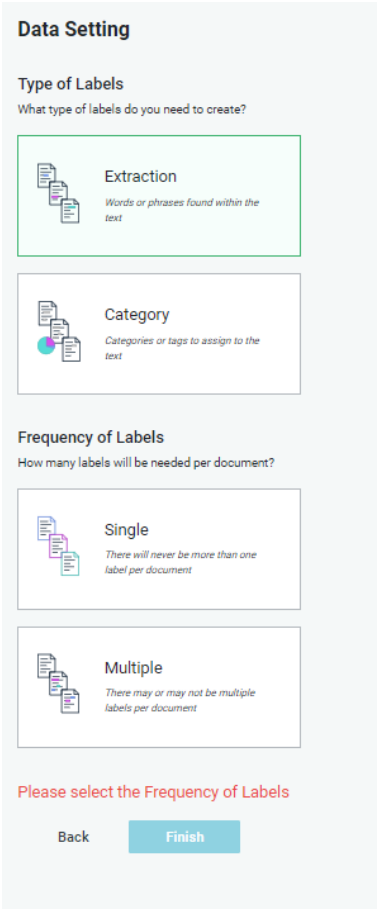
Creating a New Task

To create a new task

1. Click the *Create a Task* button display the create a task dialog:

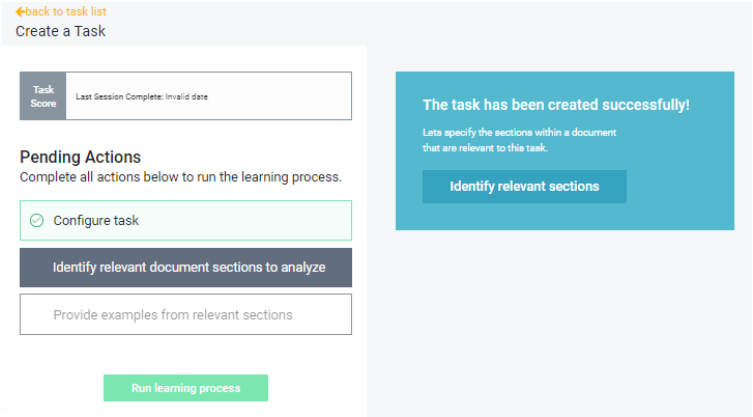


2. Name the task
3. Use the drop-down menu to select a collection as a target for the new task
   Note: the *Pending Actions* section shows the progress of the task creation.
4. Click **Next** to display the *Data Setting* dialog



5. Chose the *Type of Labels* - **Extraction** or **Category**

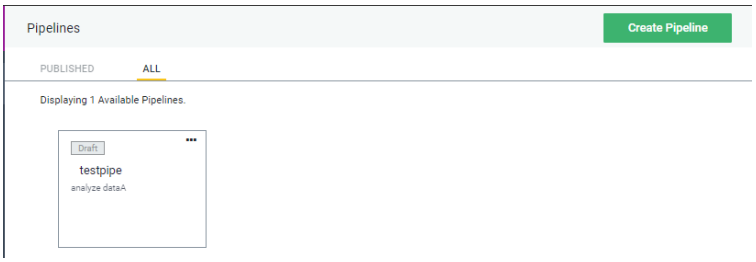
6. Chose the *Frequency of Labels* - **Single** or **Multiple**
7. Click **Finish** to display the **Success** screen



8. Consult the *Pending Actions* menu to see that 2 actions remain:
   - *Identify relevant sections...*
   - *Provide examples ...*
9. Complete each action to activate the **Run learning process** button and view your results

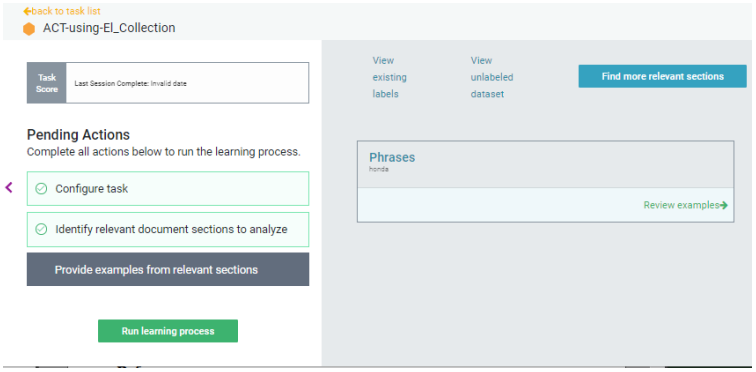
Creating and Viewing Pipelines

To view the *Pipelines* pane, select **Pipelines** from the left menu to display the existing pipeline list in the right pane. This pane shows all defined pipelines and provides the **Create Pipeline** button



Displaying a Pipeline

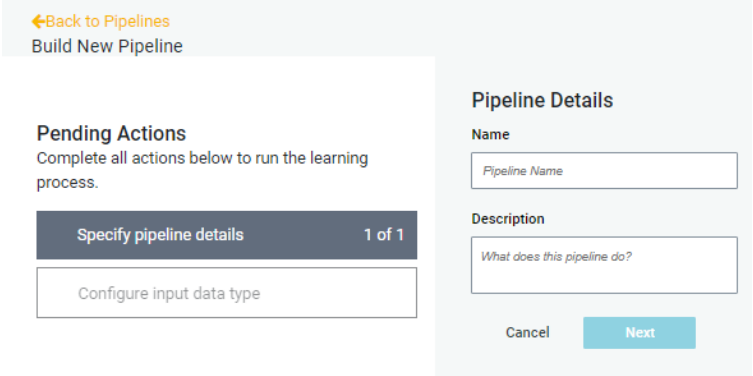
To display information about an existing (pre-defined) Pipeline, click a *pipeline name* to display information for that pipeline



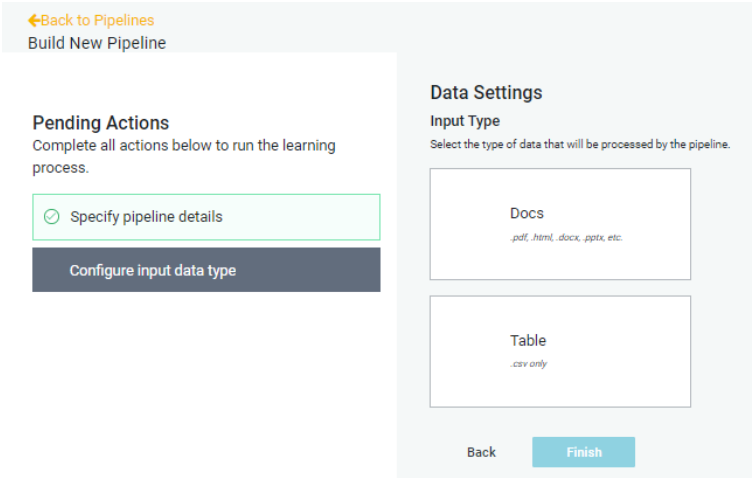
Creating a New Pipeline

To create a new pipeline

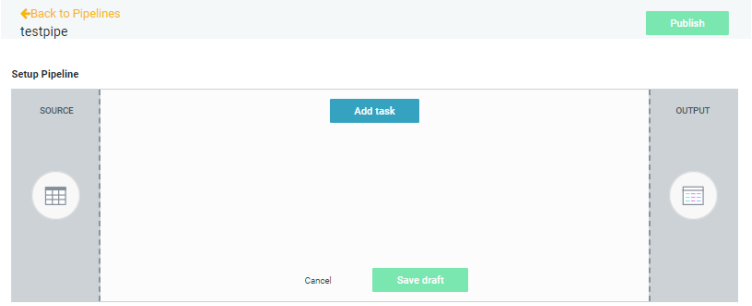
1. Click the *Create Pipeline* button display the *Build New Pipeline* dialog:



