Analyzing your data with the Content Mining application

Use the Discovery Content Mining application to analyze your data. The application shows subsets of your information in visualizations that can help you to find patterns, trends, and anomalies.

> **Note:** Only users of installed deployments (IBM Cloud Pak for Data) or Enterprise and Premium plan managed deployments can use the Content Mining application.

Overview video

View video: [IBM Watson Discovery Content Mining](#)

Video transcript

Watson Discovery Content Mining Project presented by Stuart Strolin. (Music intro) The purpose of this video is to familiarize you with the content mining project in Watson Discovery.

Content mining is one of the primary use cases for Watson Discovery and is used for analyzing and exploring both structured and unstructured data to find insights and extract hidden meaning. It is used by both the citizen analyst and the data scientist.

The content mining project can be used for all types of analysis because the user interface is not specific to a particular industry or set of data.

In this scenario, you are an analyst for a fictitious automobile company. Operational reports have alerted the company to an unusual accident rate for one of their cars. Your job is to find out why.

Using the content mining project, you begin your analysis by looking at the unstructured data from the national motor vehicle incident reports. You are presented with an interface that allows you to select the car model and begin your analysis (on the Collections page). In this case, you are interested in the Hill Walker. You could type that information into the search section at the start of the page. But it's easier just to click on the item. You can add as many search terms and conditions as you like. But in reality, you want to let the application guide your analysis.

What you see now is the navigation view (in Guided mode). It keeps track of your analysis and provides options for next steps. It also provides a count of the number of documents that match your current state of analysis. In this small collection, the number of documents relating to the Hill Walker is only 51. In a production data set, the number would usually be much larger. Analyzing trends and anomalies is often a good way to start as it allows you to see if anything seems out of the ordinary.

Immediately, you notice that the Hill Walker has problems in December and January. You decide to investigate further by narrowing this initial exploration to just the month of December.

Notice how the navigation view at the top always keep you informed of where you are in your analysis. Next, you select *Analyze cause and characteristics* because you are interested in why things are happening.

You notice that words like 'snow' and 'brake' are highlighted together (in the Part of Speech section), so you add these to your analysis.

The Content Miner project has narrowed your investigation to a small number of complaints that can be easily read. (clicks Show Documents)

The common theme here is that there is an unexpected problem with the way the brakes are working in snowy conditions. You now have the information you need to ask the engineering department to perform a detailed inspection of the braking system and determine why it is not working as expected in snowy conditions.

In this demonstration, you saw how a citizen analyst using Watson Discovery and content mining can easily discover hidden meaning in unstructured text. (list of features, functionality, and use cases)

What will you do with Watson Discovery? (Music outro)

How it works

To analyze your data, you use *facets*. Facets give you a way to slice your data and visualize a subset of information so it is easier to

comprehend.

From the data analysis page for your collection, you can choose for the data to be shown in one of the following views:

Facets

Shows facets that are derived from from annotations that are added to your documents by enrichments that are applied to your documents. Enrichments can include built-in Natural Language Processing enrichments, such as *Part of Speech* or *Entities*. They can also include custom enrichments that you add, such as dictionaries, regular expression patterns, and machine learning models.

Metadata facets

Shows facets that are derived from your data. When you add files to a collection, Discovery analyzes and indexes the data. Annotations are added to identify content types and are shown as metadata facets. The best metadata facets result when you ingest structured data, such as records from a CSV file.

Custom

Shows only the facets that you choose to add to the view. You can add a mix of enrichment-derived and content-derived facets to your custom view.

When you create a *Content Mining* project type, the *Part of Speech* facet is applied to your data automatically. This facet is a great place to start because it is valid for all data, no matter the subject. The output gives you a quick look at the terminology that is most common in the data.
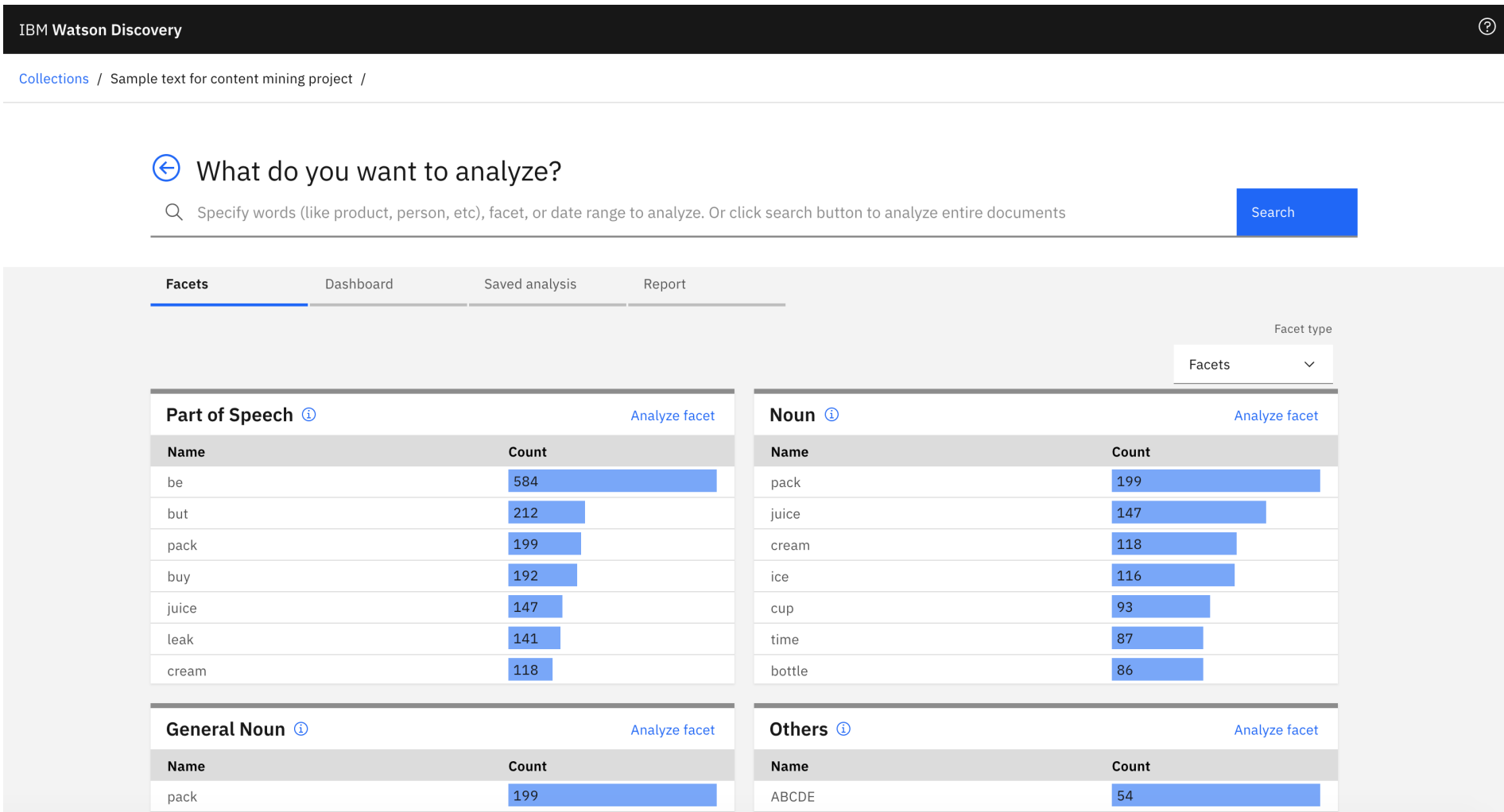

Figure 1. Watson Discovery Content Mining application home page

From this starting point, you can determine other ways to filter the data that might be useful.

If your data consists of traffic reports, for example, the *Part of Speech* facet might show that high frequency keywords include terms such as *engine*, *brake*, *fire*, *smoke*, and *spark*. Given this common terminology, you can create dictionaries to help you categorize and filter the data. The keywords from the example might lead you to create the following dictionaries:

- *component* dictionary for terms such as engine and brake
- *phenomenon* dictionary for terms such as fire, smoke, and spark

When you apply the dictionary enrichment to your data, it generates *annotations*. You can think of annotations as tags that you add to words or phrases, where the tag categorizes or identifies the meaning of the word or phrase. The resulting annotations function as new facets that you can use to filter and dissect your data further.

With your new *component* and *phenomenon* facets, for example, you can look for correlations between components and phenomena that are involved in traffic incidents.

[Learn about the ways that you can analyze your data](#)

Digging deeper

To dig even deeper into your data, apply or create AI models that can find different types of information in your documents. You can apply built-in natural language processing models, such as the **Entities** enrichment that can recognize mentions of commonly known things, such as business or location names and other types of proper nouns. Or you can apply a custom model that recognizes terms and categories that are unique to your data.

Extend your analysis by adding your own facets.

Getting started

Before you can use the application, you must create a Discovery Content Mining project. After the project is created and data is uploaded, you can open the Content Mining application.

For more information, see *Creating projects*.

Of course, you can't get out useful insights if you don't put the right type of information in. Be sure to include consistent data. If you want to find trends over time, your data must include data points that specify a date.

Data that is submitted in CSV file format is optimal. For a sample of a CSV file that provides interesting analysis capabilities, see Analyzing CSV files.

Data analysis methods

Use tools from the Content Mining application to analyze your data.

You can analyze your data in the following ways:

- Look for relevant keywords
- Find trends
- Identify anomalies in cyclical patterns
- Find characteristic words
- Analyze relationships
- Analyze relationships between many facets

As you review the results of your analysis, you can flag documents that you want to research further later. For more information, see Flagging documents.

When you find important insights, you can take a snapshot of the view, and then add it to a report to share with others. For more information, see Creating a report.

Start your analysis

Use the content mining application to analyze documents in your collection based on the document text and any annotations or enrichments that are stored in the documents.

To start your analysis, complete the following steps:

1. Enter a search term, click a facet with which to filter the documents, or leave the search field blank to return all of your documents.
2. Click **Search**.

The guided mode view of the results shows suggested next steps that you can take to analyze your data further. If you don't want to see suggestions, you can switch to **Expert mode**. In Expert mode, the **Documents** view that lists the search results is returned whenever you submit a search.

The tasks in this topic describe how to use the application in guided mode.