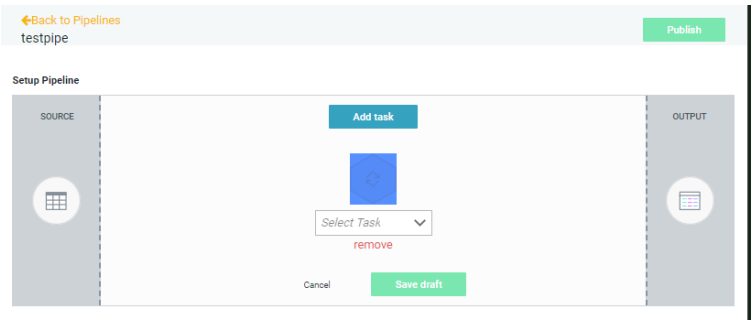
2. Provide a *Name* and a *Description* for your pipeline
3. Click **Next** to display the *Data Settings* dialog.
   Note: The *Pending Actions* list shows the list of required actions in sequence and the creation progress.



4. Click to select the *Input type - Docs* or *Table*
5. Click **Finish** to display the *Add Task* dialog to build the pipeline



6. Click **Add Task**
7. Select a Task from the drop-down menu of defined tasks

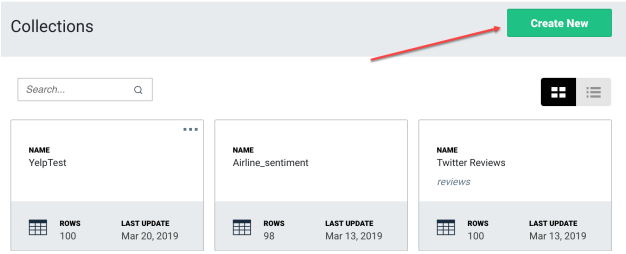


8. Repeat until the appropriate tasks are added
9. Use the drop-down menu to select a collection as a target for the new task
   Note: the *Pending Actions* section shows the progress of the task creation.
10. To publish the completed task, click the **Publish** button

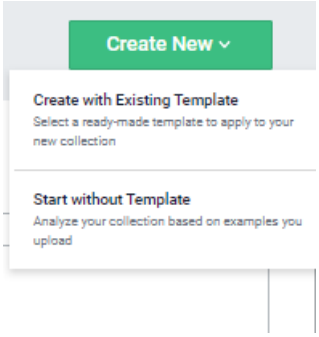
Creating Collections

The following procedure describes the process of creating a collection of documents so they can be ingested and analyzed by DeepNLP.

1. If necessary, copy the documents to the `/datasets` directory
2. Point a browser to the DeepNLP URL to access the user interface
3. Log into DeepNLP
4. Click **Collections** in the left pane to display the *Collections pane*. This pane shows all defined collections and provides the **Create New** (collection) button
5. Click the **Create New** button in the upper right hand corner of the window.



The *Create New Collection* drop down menu displays:

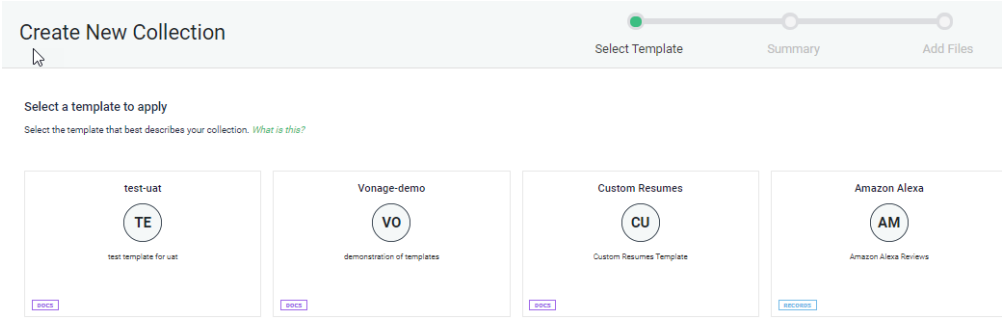


6. Select between the 2 options to display the associated pane:

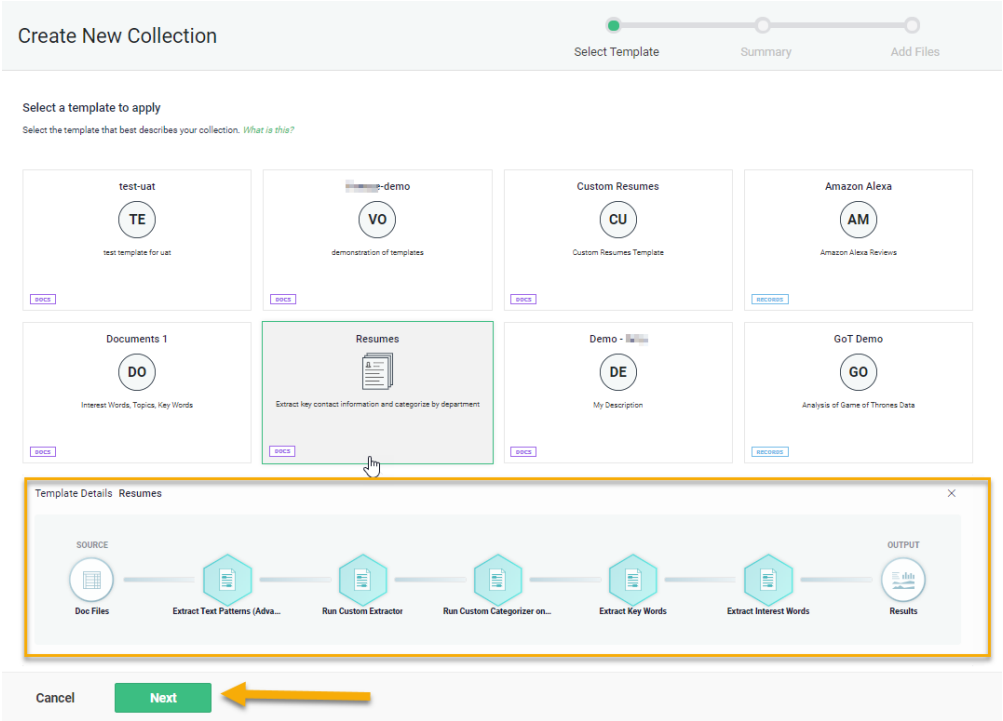
   - *Create with Existing pipeline*
     OR
   - *Start without pipeline*

Create with Existing Pipeline

a. Select *Create with Existing Pipeline* to display the *Select Pipeline* pane:



b. Select an existing pipeline to use, for example:

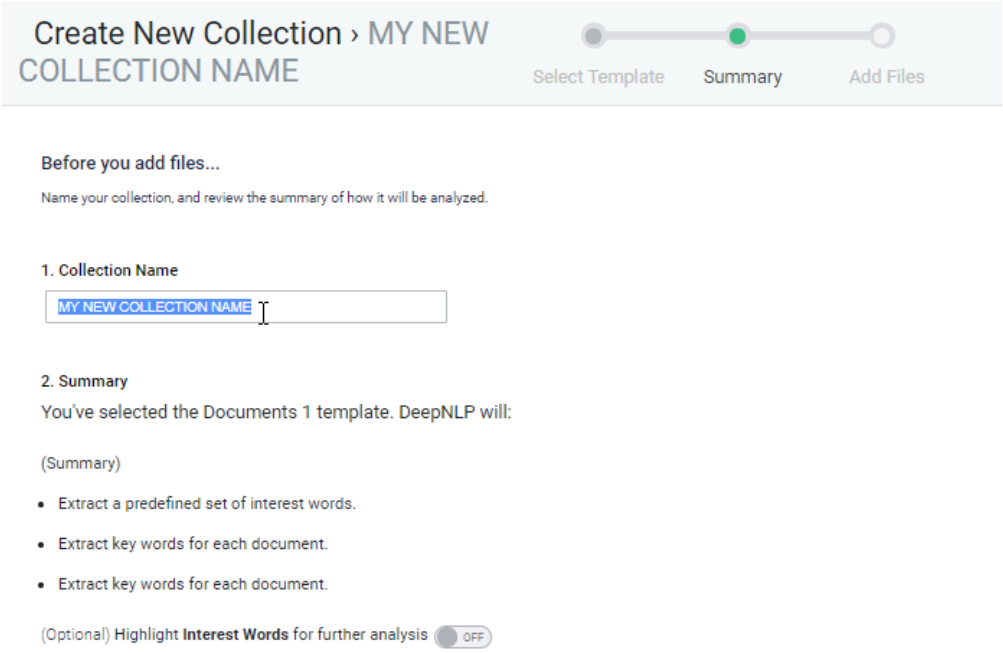


Note the pipeline progress bar (highlighted in yellow, above) displays to indicate the details of the selected pipeline. The icons on the pipeline progress bar also indicate the relative location within the creation process as the process continues.

c. Click **Next** to display the *Create New Collection* dialog, or **Cancel** to abandon the new collection

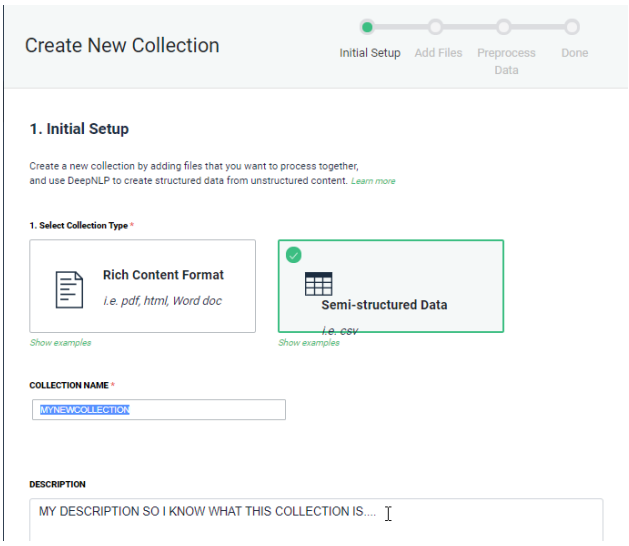
d. In the **Create New Collection** pane:

   a. Enter a *Collection Name*
   b. Verify the *Summary*
   c. Toggle the option to choose whether to *Highlight Interest Words*
   d. Click **Next** to proceed (or **Back** to choose a different pipeline



Start without Pipeline

a. Select *Start without Pipeline* to display the *Create New Collection- Initial Setup* pane:



b. Click a **Collection Type** to tell DeepNLP what kind of data to expect in the collection, in this example, *Semi-structured Data* is selected

c. Enter a **Collection Name**

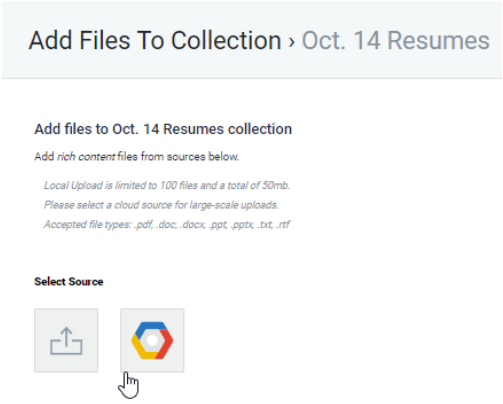
d. Enter a **Description** of the collection

e. Click **Next** to continue (or **Cancel** to abandon the new collection)

7. In the *Add files to ... collection* pane, select a source for the files. The choices include:

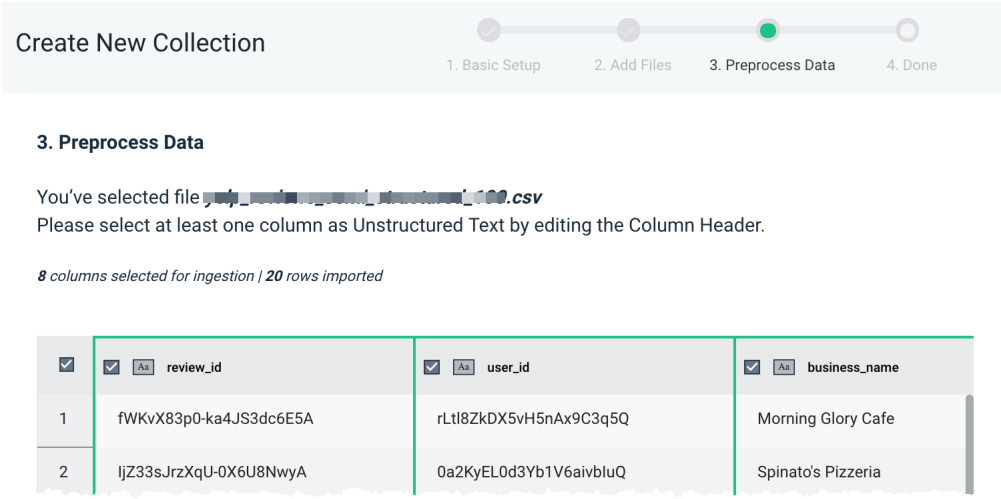
- *Local File Upload* - prompts you to browse for the directory that contains the files

- Google Cloud Platform - displays the *Folder Path* dialog to input a Google directory location, for example:
  `gs://MyGoogle_Directory`



8. Click **Finish** to continue on to *Preprocessing Data* (or **Back** to choose a different *Collection Name*)

   The *Preprocess Data* pane displays a preview table of the data.