Look for relevant keywords

To analyze keyword relevance, complete the following steps:

1. From the initial search page, submit a keyword search to filter the documents.

2. From the search results page in guided mode, click **Analyze cause or characteristics**.

   After the characteristic words pane, a pane with relevancy information for each facet type is displayed.
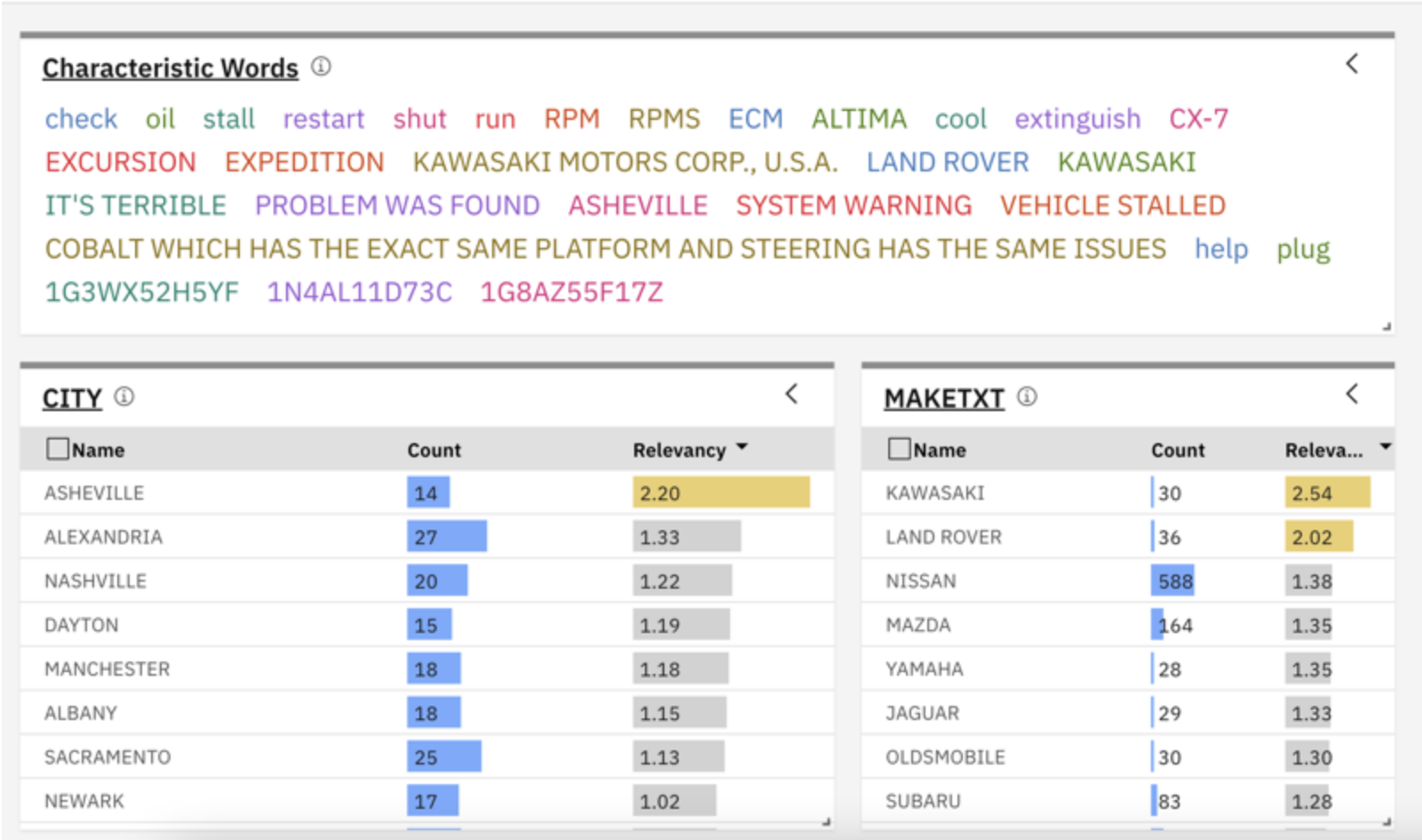
Figure 1. Facet relevancy

Each relevancy pane shows a list of the keywords that occur in the documents that match the facet type.

The **Count** column shows the number of documents in the current result set that contain the keyword. The *Relevancy* column shows the level of uniqueness of the frequency count compared to other documents that match your query. High relevancy values are shown in shades of color with increasing intensity. The color begins at yellow, then increases to orange, and then to red.

Find trends

Use *Trends* analysis to find trends in your data. For example, you might see that a new product release aligns with an uptick in customer interest. Or that a new customer care approach is followed by an increase in customer satisfaction.

⚠ **Important:** Your documents must contain at least one date field for trend information to be available.

To find trends, complete the following steps:

1. From the initial search page, enter a keyword or select a facet with number values to filter the documents.

2. Click **Find trends and anomaly** from the list of suggested next steps that is displayed in the guided mode view.

   The resulting bar graph shows the number of documents that mention the term or facet value that you specified in the search query over time.
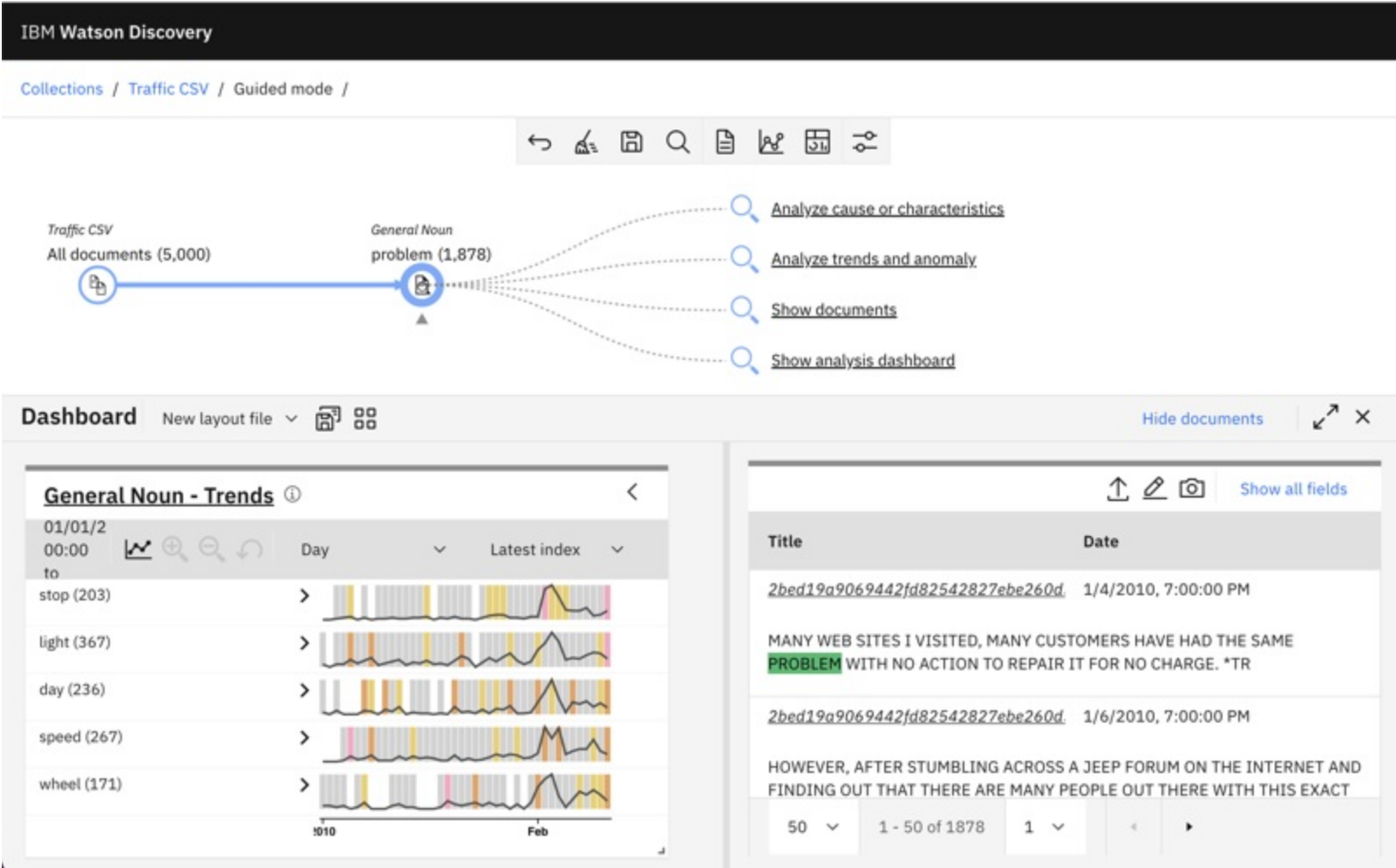
Figure 2. Facet trend graph

The time series chart is rendered as a heat map. Each cell color indicates a level of relevancy.

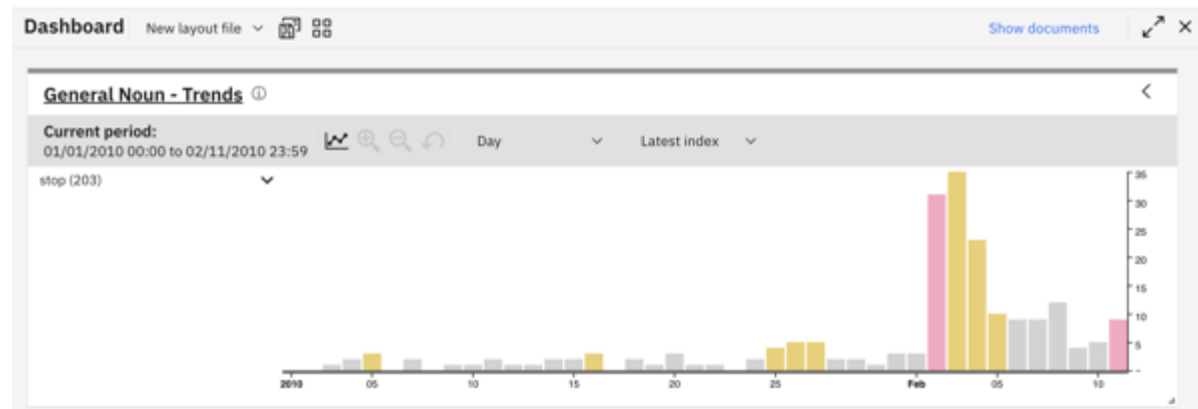3. You can click a facet to investigate it more closely. The facet is shown in a bar graph.

Figure 3. Facet trend detail in bar graph

Each individual bar graph highlights trends in your data that deviate from the normal distribution by displaying *increase indicators*.

Increase indicators measure how much the frequency of a facet value on a specific date or in a particular time interval deviates from the expected average frequency. The average is calculated based on the changes in the past time interval frequencies.

You can click individual items in a visualization or click and drag the cursor to select contiguous items.

The cyclical data is calculated from the current time zone setting of your collection. If you want to change the time zone that is used by the graph, see Change the time zone .

Identify anomalies in cyclical patterns

Use *Topic* analysis to find anomalies in seasonal, monthly, or even daily patterns that are present in your data.

> ⚠ **Important:** Your documents must contain at least one date or time field for topic information to be available.

Topic analysis focuses on how much the frequency of a keyword deviates from the expected average frequency in a specific time period. The expected average uses all of the averages of the frequency counts for other keywords in the same time period. This method of analysis is useful for identifying patterns that occur cyclically and highlights any unexpected changes that might occur in these cyclical patterns.

To find anomalies, complete the following steps:

1. From the initial search page, enter a keyword or select a facet with number values to filter the documents.

2. From the search results page in guided mode, click **Analyze cause or characteristics**.

3. From the *Facet analysis* pane, select **Topic**.

4. Adjust the following values to suit your analysis:

   - Number of results
   - Date facet
   - Time scale
   - Date range

5. Choose a target facet or subfacet, and then click **Analyze**.

   The resulting time series graph shows changes in the frequency of keyword mentions over time.

Figure 4. Topic analysis time series view

Color coding is used to highlight when the number of mentions deviates from the expected frequency. The higher the deviation, the more intense the color, from yellow to orange to red. The average is calculated based on the frequency of occurrence of other keywords in the same time period.

The cyclical data is calculated from the current time zone setting of your collection. If you want to change the time zone that is used by the graph, see Change the time zone .

Find significant terms

Find characteristic words from your data set. The characteristic words view is a word cloud that shows terms that are mentioned frequently in the documents you are analyzing.

You can click a word from the word cloud to add it to the existing query and filter the current document set to include only documents that also mention the specified word.

To find significant terms, complete the following steps:

1. From the search results page in guided mode, click **Analyze cause or characteristics**.

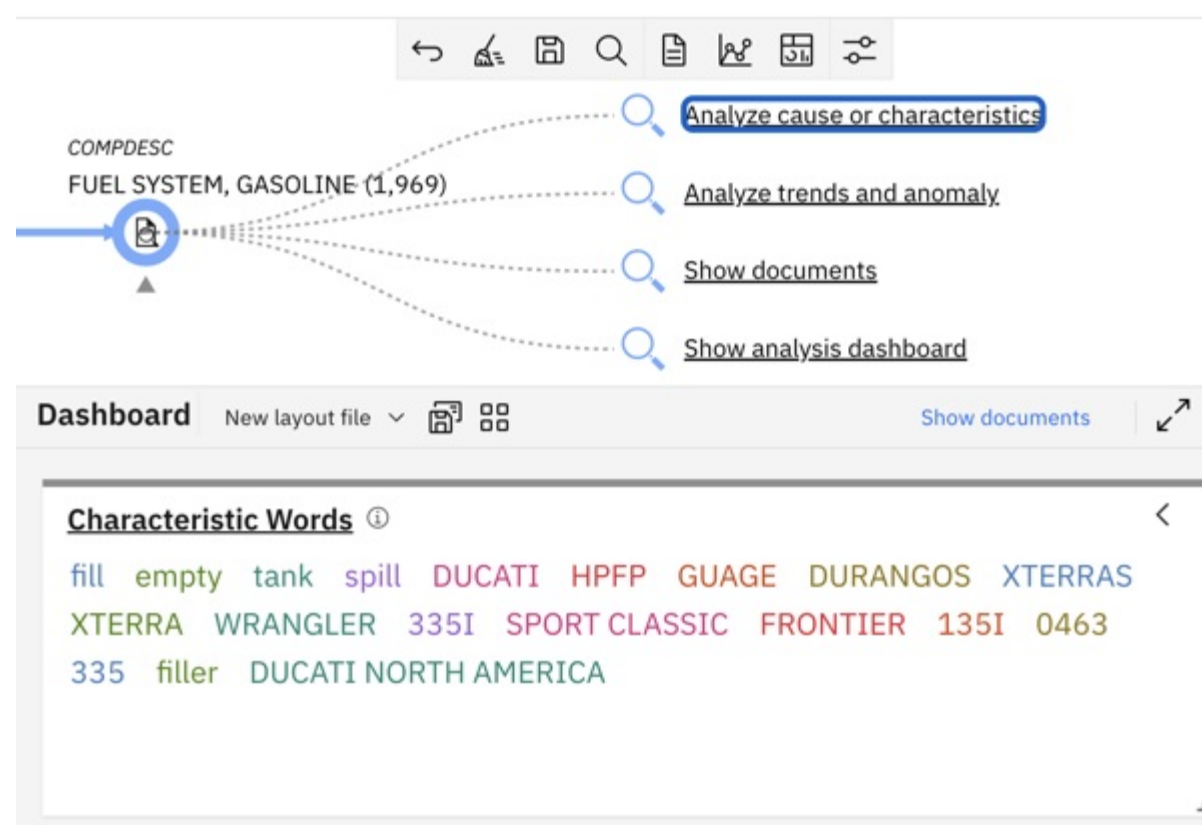   The characteristic words view is displayed.

Figure 5. Characteristic word cloud

> ▌ **Note:** The different font colors help to distinguish the words from one another; they have no statistical meaning.

2. Click a word in the cloud to limit the document set to include only documents that mention the word.

Analyze relationships between two facets

Use *Pairs* analysis to see how two facets are related to one another.

To compare two facets, complete the following steps:

1. From the *Facet analysis* pane, select **Pairs**.

2. Find the first facet that you want to compare in the list. Click either the X- or Y-axis icon that is associated with the facet to indicate where you want the facet values to be displayed in a two-dimensional graph.

3. Find the second facet, and then click the remaining axis icon. For example, if you selected the X-axis icon previously, select the Y-axis icon for the second facet.

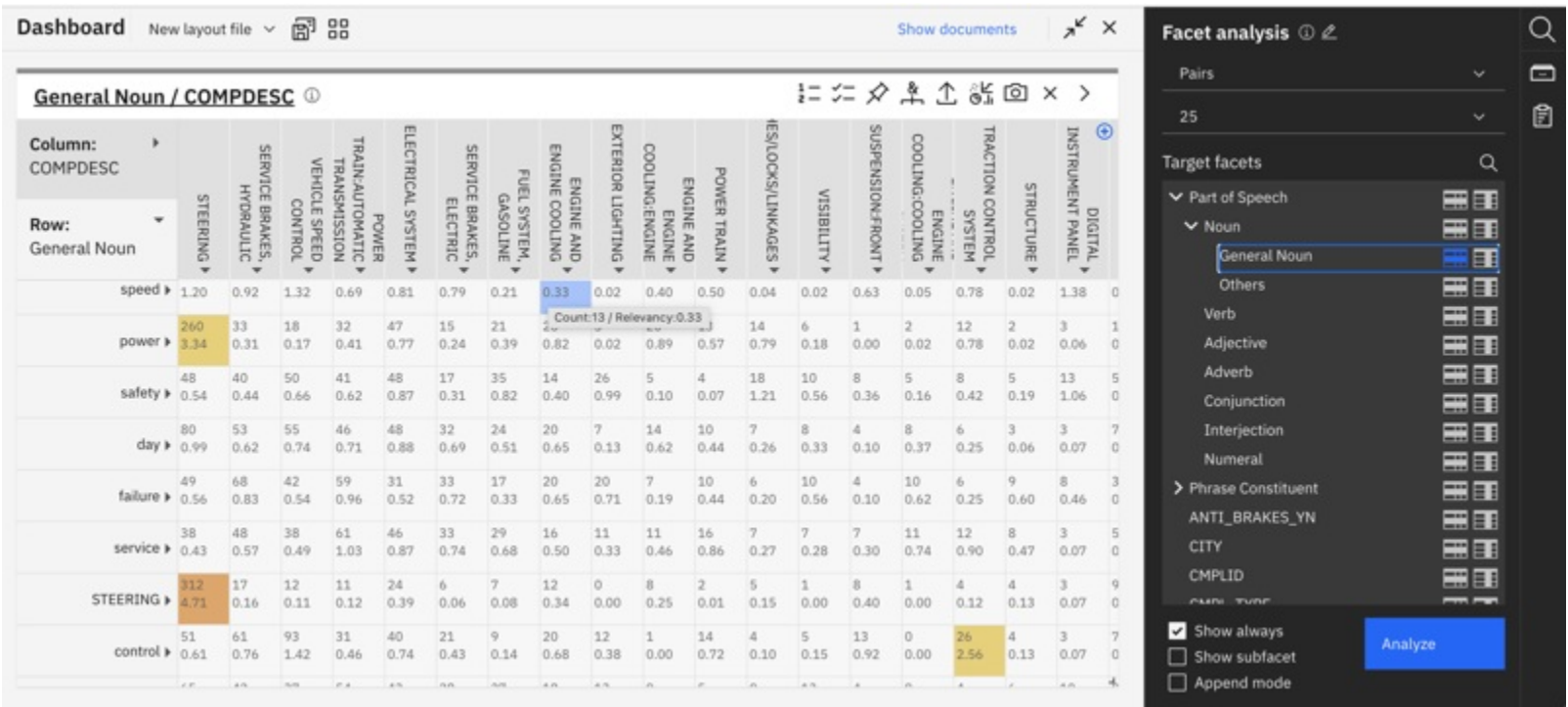   Data from the two facets is displayed in a graph.

Figure 6. Facet comparison graph

The graph shows two numbers. The first number is a frequency count and the second number is a relevancy value. The frequency count measures how many times the two data points are found together in a document. Relevancy measures the level of uniqueness of the frequency count compared to other documents that match your query. If the relevancy shows 2.0, it means that the number of times that the two data points intersect is 2 times larger than expected. To help you identify anomalies that might require more in-depth analysis, high relevancy values are shown in shades of color with increasing intensity, from yellow to orange to red.

Analyze relationships between many facets

Use *Connections* analysis to see how multiple facets are related to each other.