Preprocessing Data

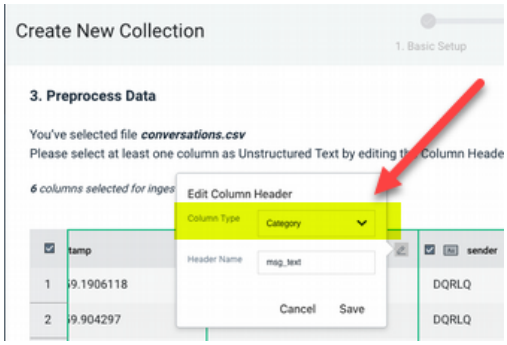
When DeepNLP ingests data, various preprocessing chores can enhance data extraction and improve accuracy. These chores include specifying data column types.

Column Data Types When ingesting a CSV it is important to define the correct column data types within the file. Verify the column type and if necessary, change the column data type definition to enhanced data interpretation.
Note: Specifying column types is normally performed in the *Preprocess Data* step during Collection creation.

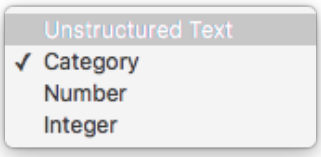
For example, if a file has a column that contains text such as chat text, ensure the column is marked correctly - in this example the column must be re-specified to be *Unstructured Text*.

Perform the following to change a column type. For example, to change the column type to the *Unstructured Text* type:

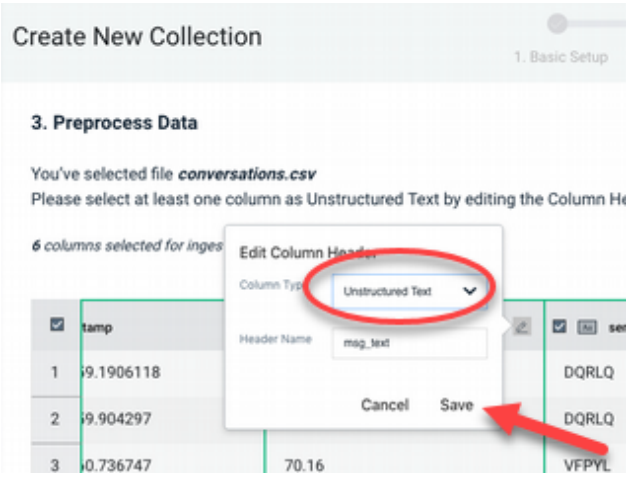
1. Click the pencil to the right of the column to display the *Edit Column Header* dialog
2. Within the dialog window, click the *Column Type* chevron to display the selection drop-down



3. Locate the appropriate option, *Unstructured Text* for example
   Note: the designation *Unstructured Text* enables DeepNLP to use any column(s) with that designation as a searchable field.
4. Select (highlight) the required option, in this case *Unstructured Text*
   ***Note**: the currently selected column type is indicated with a check mark - in this example* Category is selected before the change.

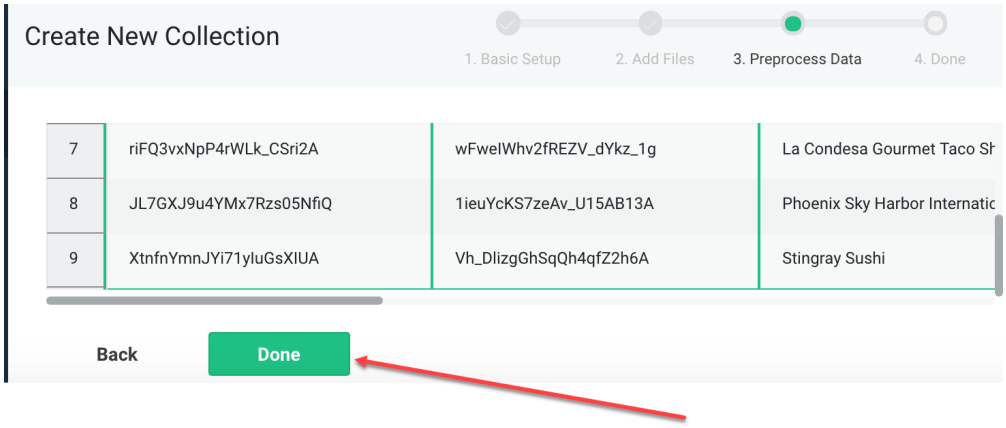


5. Verify the type is changed, for example to *Unstructured Text*
6. Click **Save** to apply the change and exit the drop-down.

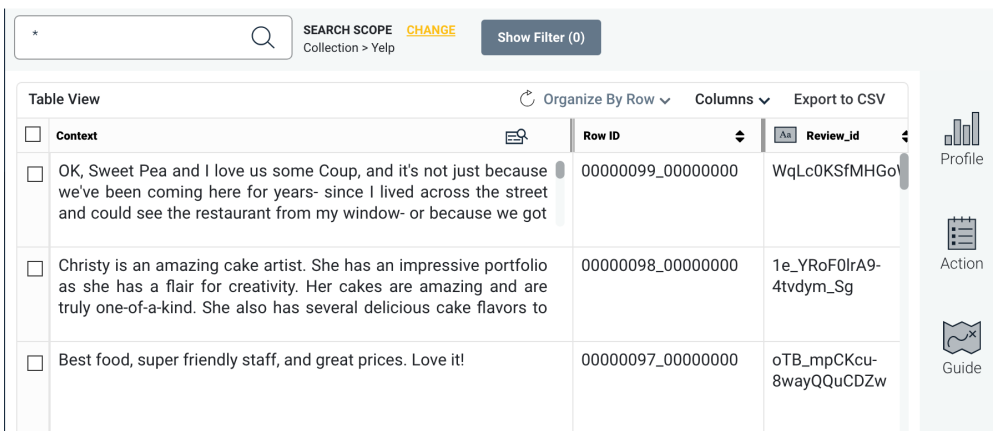


7. Click **Done** to begin DeepNLP processing (ingestion of the file). The progress of the process

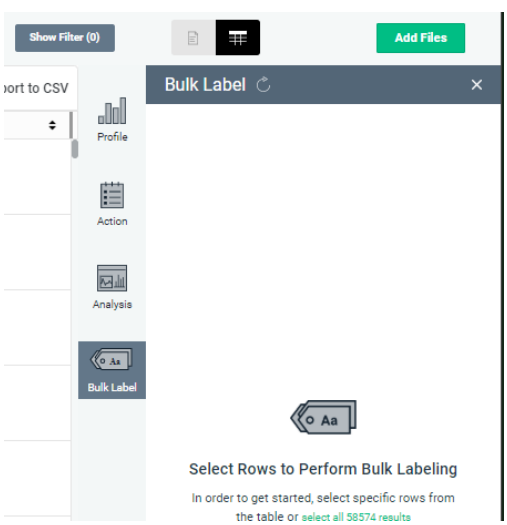
displays until processing completes



8. When processing completes, click the collection name on the main page to display the results and begin investigating the collection:



9. If it is necessary to make bulk changes to row labels, or perform bulk row labeling, click the **Bulk Label** icon to display the *Bulk Label* dialog:



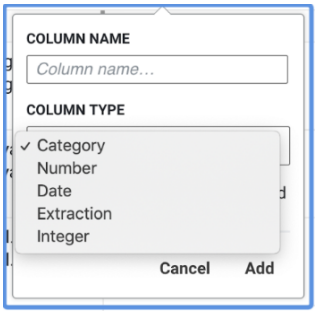
Use the dialog to make, save and apply your changes.