To compare two or more facets, complete the following steps:

1. From the *Facet analysis* pane, select **Connections**.

2. Select the root facet that you want to compare to other facets first.

3. Select up to 4 more facets from the list, and then click **Analyze**.

   Pair analysis is done between the first facet and each other facet in turn.
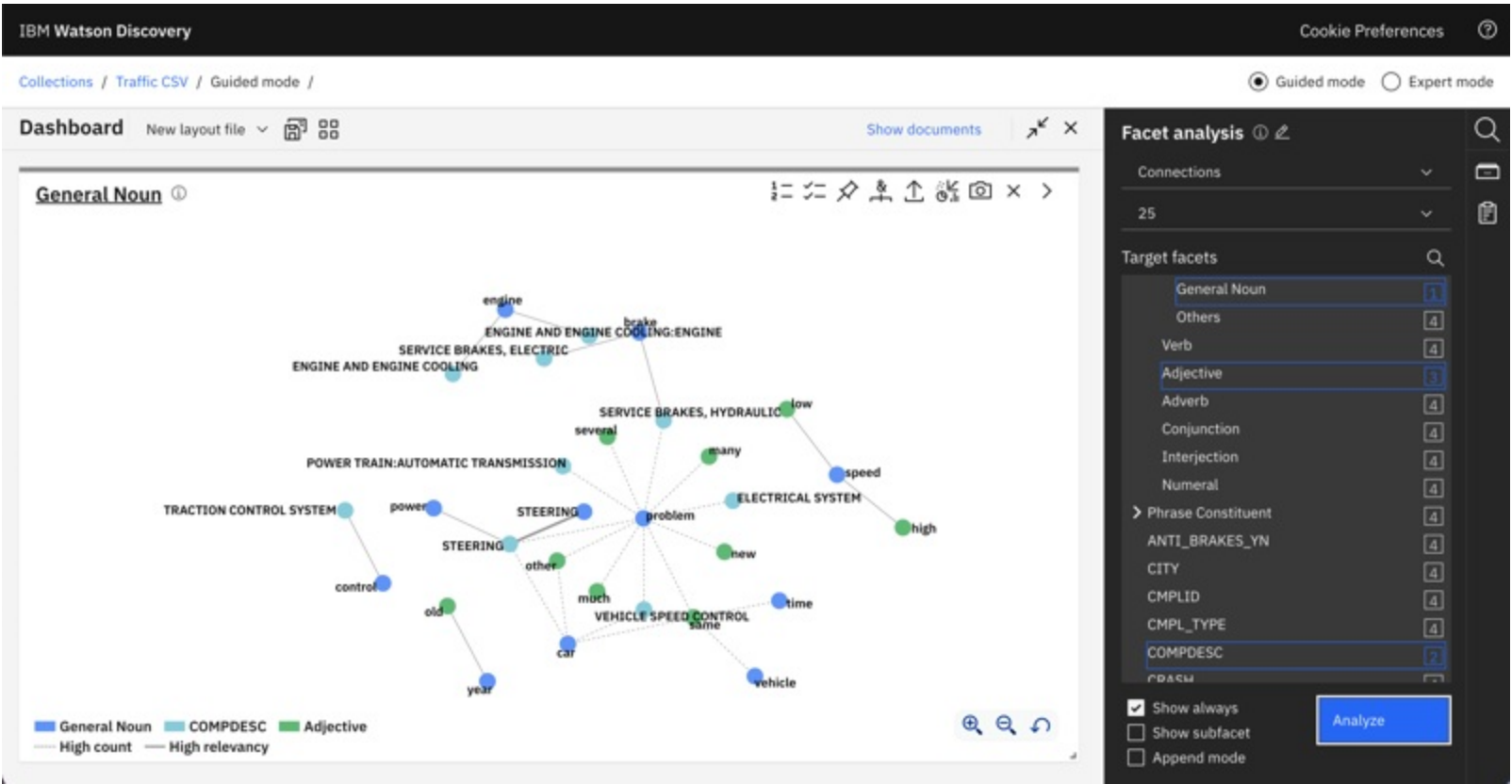
Figure 7. Facet network graph

The resulting network graph shows only highly relevant and high-frequency pairs. Each node represents a facet value. The node color reflects the facet type. A solid-line connection between nodes identifies highly relevant pairs. A dotted-line connection identifies high-frequency pairs.

Changing number ranges

If the scale of a graph is not optimized for your data, you can change it. For example, to plot vehicle speeds, you might want a range that increments by tens or twenties rather than by thousands.

To change the scale of a graph for a facet, complete the following steps:

1. Click **Collections** link in the page header.

2. In the tile for your collection, click the *Open and close list of options* icon, and then choose **Edit collection**.

3. In the *Facet* tab, find the facet for which you want to change the number range.

4. In the Range field, click **Edit**.

5. Define each range that you want to use as a JSON object. You can add or remove objects to change the number of data points in the range.

   For example, the JSON objects that identify the ranges for vehicle speeds might look as follows:

   ```
   [
     {
       "query": "[1, 20)",
       "label": "1 - 19"
     },
     {
       "query": "[20, 40)",
       "label": "20 - 39"
     },
     {
       "query": "[40, 60)",
       "label": "40 - 59"
     },
     {
       "query": "[60, 80)",
       "label": "60 - 79"
     },
     {
       "query": "[80, 100000)",
       "label": "80+"
     }
   ]
   ```

6. Click **Apply**.

7. Click **Save**, and then click **Close**.

8. Click your collection tile to return to the collection and continue your analysis.

The changes to the number ranges for vehicle speeds introduce more opportunities for relationships or anomalies in the data to be highlighted.

STATE / VEH_SPEED ⓘ

Column: VEH_SPEED — Row: STATE	1 - 19	20 - 39	40 - 59	60 - 79	80+
AZ ▶	0.70	0.36	0.74	0.54	0.00
IN ▶	25 0.41	34 0.63	34 1.00	11 0.40	0 0.00
CO ▶	31 0.55	35 0.66	16 0.39	17 0.72	0 0.00
WI ▶	33 0.59	38 0.73	18 0.46	16 0.67	1 0.02
CT ▶	35 0.65	49 1.02	9 0.18	11 0.41	0 0.00
MN ▶	55 1.15	38 0.76	14 0.34	10 0.37	0 0.00
TN ▶	27 0.52	29 0.59	21 0.63	14 0.63	5 2.04
SC ▶	20 0.49	19 0.47	16 0.60	10 0.54	0 0.00
AL ▶	7 0.10	15 0.36	20 0.85	10 0.57	0 0.00

Figure 8. Results after changed number range

Showing results in a map visualization

Facets that represent geographical locations can be shown in a map visualization. For example, if you have a collection with a US states facet, you might want to display data per state from a visualization that enables users to select each state from a map.

STATE ⓘ

○ Relevancy ○ Count

Label: West Virginia
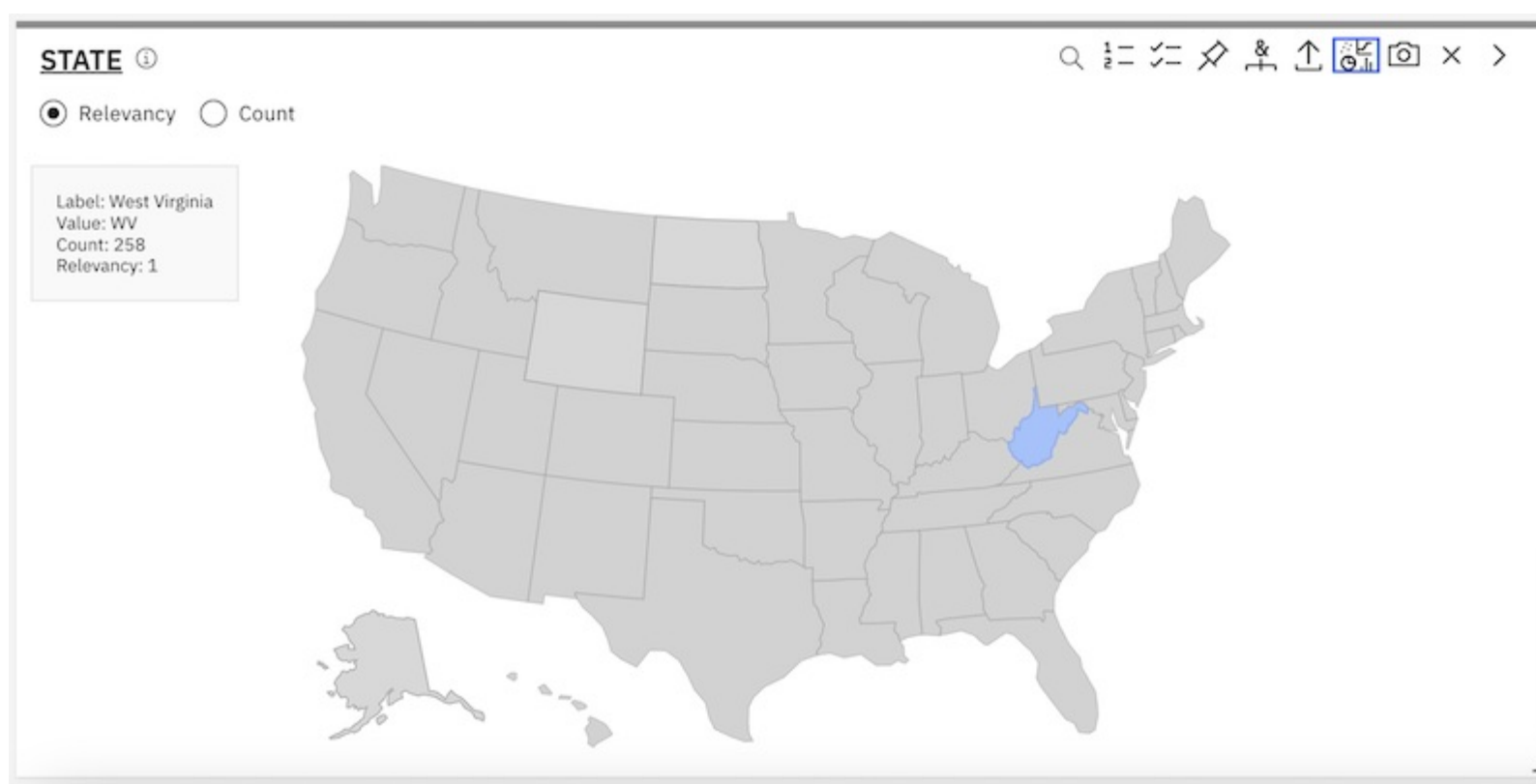Value: WV
Count: 258
Relevancy: 1

Figure 9. Results shown in a map visualization

A US Map is available by default. You can add a custom map that is built in GeoJSON format. For more information, see RFC7946.

To use a map that you define, complete the following steps to import the map definition:

1. From the Content Mining application home page, click **Collections** from the breadcrumb in the page header.
2. Click the *Settings* icon at the start of the page.
3. Click **Manage customization resources**.
4. Click **Add resource**.
5. Name the resource, and then click **Next**.
6. Add your map file, and then click **Save**.

To make the map that you added available as a visualization option for a facet, you must edit the facet.

1. Click **Home** from the breadcrumb in the page header.
2. Right-click the overflow menu for your collection, and then choose **Edit Collection**
3. Open the *Facet* tab, and then find the facet with which you want to associated the map visualization.
4. Change the *Visualization type* value to **Map**, and then pick the map that you added from the list in the **Resource** field.
5. Click **Save**, and then click **Close**.

Flag documents of interest

Use document flags to assign a custom flag to a document or a group of documents for classification, export, or further analysis.

Flagging documents is a useful way to highlight documents that you want to examine further later.

Before you can flag documents, you must create flags for your collection. For more information, see [Add document flags](#).

To apply flags, complete the following steps:

1. From the analysis view of your collection, create a query that returns a set of documents with specific characteristics.

2. From the documents view, click the *Document flags* icon.

3. Select a flag.

4. You can choose to apply the flag to all query results or to selected documents, and then click **Apply**.

> 🔖 **Note:** You can't set a document flag more than 50 times per collection. Whether you flag one document that you select individually or flag a query, which might return many documents, each action counts as setting a flag one time.

A flagged document set dynamically changes as the collection is updated. Flagged document sets are stored as queries in the index. Each flag has a query that represents the document set that it is associated with. For example, after you create the document flag and you search for the term `ice cream` and apply a red flag to all of the documents that have this word, `ice cream` is stored as the query that represents the flag. Then, if you search for the term `coffee` and apply the red flag to all of the documents that have that word, the internal flag query changes to `(ice cream) OR coffee`. Therefore, if new documents that contain the word `coffee` are ingested, the red flag is applied to those documents automatically.

Viewing flagged documents

To view the documents to which a flag is applied, complete the following steps:

1. In the *Facet analysis* panel, scroll down to the *Document flags* facet.
2. Select the facet, and then click **Analyze** to open the *Document flags* dashboard.
3. Click one of the flags, click **Analyze more**, and then click **Show documents**.

Removing document flags from a Document Flags query

To remove a document flag, complete the following steps:

1. From the *What do you want to analyze?* page, submit an empty query by clicking **Search**.

   The empty query returns all of the documents in your collection.

2. Click **Show documents**.

3. Click the **Document flags** icon on the toolbar, clear the checkbox of the document flag, and then click **Apply**.

   The document flags are removed from your documents.

Adding facets

Add more facets that you can use to filter your data.

When you apply custom enrichments to your collection, annotations are added to its documents. The annotations feed into new facets that you can use to sort your data.

The following table describes the types of facets that you can create from annotations.

Information to recognize	Annotator type
Commonly understood terms, such as organization or people names.	Built-in Natural Language Processing models

Phrases that express an opinion and evaluate whether the opinion is positive or negative.	[Phrase sentiment](#)
Alternative words that share a meaning with terms in a finite list.	[Dictionary](#)
Terms that match a syntactical pattern	[Regular expression](#)
Custom terms by the context in which they are used.	[Machine learning model](#)
Documents that fit into categories that you define.	[Document classifier](#)

Custom facet types

Grouping facets

To organize your facets, you can group them in folders.

Grouping facets does not combine the data from the facets. It merely makes the facets easier to find because they are organized in named folders.

To associate facets such that you can combine data from multiple facets, the facets must have a facet and subfacet relationship. Such hierarchical relationships must be defined at the time that the facet enrichment or annotation is created and applied to the collection.

To group facets, complete the following steps:

1. From the initial search page, submit a search.

2. From the *Facet analysis* pane, click the *Edit* icon.

3. Name the group, and then select the facets that you want to group together.
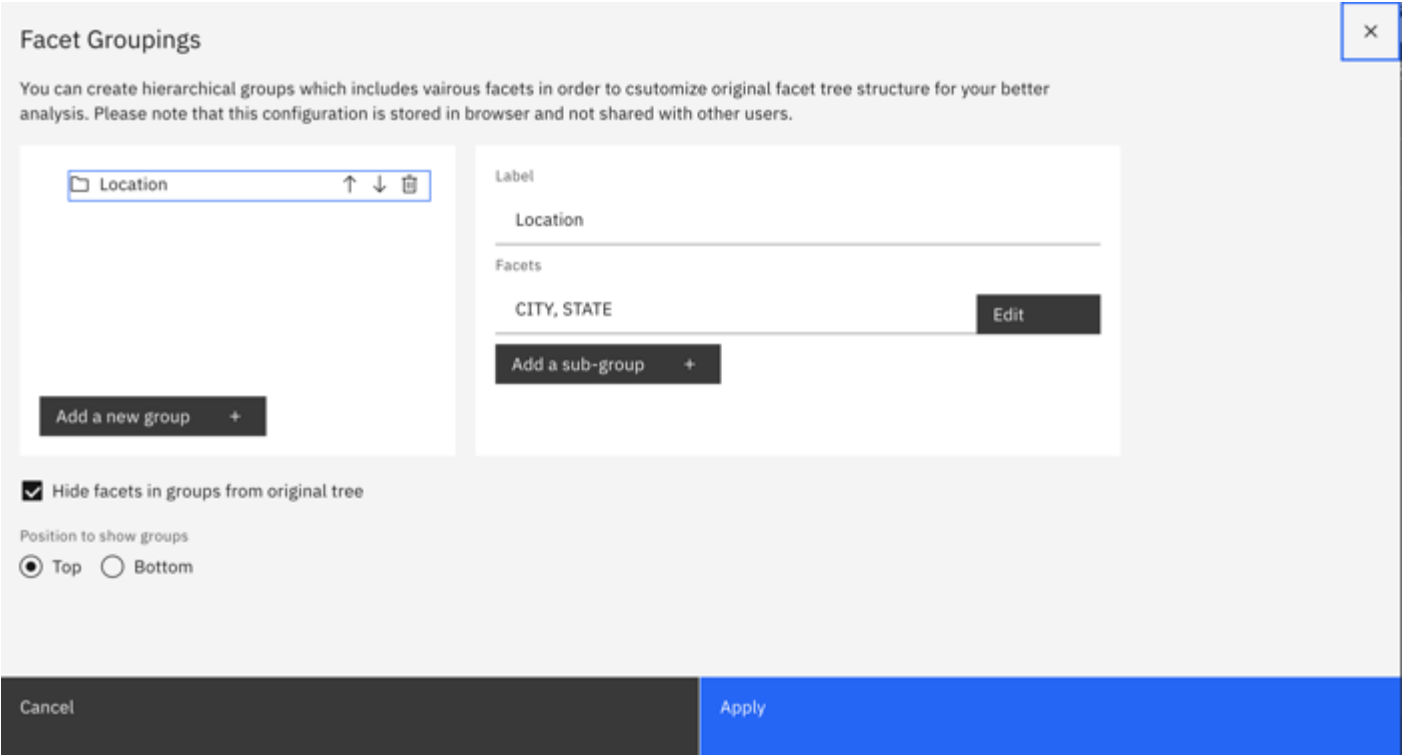

Figure 1. Facet grouping dialog

4. Click **Apply**.

5. The facets that you grouped are now available from a folder with the group name.
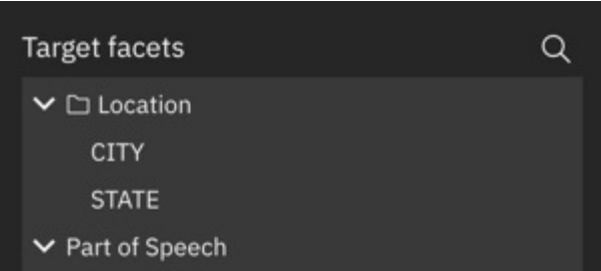

Figure 2. Facet folder from the facet list

Creating custom annotators

You can create a dictionary, regular expression, or machine learning annotator to generate new facets that can help you to analyze your data.

Before you begin, have the following data ready.

Annotator type	Description	Data
Dictionary	Assigns facets to terms that match dictionary entries that you define or upload.	You can optionally upload a file of dictionary terms.
Machine learning	Assigns facets to mentions that are recognized by a machine learning model that you upload.	A compressed file of a machine learning model is required.
Regular expression	Assigns facets to text that matches Java regular expression patterns that you define or upload.	You can optionally upload a JSON file that contains regular expression patterns.

Custom annotator prerequisite data

To create a custom annotator, complete the following steps:

1. From the analysis view of your collection, click the **Collections** link in the breadcrumb to open the *Create a collection for your analytics solutions* page of the Content Mining application.

2. To create an annotator, click **collection**, and then select **custom annotator** from the list.
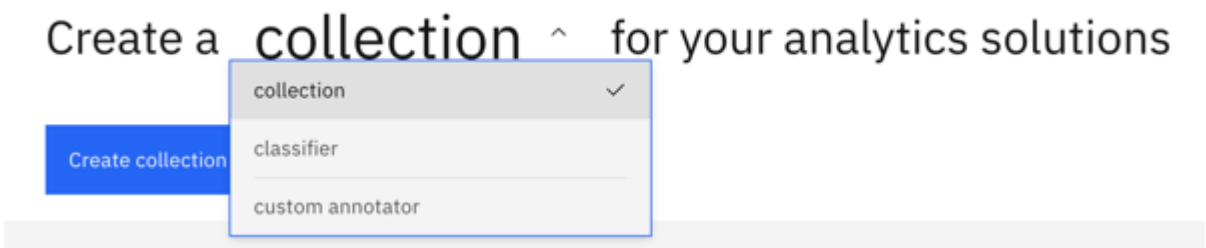

Figure 1. Collection menu

3. Click **Create custom annotator**.

4. Name your annotator, and then optionally add a description.

5. Choose the annotator type, and then click **Next**.

6. Follow the on-screen instructions.

   For more information about how to configure each annotator type, see one of the following sections:

   - [Dictionary](#)
   - [Machine learning model](#)
   - [Regular expressions](#)

Dictionary configuration

You can import an existing dictionary by uploading it or you can create a dictionary by adding terms one at a time.

If you plan to import a dictionary, the dictionary terms must be defined in a CSV file. Specify each term and its synonyms in a separate line. Use the following syntax to specify each term:

```
{term},{synonym},{synonym},...
```

To add a dictionary, complete the following steps:

1. Do one of the following things:

   - To import the dictionary terms:

     1. Click **Import**, and then browse for the file with your dictionary terms.
     2. Click **Import**.

   - To define the dictionary terms:

     1. Click **Add**.
     2. Click *Word list* to add the dictionary terms.
     3. Click **Add**, and then add the term in the **Base word** field and any synonyms that you want to define for the term in the **Other words** field. Separate multiple synonyms with commas. Click **OK**.
     4. Repeat the previous step to add more dictionary terms.
     5. After you finish adding dictionary terms, click *Basic settings*.

2. Name the dictionary.

3. If you plan to define terms with a part of speech other than a noun, specify the part of speech.

4. Decide how you want to handle case.

   When case is ignored, the terms `Sat` , `SAT` , and `sat` are all labeled as occurrences of the `Sat` dictionary term.

   When you deselect the **Ignore case** checkbox to create a case-sensitive dictionary, the surface form of the term with uppercase match is used. Annotations are added for the term exactly as written and for variations of the term in which the letters are uppercase.

   For example, a `sat` entry in the dictionary results in annotations for `sat` , `Sat` , or `SAT` mentions when they occur in text. For a `Sat` entry in the dictionary, annotations are added for occurrences of `Sat` and `SAT` , but not for `sat` .