Enriching Collections

Enriching and managing the enrichment process is exclusive to the *DNLP Integrator*.

Enrichment is the act of creating/extracting new structured meta-data from natural language content that can be used to enhance search, generate analytics or create downstream automation.

In DeepNLP, use *Actions* to create/extract new structured data from natural language content. Typically, it is necessary to create new columns to hold/receive the new structured data.



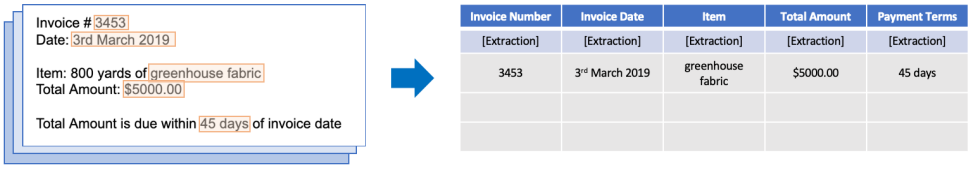
Column Types

When creating a new column to hold structured data, choose the appropriate column type:

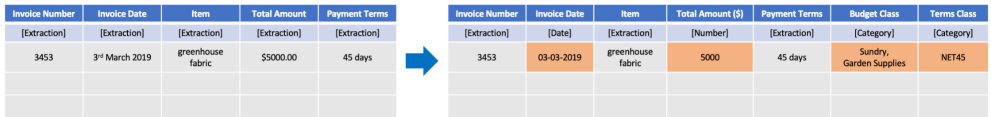
- **Extraction:**
  Extraction results in a text string that is found in the natural language content in the *Context* column. Internally, extractions track the offset of the string in the content for highlighting in documents and for model building.
- **Category:**
  Category is a text string that describes or qualifies the natural language content but does not necessarily appear in the Context column.
- **Number:**
  Number can contain positive or negative, fractional or whole numbers.
- **Date:**
  Date is stored in *MM-DD-YYYY* format.

Example To illustrate the concept of how column types are used in content enrichment, consider the following example that involves a business process for automating invoice processing:

- From a set of incoming invoices, the first round of enrichment extracts key pieces of information as *Extractions*. Note that Extractions are *text strings* that match the natural language content, verbatim.



- The second round of enrichment transforms certain Extractions into *Dates* and *Numbers*. Additionally, based on extracted data, an invoice can be categorized into *Budget and Terms*, classed as *Categories*.

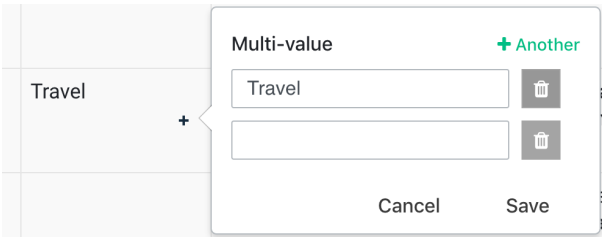


Invoice Number	Invoice Date	Item	Total Amount	Payment Terms
[Extraction]	[Extraction]	[Extraction]	[Extraction]	[Extraction]
3453	3rd March 2019	greenhouse fabric	$5000.00	45 days

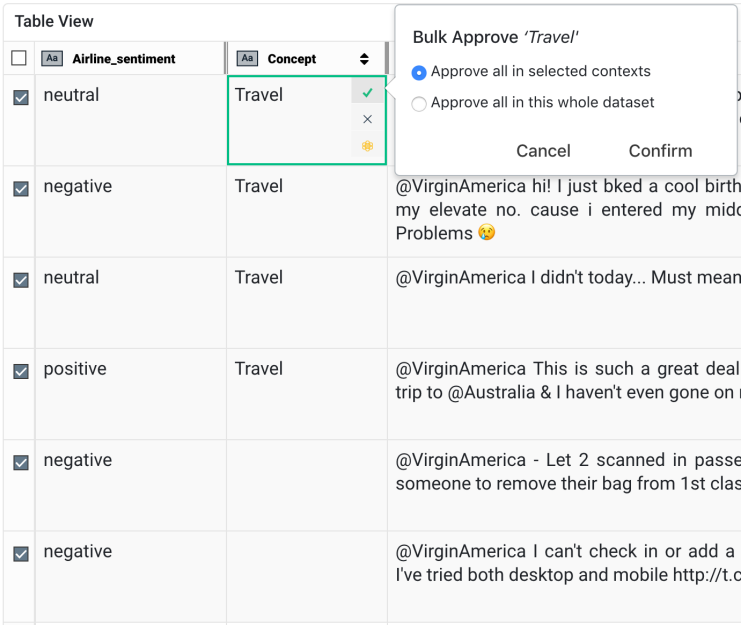
Invoice Number	Invoice Date	Item	Total Amount ($)	Payment Terms	Budget Class	Terms Class
[Extraction]	[Date]	[Extraction]	[Number]	[Extraction]	[Category]	[Category]
3453	03-03-2019	greenhouse fabric	5000	45 days	Sundry, Garden Supplies	NET45

Manually Adding/Editing Extractions or Categories

Contents of individual cells can be edited manually at any time. All cells in the table can hold multiple values, accessed using the (+) control in edit-mode.



To apply values to multiple rows, each cell presents *Bulk Approve* and *Bulk Reject* options in edit mode



Note: *Bulk Approve* and *Bulk Reject* options work differently on Extraction columns than on *Category, Date and Number* columns:

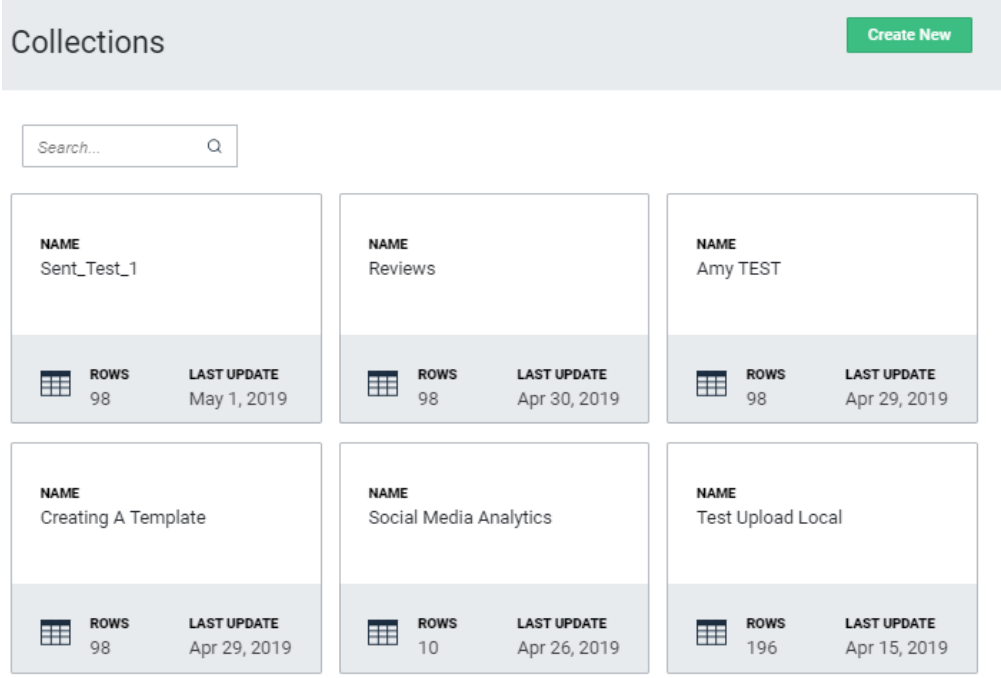
- On **Extraction** columns -
  Bulk Approve populates every in-scope row where the value in the currently selected cell appears in the Context column.
- On **Category, Number and Date** columns -

Bulk Approve populates every in-scope row with the value in the selected cell.
Where, *in-scope* is selected from the *Bulk Approve* menu to mean either all selected rows or all rows currently loaded into the Table view.

DeepNLP User Activities

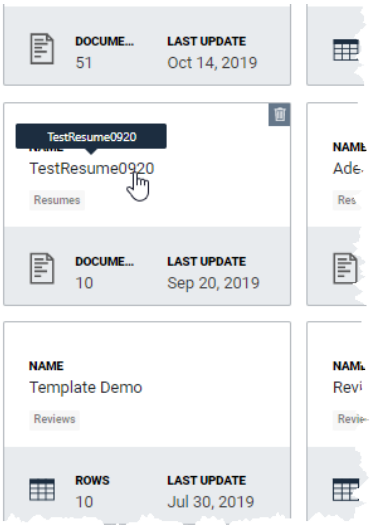
The following section describes the various activities of the DeepNLP User. Activities of the DeepNLP Creator role are defined in DeepNLP Creator Activities.

When a DeepNLP User logs into the DeepNLP browser-based GUI, the initial page displays existing collections, a *Create New* button, a Search field and a *Logout* button:

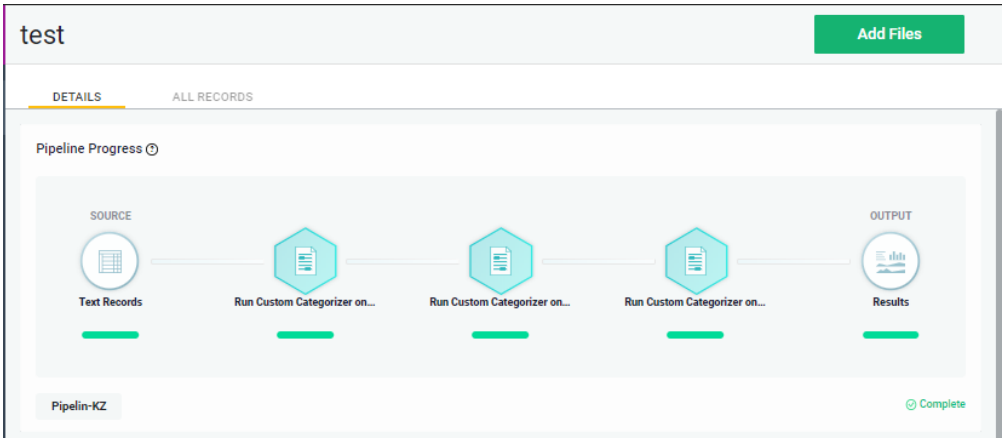


Displaying Collection Results

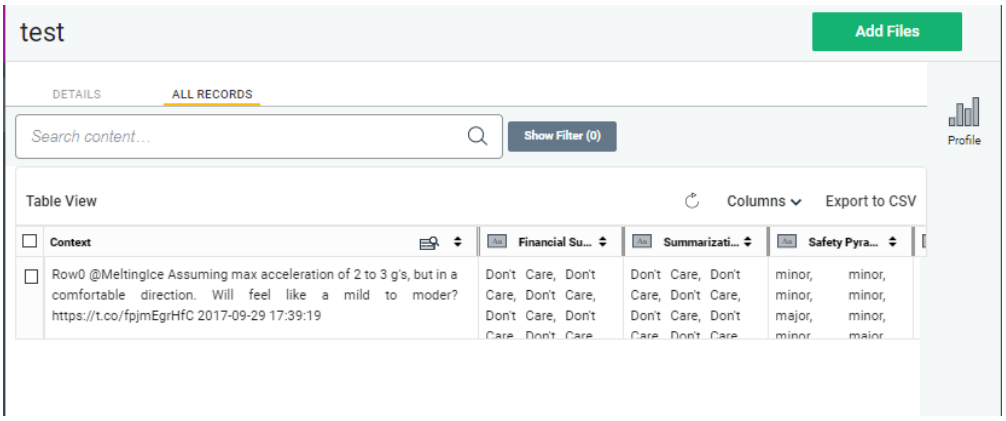
To Browse existing collection analysis results, click a collection name to display the *Details* pane



The *Details* tab displays the *Pipeline Progress* for that collection. Mousing over icons displays informational pop-ups.

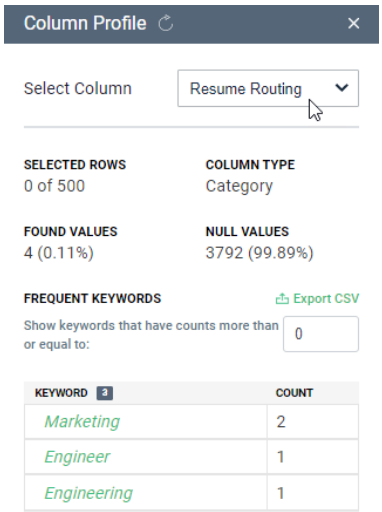


The *All Records* tab displays the table view results for that collection.