5. Identify the facet name to use for this dictionary.

   The facet name that you specify for the annotator is the facet name that is displayed from the collection search view.

   You can create a hierarchical facet by including a period (.) in the facet name. For example, you might create one dictionary with the facet path `Food.Vegetables` and others with the facet paths `Food.Fruits` and `Food.Proteins` . Add more facet groups with more periods. For example, you can add `Food.Proteins.Nuts` and `Food.Proteins.Meats` to categorize proteins even further.

Figure 2. Adding a dictionary

6. If you want documents that are returned for a subfacet to be included when a user filters on the root facet, select **Lift up words**.

   For example, you might enable *Lift up words* for `Food.Fruits` and `Food.Proteins` but not `Food.Vegetables` . As a result, when a user clicks the Food facet, the returned documents include documents that mention terms included in the Fruits and Meats dictionaries, such as *apples* and *beef*.

Figure 3. Dictionary enrichment application

However, a user must click the *Food>Vegetables* facet explicitly to get documents that mention terms in the Vegetables dictionary, such as *lettuce*, to be returned.
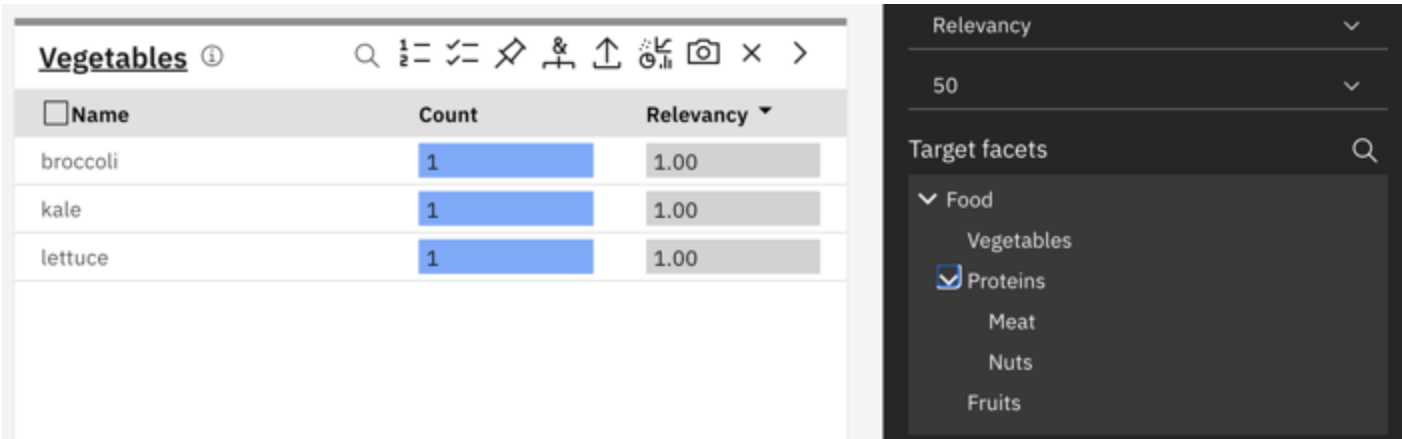
Figure 4. Subfacets

7. Repeat previous steps to add more dictionaries.

8. Click **Save**.

From the custom annotator page, you can see dictionaries that were created in other projects, including non-Content Mining projects. Dictionaries from other project types show the enrichment name as the annotator name. The *Ignore case* and *Lift up words* settings are disabled and the dictionary is named `custom dict`.

Dictionary limits

Plan	Number of dictionaries per service instance	Number of base words per dictionary	Number of terms for which suggestions can be generated
Cloud Pak for Data	Unlimited	Unlimited	1,000
Premium	100	10,000	1,000
Enterprise	100	10,000	1,000

Dictionary plan limits

Totals include enrichments that you create in this Content Mining project and in other projects in the same service instance.

Machine learning configuration

You can import an existing machine learning model.

To use Discovery to create a model, see *Entity extractor*.

To import a model, complete the following steps:

1. Click **Select file**, and then browse for the machine learning model file.

2. In the **Facet path** field, specify the root facet name to use for the model.

   The facet name that you specify for the annotator is the facet name that is displayed from the collection search view.

3. Click **Save**.

Machine learning model limits

Plan	ML models per service instance
Cloud Pak for Data	Unlimited
Premium	10
Enterprise	10

ML model plan limits

Totals include enrichments that you create in this Content Mining project and in other projects in the same service instance.

Regular expressions configuration

You can import existing patterns by uploading them in a JSON file or you can add patterns.

To add patterns, complete the following steps:

1. Add the regular expression pattern to the **New pattern** field, and then click **Add**.

2. Specify a name for the pattern, and then identify the facet name to use for this pattern.

   The facet name that you specify for the annotator is the facet name that is displayed from the collection search view.

3. **Optional**: Specify a facet value. You can specify a value from the options that are described in the table.

Facet value	Description
`$0`	Displays the matched text as-is.
`$n`	If your regular expression pattern contains groups, you can specify a group number to return the matched text from the pattern group only. For example, if your regular expression consists of 3 groups that define a US phone number pattern, such as `(\d{3})-(\d{3})-(\d{4})`, and you want to return only the area code portion of the phone number, you can specify `$1`. If the matched text is `212-555-1234`, then the facet value is displayed as `212`. Only specify a group as the facet value for patterns that you know will return matches.
`{prefix-text}:$0`	Adds hardcoded text in front of the facet name. You might want to use this option if you want to distinguish facets that are generated by this regular expression from facets that are similar but generated in some other way. For example, `MyRegex:$0` results in a facet named `MyRegex:212-555-1234`.

Regular expression facet value options

4. Click **Save**.

To import patterns, complete the following steps:

1. Define the patterns that you want to add in a JSON file.

   The pattern definition must use the following syntax:

   ```
   [
     {
       "name": "US Phone number",
       "description": "US mobile phone number",
       "pattern": "(\\d{3})-(\\d{3})-(\\d{4})",
       "facetPath": ".regex.usphonenumber",
       "facetValue": "$0"
     }
   ]
   ```

   Keep the following notes in mind:

   - The patterns must be defined in an array, even if you plan to define only one pattern.
   - Escape any backslash (`\`) characters with a backslash.
   - For more information about the facet value options, see the *Regular expression facet value options* table.

2. Click **Import**, and then choose the JSON file where the patterns are defined.

3. Click **Save**.

Regular expression limits

Plan	Regex enrichments per service instance	Regex patterns per service instance
Cloud Pak for Data	Unlimited	Unlimited
Premium	100	50
Enterprise	100	50

Regular expression plan limits

Totals include enrichments that you create in this Content Mining project and in other projects in the same service instance.

Applying the annotator

After the annotator is created, you must apply it to your collection.

1. From the *Create a custom annotator for your analytics solutions* page of the Content Mining application, click **custom annotator**, and then select **collection** from the list.

2. In the tile for your collection, click the *options* icon, and choose **Edit collection**.

3. Click the **Enrichment** tab, and then select the annotator that you created.

   You might need to scroll to find it.

4. Click **Save**, and then confirm the action.

Give the index time to rebuild.

Filtering documents with your facet

1. Click the collection tile to open your collection in the data analysis page.

2. Do one of the following things:

   - Your custom facets are listed in the *Facets* view. Scroll and click **Load more** repeatedly until your facets are displayed.

   - Submit an empty search to return all documents. In the *Facet analysis* pane, select the facet that you created.

   - To access your custom facets more quickly, add them to a custom view. Select **Custom** as the view, and then click **Edit**. Select one or more facets to add to the view, and then click **Save**.
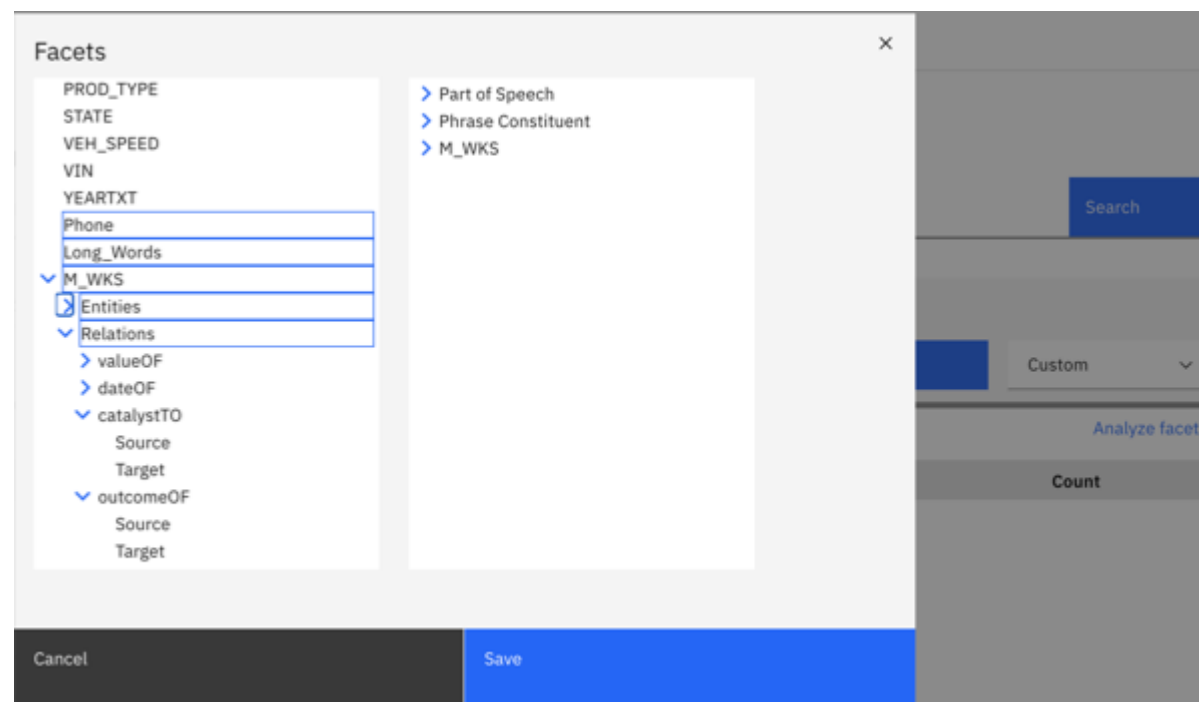

Figure 1. Collection menu

Classifying documents

A document classifier machine learning model analyzes documents and tags them with the appropriate label from a set of labels that you define.

Classifying documents is useful when you want to sort many documents into groups programmatically. For example, you might have a collection that contains customer comments about products that you sell. If you can automatically sort the feedback into classes, you can isolate urgent issues that customers mention and tackle them first. Based on previous feedback, you might define classes such as the following labels:

- Not functioning correctly
- Features not as advertised
- Difficult to use
- Missing parts
- Parts shipped don't match parts list in assembly instructions

To create a document classifier, you build a machine learning model that can recognize which class best captures the point of customer feedback that is specified in natural language. You pair them with class labels that represent real scenarios that make sense for your business.

What's the difference between a document classifier and a text classifier?

A document classifier can classify documents based on words and phrases extracted from the body text fields with information from their part of speech and the other enrichments that are applied to the body text taken into account. The information from the other non-body fields are also used. A text classifier can classify documents based on words and phrases extracted from the

body text with their part of speech information taken into account. For more information about how to create a text classifier, see *Classifier*.

Before you begin

To train the document classifier model, you must provide sample documents that are labeled appropriately. Prepare the following files:

Training data

Required. CSV file that is used to train the document classifier machine learning model. The file can contain key data points per column. The data points can vary, but the file must include the following columns:

- Natural language text that you want to classify or label.
- Label or class name that categorizes the idea that is expressed in the document text. You can apply more than one label to a text sample. Separate multiple label values with a semicolon.

Test data

Optional. CSV file that is used to test the document classifier machine learning model after it is trained. If you don't specify a separate file for testing, a subset of the training data content is used for testing purposes.

Target data

Required. CSV file with the data that you want to classify.

⚠ **Important:** All of the CSV files (training, test, and target) must have the same column names. The data in the columns must have the same data types, such as string, number, and so on.

You can use a CSV file that you uploaded at the time that you created the Content Mining project or you can create a new collection.

For more information, see the following topics:

- [Adding collections](#)
- [Analyzing CSV files](#)

Document classifier training data sample

The following table shows an example of the type of content that might be stored in CSV files that are used to train a document classifier.

Claim_id	Date	Product_line	Product	Client_segments	Client_location	Client age	Feedback	Label
0	2016/1/1	tea	lemon tea	Not Member	Manhattan	20	The straw was peeled off from the juice pack.	package_container
1	2016/1/2	ice cream	vanilla ice cream	Silver Member	Queens	20	I got some ice cream for my children, but there was something like a piece of thread inside the cup.	contamination_tampering

Table 1. Sample data for CSV files

Note that the two required fields are present in the sample. The required fields have the following names:

- `Feedback`: Natural language text to label.
- `Label`: Label to apply to the feedback.

Opening the Content Mining application

If you didn't do so, create the project and add a collection to it. If you already created the project and collection, you can skip this procedure and create the document classifier.

1. In Discovery, create a Content Mining project.

2. Choose to upload data to create the collection. Name your collection, and click **Next**.

3. Upload the CSV file that contains your training data.

   The training data file must contain the following information at a minimum:

   - A column that contains sample text that you want to classify. For example, the sample text might be a product review.
   - A column that contains a class or category label that is assigned to the sample text.

4. After collection processing is complete, click **Launch application** to open the Content Mining application.

   The facet details are displayed for the collection.

Creating a document classifier

To create a document classifier, complete the following steps:

1. From the Content Mining application, click the **Collections** link in the breadcrumb to open the *Create a collection* page.

   The status of index creation is displayed. Wait for the collection to be fully indexed before you continue with this procedure.

2. To create a classifier, click **collection**, and then choose **classifier** from the list.
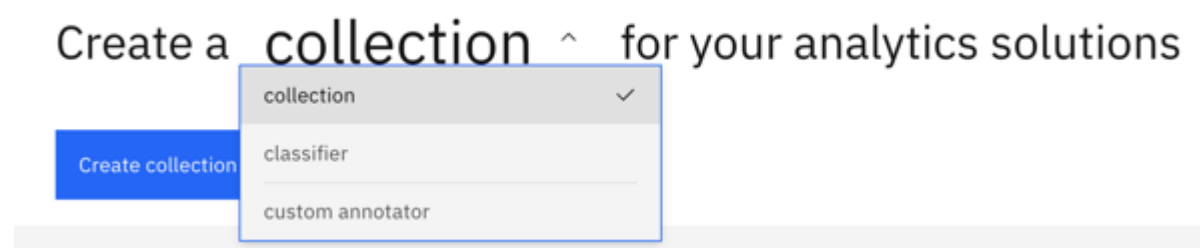

Figure 1. Collection menu

3. Click **Create classifier**.

4. Name your classifier.

   When you deploy the model as an enrichment later, the enrichment is given a name with the format `{classifier name} - {model name}`. For example, if your classifier is named `Product reviews` and the model is named `v0.1`, then the enrichment name is `Product reviews - v0.1`.

   Optionally, add a description and identify the language of your training data by selecting it from the *Language* field.

5. Click **Next**

6. On the *Training data* page, select the file that you uploaded previously from the list, and then click **Next**.

   Alternatively, you can upload a CSV file that contains your training data.

   The *Fields* page is displayed. It shows details about the fields that are generated from the file that you added. Typically, each column in a CSV file is converted into a field and is assigned a name that is copied from the column header.

7. Deselect any metadata fields that you want to exclude from the data set for your document classifier to learn from, and then click **Next**.

   Any fields that you include are used as additional features in the classification. All of the fields are selected by default. You might need to scroll horizontally to review all of the fields.

8. On the *Classifier* page, specify the fields to use for machine learning training and prediction.

   Answer field

   Select the field from your training data file with the classification label. From the earlier example, the `Label` field is the best

choice.

Predicted field

The name of the facet that is generated for the predicted class values. By default, the facet name has the syntax **<Answer field value>_predicted**. For example, **Label_predicted**.

Test dataset

Specifies the data set to use to test the classifier model. By default, the training data CSV file that you uploaded and configured is split into three data sets that are used for training, validation, and test respectively. However, you can optionally specify a separate data set to use for testing the model.

Train federated model

Creates more than one model, based on values from a specific field in the data set. For example, if the document has a **Product** field, you can configure the classifier to create a separate classifier model for each product name value that is specified in the field. By default, the classifier creates one machine learning classifier model.

> 📑 **Note:** You don't need to specify the field that contains the text to be classified. The system detects this field automatically. You can check which field the analyzable text is extracted from and change it or augment it by changing index type of another field. For more information, see [Identifying the text field](#).

Click **Next**.

9. If you want to apply an enrichment to the text in your training data, select at least one field from the **Target fields** list where you want to apply enrichments.

   Typically, you want to choose the field that contains the body of text that you want to classify. From the earlier example, the **Feedback** field is the best choice.

   Next, select any annotators that you want to apply to enrich the text in the target field or fields, and then click **Next**.

   The *Part of speech* annotator is selected by default.

10. On the *Confirm* page, review your classifier configuration settings. To make changes, use the **Back** button. Otherwise, click **Save**.

    An *Overview* page is displayed.

11. Click **New model** to create and train your machine learning model.

12. You can optionally change the name of the model and add a description.

    You can change the default ratio values that are specified for the following data sets:

    - Training dataset: Updates the weights of the training model.
    - Validation set: Monitors the accuracy of the training model during training. The accuracy result is used to draw a training loss graph.
    - Test dataset: Calculates the score of the trained model.

13. Click **Create**.

    It might take several minutes for model training to complete.

Deploying the document classifier model

After the model is trained, deploy the model as an enrichment.

1. Click the overflow menu icon in the **Actions** column, and then click **Deploy model**. Specify the name and other details, and then click **Deploy**.

2. Do one of the following things:

   - To apply the document classifier to a collection in your Content Mining project, see *Enriching your collection*.

   - To apply the document classifier to a collection in a different project, complete the following steps:

     1. In Discovery, create or open the collection that has the documents that you want to classify.

        > 📑 **Note:** The data in the collection where you apply the enrichment must have the same fields as the collection that you used to train the model.

2. In the **Enrichments** tab, locate your classifier in the **Name** column. From the **Fields to enrich** field, choose the same text field that was used to train the model. (This field is determined by the system and is indexed as the *Analyzable text content* field. For more information, see Identifying the text field.)

3. Click **Apply changes and reprocess**.

Results of classification

After the enrichment is applied to a collection, a facet is generated that you can use to find the predicted classes. In this example, the predicted field is named `label_answer_predicted`.
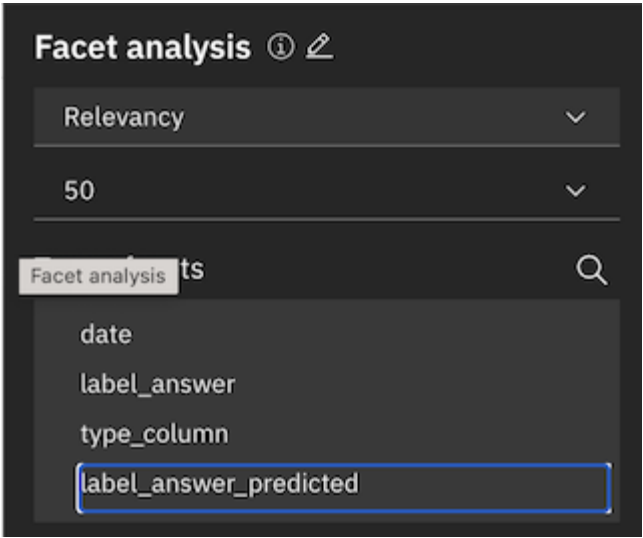

Figure 2. A Label_answer_predicted facet is generated

Use the generated facet to filter documents by classification and analyze subsets of documents. Doing so helps you to find patterns and discover other insights. You can export these target documents to share with team members or to analyze further. For more information, see Exporting data.

When the document classifier classifies a document, it stores the classification in the `document_level_enrichment.classes.class_name` field.