This pane enables the following actions:

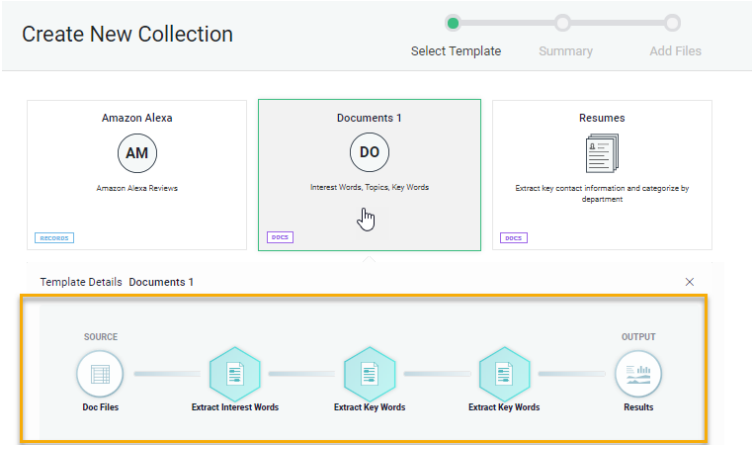
- Perform a search of the documents in the results table
- Enable/specify display filter(s)
- Click a **File Name** to display the associated document
- Choose which columns display in the results table
- Click the **chevrons** on columns to sort
- Export a *.csv formatted file* of the results
- Click **Add Files** to expand the collection with additional files
- Click the **Profile** icon to display the *Column Profile* pane. Within this pane, use the drop-down menu to select a column to learn about - in this example the column titled *Resume Routing* is selected:



Creating New Collections

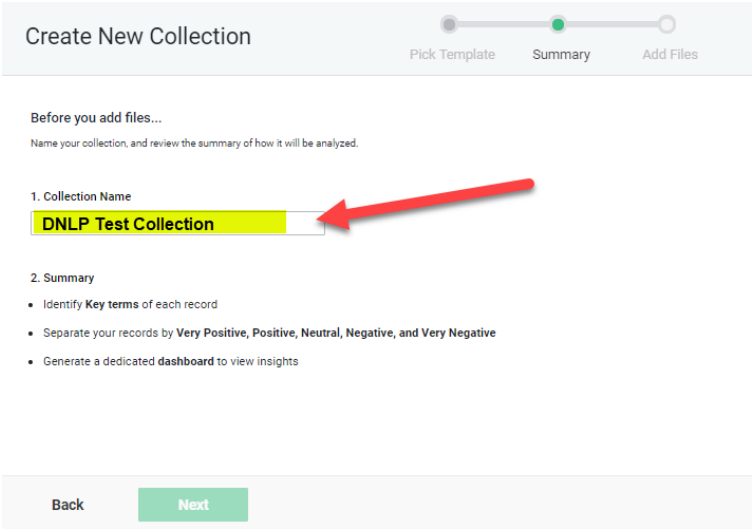
To Create a new collection to analyze with DeepNLP, perform the following:

1. Click **Create New** to display the *Select a pipeline* pane
2. Browse the available collection pipelines.
   Note: Pipelines are defined by the *DeepNLP Creator*.
3. Select and click a pipeline to use for the collection

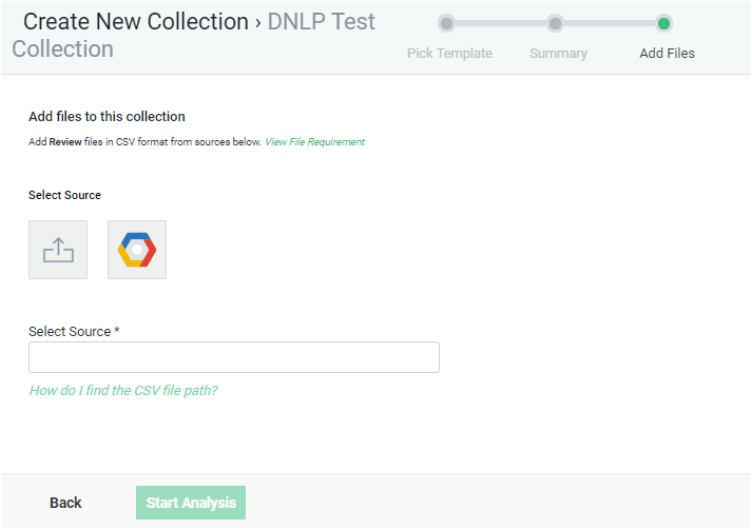


Note the pipeline progress bar (highlighted in yellow, above) displays to indicate the details of the selected pipeline. The icons on the pipeline progress bar also indicate the relative location within the creation process as the process continues.

4. Click **Next** to display the *Collection Summary* pane
5. Add a name for the new collection and review the settings the pipeline will use to analyze the data.



6. Click **Next** to display the *Add Files to collection* pane.



   a. Select a source for the collection files. The options currently are upload from your local machine or specify a Google storage path.
   b. Enter required file information in the *Selected Source* field

7. Click **Start Analysis** to start the DeepNLP analysis
8. Review/Browse the analysis results. See Exploring the Collection

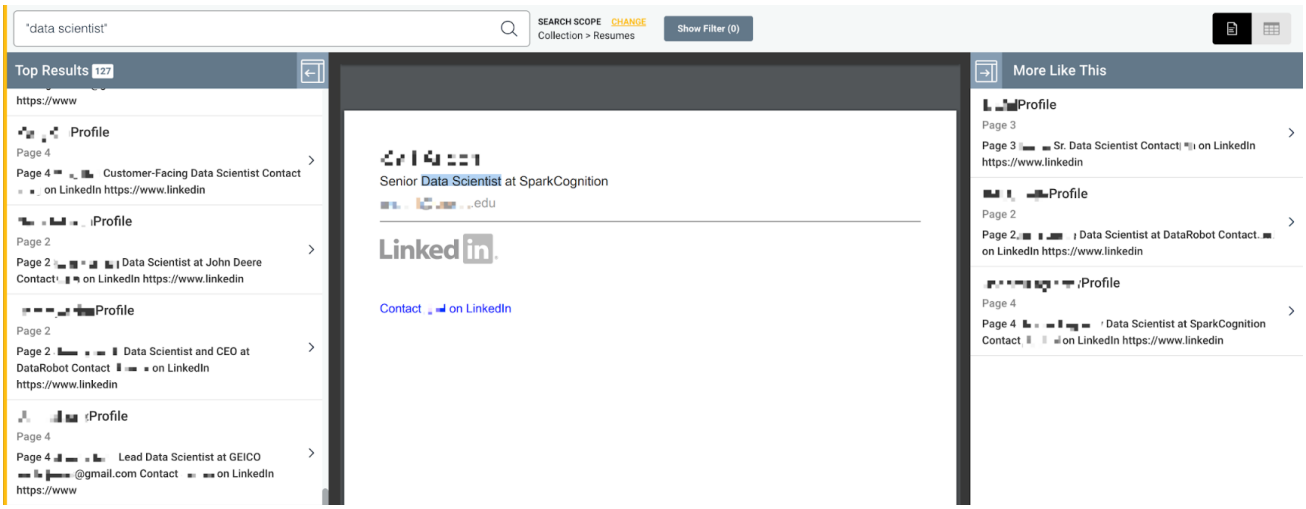
Exploring Collections

Use the *Search Bar* to query (search) the collection.

All other columns represent the properties of the content in the *Context* column and can be used as *Facets* to adjust the scope or filter the search results.

Document View

You can explore their collection using either a *faceted* or *natural language* search. The search results from within a document collection are presented in their full fidelity within the *Document View*, for example:



Note: There is nothing to display in the document view if only *semi-structured* data records (documents with tabular data) are ingested.

Table View The process of enrim> in DeepNLP begins with the Table view.

Note: This is the practical view to display if only semi-structured data records are ingested.

- **For Documents -**
  DeepNLP automatically segments (splits) documents into pages and sentences to make them suitable for enrichment. A tabular view of all segments in the collection can be used to begin enrichment in the Table view.

  Every document collection starts as a table containing the following information:

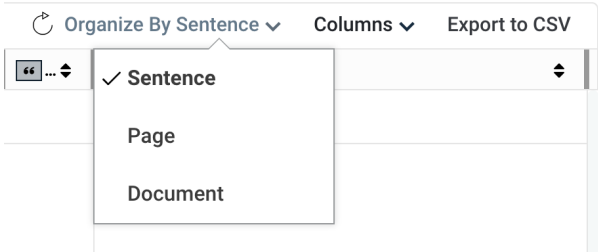
  - Sentences
  - Name + Link (to the parent document)
  - Page Number (in the parent document)
  - Row ID

- **For Records -**
  Record collections contain any columns read in from their source CSV file(s).

Context Column Context contains the unstructured natural language content that can be categorized or from which information can be extracted.

- **For Documents -**
  DeepNLP automatically assigns sentences into the *Context* column.
  Notes:

  - For *Document Collections*, you can configure the Table view to display *one Document*, *one Page* or *one Sentence* per row.
  - Currently, enrichment can only be performed for *Sentences*. Organizing the table as one *Document* or one *Page* per row is only useful for aggregation (??) and export.



- **For Records -**
  You can configure one or more columns as *Unstructured Text*. The contents of all *Unstructured Text* columns will be concatenated into the *Context* column.
  Note: You can directly search content in the *Context* column using the search bar. Search results can be refined by using filters on other columns.

Contact Support

The following resources enable you to research issues, create a support ticket, or contact SparkCognition:

- **FAQ** - Frequently Asked Questions
- Use the DeepNLP support portal - Create a support ticket and log your issue
- **Email Contact** - Send email to deepnlp_support@sparkcognition.com
- **Call Support** - The DeepNLP support line is +1-512-400-2001

Reference

Methods to Enhance Accuracy

In the case where an accuracy value is deemed low, consider the following recommendations to enhance the accuracy:

- Ensure that pdf formatted documents or documents that are the product of scanning are processed with a high-quality OCR (optical character recognition) software before DeepNLP ingestions.
- Ensure the careful assignment of documents to correct folders.
- If the documents in two or more folders are very similar, create a new folder and move those similar folders into the new folder so they become subfolders.
- If documents within a folder belong to more than one category, break each of those categories out as separate folders.
- The number of documents assigned to DeepNLP affect accuracy. As the number of assigned

documents increases, the accuracy rises. This means it is important to assign documents to folders with a lower number of documents, denoted in the *trained* column.

DeepNLP Actions Reference

The following table defines the basic actions and associated options available through the DeepNLP user interface.

Action	Description
Categorize by Topic (Advanced)	This action analyzes text in the specified Source column and assigns each record a dominant topic based on its content. Topics are a collection of words that represent a context. This action includes the ability to specify: • The target Number column to receive the dominant topic number • The Amount that specifies a value that describes the number of topics to categorize text into • The column that contains the text to analyze for topic distribution
Compute Scores (Advanced)	This action computes precision, recall and a confusion matrix for specified Label and Prediction columns. This action includes the ability to specify a Category column that contains Labels and a Category column that contains Predictions to compare.
Extract Sentiment from Column	This action provides a sentiment score together with the top positive and negative terms identified from the text in a specified Category column. This action includes the ability to specify: • The Category column that contains the text to analyze • The target Number column to receive the sentiment output score • The target Extraction column to receive the top terms with a positive sentiment output • The target Extraction column to receive the top terms with a negative sentiment • Wordiness a number that indicates the average length of text in the Context column. Suggested value is **30** for texts longer than a tweet but shorter than a paragraph.

Action	Description
Extract Sentiment from Context	This action provides a sentiment score together with the top positive and negative terms identified from the text in the Context column. This action includes the ability to specify: • The target Number column to receive (output) the sentiment score • The target Extraction column to receive (output) the top terms with a positive sentiment • The target Extraction column to receive (output) the top terms with a negative sentiment • Wordiness a number that indicates the average length of text in the Context column. Suggested value is **30** for texts longer than a tweet but shorter than a paragraph.
Extract Significant Terms	This action automatically extracts significant terms from the Context column and enables specifying the target Extraction column to receive the discovered terms as output.
Extract Specified Terms	This action automatically labels specified terms from the Context column. This action includes the ability to specify the target Extraction column to receive the specified terms output and optionally, to specify specific words or phrases to extract.
Extract Text Patterns (Advanced)	This action extracts fragments that match a specified SpaCy pattern from the text in the Context column. This action includes the ability to specify the target Extraction column to output the matched fragment(s) and the SpaCy pattern to extract.
Retrain Custom Extractor	This action retrains an extraction model to automatically extract contextual terms from text in the Context column. When training completes, run Run Custom Extractor to extract contextual terms from text. This action includes the ability to specify the target Extraction column containing the labeled data for the model to train on.

Action	Description
Run Custom Categorizer on Column	This action runs a pre-trained classification model to automatically predict categories for text in a specified Category column. This action includes the ability to specify the target Category column for the model to place the predicted categories **Note**: Train Custom Categorizer on Column must be run prior to running this action.
Run Custom Categorizer on Context	This action runs a pre-trained classification model to automatically predict categories for text in the Context column. This action includes the ability to specify the target Category column for the model to place the predicted categories. **Note**: Train Custom Categorizer on Context must be run prior to running this action.
Run Custom Extractor	This action runs a pre-trained spaCy extraction model to automatically predict labels from text in the Context column. This action includes the ability to specify the target Extraction column for the model to place the predicted labels. **Note**: Train New Custom Extractor must be run prior to running this action.
Split Train-Test (Advanced)	This action analyzes text in the specified Source column and assigns each record to train or test categories in the specified ratio. This action includes the ability to specify: • The column that contains the text to be analyzed • The target Category column to receive the train or test assignment output • The fraction of records to assign to the train category **Note**: The suggested value is **0.8** to create an 80:20 train-test split
Train Custom Categorizer on Column	This action trains a classification model to categorize text in a specified Category column. When training completes, select Run Custom Categorizer on Column to predict categories on unlabeled text. This action includes the ability to select the column that contains the text to categorize and the target Category (column) that contains the labeled categories for the model to train on.

Action	Description
Train Custom Categorizer on Context	This action trains a classification model to categorize text in the Context column. When training completes, select Run Custom Categorizer on Context to predict categories on unlabeled text. This action includes the ability to specify the target Category column that contains the labeled categories to train the model on.
Train New Custom Extractor	This action trains an extraction model to automatically extract contextual terms from text in the Context column. When training completes, run Run Custom Extractor to extract contextual terms from text. This action includes the ability to specify the target Extraction column that contains the labeled data for the model to train on. **Note**: If the model name already exists, it will be overwritten!