For example, the following JSON excerpt shows a document that was classified with the `package_container` class.
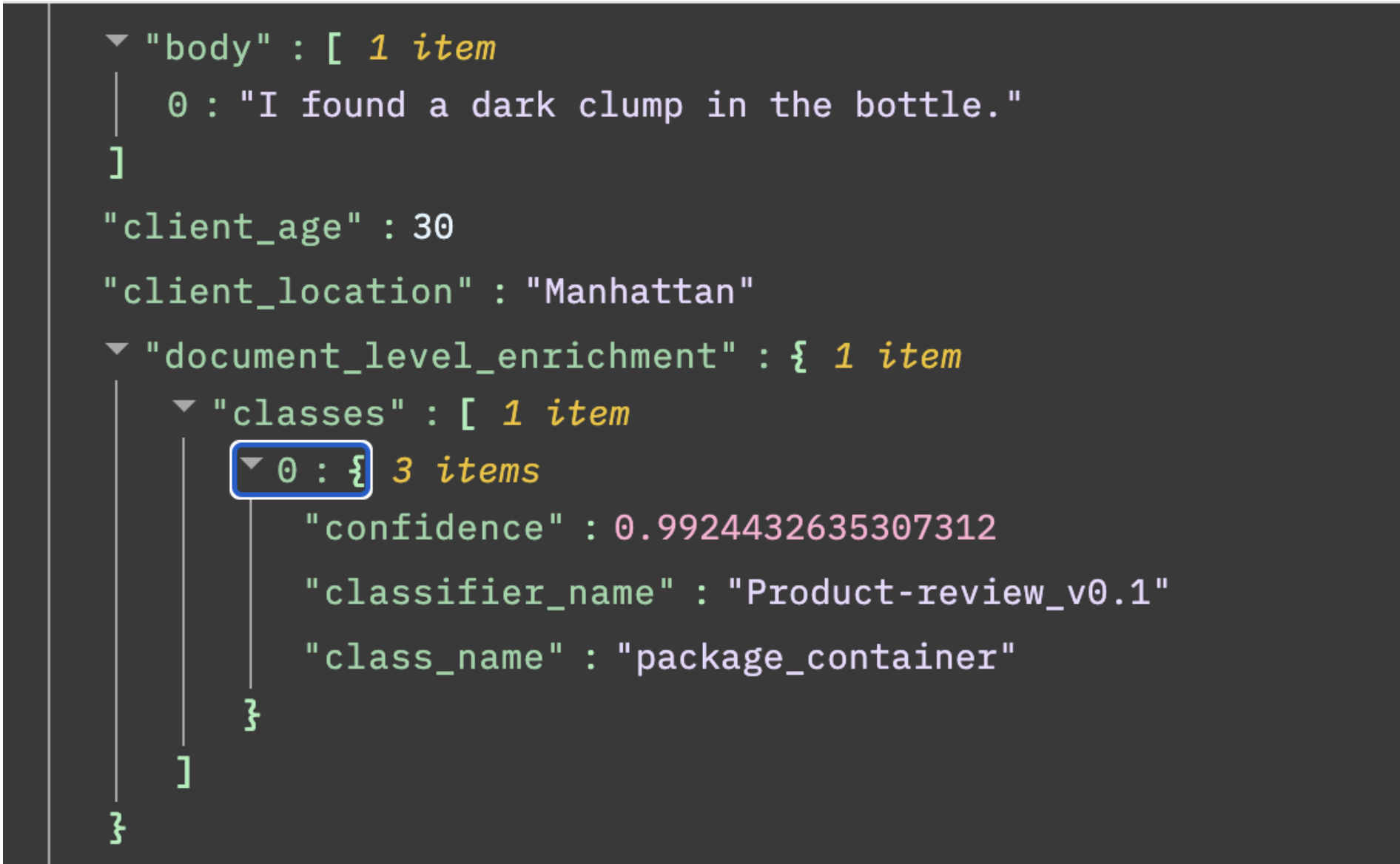

Figure 2. Document classifier enrichment syntax

Document classifier limits

The number of document classifiers and labels that you can create per service instance depends on your Discovery plan type.

Limit	Enterprise	Premium	Cloud Pak for Data

Number of document classifiers per service instance	20	20	Unlimited
Number of labeled data rows	20,000	20,000	20,000
Maximum size in MB of training data after enrichment	1,024	1,024	1,024
Number of labels	1,000	1,000	1,000
Number of target fields	50	50	50

Document classifier plan limits

Detecting phrases that express sentiment

Analyze a document to find phrases that express an opinion or reaction and assess whether the sentiment expressed is positive, neutral, or negative. For English and Japanese, you can also detect specific sentiment targets. The Content Mining application marks these extractions as annotations.

For example, if a product feedback form contains the following sentence, you want to find it and indicate that it is a `positive` statement.

`I love my XYZ blender...`

What's the difference between phrase and document sentiment?

> Document sentiment is a built-in Natural Language Processing enrichment that is available for all project types. Document sentiment evaluates the overall sentiment that is expressed in a document to determine whether it is positive, neutral, or negative. Phrase sentiment does the same. However, phrase sentiment can detect and assess multiple opinions in a single document and, in English and Japanese documents, can find specific phrases. For more information about the document sentiment enrichment, see *Sentiment*.

Complete the following steps to enable phrase sentiment analysis:

1. From the analysis view of your collection, click the **Collections** breadcrumb link in the page header.

2. In the tile for your collection, click the *open and close list of options* icon, and then choose **Edit collection**.

3. Click the **Enrichment** tab, and then select the **Sentiment of phrases** annotator.

4. Click **Save**, and then click **OK** to verify the change.

   The collection is reindexed. Wait for processing to be completed.

5. Click **Close** to return to the *Collections* page, and then click your collection tile.

6. In the *What do you want to analyze?* field, enter a term to search for in your documents or select one or more facets, and then click **Search** to filter the documents.

   The search results are displayed in the mining graph. The *Facet analysis* pane is displayed also. By default, *Relevancy* analysis is shown.

7. In the drop-down menu from the *Facet analysis* pane, select **Sentiment**.

8. In *Target facets* from the *Facet analysis* pane, expand the **Sentiment Analysis** option to see facets that are available for analysis in your documents.

9. Click a facet to explore.

   For example, if you click **Positive Expression**, you can see the following information:

   - Positive expressions that were identified in your documents
   - Sentiment percentage
   - Side-by-side comparison of positive and negative expressions
   - Number of instances of the expression
   - Expression relevancy

10. Click one or more options in the facet list, or select one or both facet lists, and then click **Analyze more**.

View the phrase, expression, or target in the *Documents* or *Trends* views.

> 🔖 **Note:** Text from the body field of the document is analyzed. For more information about which field is used for the body text, see [Identifying the text field](#).

Adding collections

You can add a collection directly to the Content Mining application.

You might want to add a collection from within the Content Mining application to make data available for use as training data for a document classifier, for example.

The collection can contain an uploaded CSV file only. For information about file guidelines, see [Analyzing CSV files](#).

The collection that you create is not added to your existing Content Mining project. A new Content Mining project is created to store the collection. The project that is generated is given the name that you specify for the collection.

To add a collection, complete the following steps:

1. From the analysis view of your collection, click the **Collections** link in the breadcrumb to open the *Create a collection for your analytics solutions* page of the Content Mining application.

2. Click **Create collection**.

3. Drag your CSV file to the *Import your files* dialog, or click Open to browse for the file. When the button is available, click **Next**.

4. You can optionally customize the columns that you want to include or exclude from the collection, and adjust the data types of the fields from the *Fields* page. Click **Next**.

5. From the *Enrichments* page, you can optionally apply or remove any enrichments from the collection, and then click **Next**.

   The *Part of Speech* enrichment is applied automatically.

6. On the *Facets* page, you can optionally customize the data that is displayed for facets. Click **Next**.

7. Click **Save** to save and index the collection.

Editing your collection

You can change the characteristics of your collection from the Content Mining application.

You can change the following characteristics:

- [Change the time zone of your collection](#)
- [Add document flags that you can use to tag documents of interest in your collection](#)
- [Change or augment the field that is designated as the source for the text body of your documents](#)
- [Group text body fields](#)
- [Add, remove, or change the enrichments that are applied to the collection](#)

Edit a collection

1. From the analysis view of your collection, click the **Collections** link in the page header.

2. In the tile for your collection, click the *Open and close list of options* icon, and then choose **Edit collection**.

3. Use the appropriate tab to change characteristics of the collection.

4. When you are done making changes, click **Save**.

   The following message is displayed:

   ```
   You need to clear index to make these changes.
   After clearing index, fully build the index to
   analyze using this collection.
   ```

   You can ignore the message. The index is rebuilt automatically when you click *OK*.

5. Click **OK** to verify the change.

6. Click **Close** to return to the *Collections* page.

   > ☑ **Tip:** Wait for the index to be rebuilt before you continue your analysis. From the *Collections* page, you can see the progress

7. Click your collection tile to return to the data analysis page.

Change the time zone

To change the time zone that is used by the trend graph, you must edit the default time zone for the collection.

1. Complete the steps in <u>Editing a collection</u> to get the collection into edit mode.
2. In the *Edit* tab, change the value of the **Time zone** field, and then click **Save**.

Add document flags

To add document flags, complete the following steps:

1. Complete the steps in <u>Editing a collection</u> to get the collection into edit mode.
2. Click the **Document flags** tab, and then click **Add flag**.
3. In the *Document flag* dialog box, name the flag, add a description, choose a flag color, and then click **Add**.
4. Repeat the previous steps to add more flags.
5. From the *Document flags* view, select **Enabled** so that the flags appear in your documents, and then click **Save** to make them available in your collection.

For more information about how to flag documents, see <u>Flag documents of interest</u>.

Identify the text field

When you analyze data with the Content Mining application, Discovery determines which field contains the *body* of the text to be analyzed. It does so by looking for the field with the highest average word count.

You can check which field is designated as the main text body field, and change it or augment it by changing the index type of another field.

1. Complete the steps in <u>Editing a collection</u> to get the collection into edit mode.

2. Click the **Fields** tab. Check the **Index type** column to find the field designated with the **Analyzable text content** index type.

   You can change the field or set more than one text field to be an **Analyzable text content** index type.

3. Click **Save**.

If you select multiple fields to analyze, you cannot see the facet analysis for only one field. To view the analysis for multiple fields, you must group them.

Group multiple text fields

1. Complete the steps in <u>Editing a collection</u> to get the collection into edit mode.

2. Click the **Contextual view** tab, and then click **Add view**.

3. Complete the following fields:

   - **Name**: The name or label of your grouped view.
   - **Id**: The alphanumeric ID that Discovery uses when you submit a text query. For example, `ans1`.
   - **Fields**: The text fields that have the **Analyzable text content** setting applied. Select one or multiple text fields that you want to group for facet analysis.

4. Click **Add**.

   Repeat this task if you want to add more text fields that you want to group for facet analysis.

5. Click **Save**.

Now you can return to the data analysis page for your collection. From the **Facet analysis** panel, you can click **Contextual view selection** to see the text fields that you grouped. You can select one of the text fields to view the facet analysis for that field.

Enrich your collection

Discovery provides built-in natural language processing models, such as the *Entities* enrichment that can recognize mentions of commonly known things, such as business or location names and other types of proper nouns. You can apply these built-in NLP enrichments to your collection.

You can also apply a document classifier enrichment that you created in the Content Mining application to your collection.

Alternatively, you can apply enrichments that were built in other projects in the same service instance to the collection in your content mining project. For example, you can apply a dictionary or text classifier that was built in another project in the same service instance to your collection.

To apply enrichments to your collection, complete the following steps:

1. Complete the steps in Editing a collection to get the collection into edit mode.
2. Click the **Enrichment** tab, and then select the enrichments that you want to apply to your collection.
3. Click **Save**.

Analyzing CSV files

You can add the data that you want to analyze as a comma-separated value (CSV) formatted file.

The content mining project works well with CSV files. When your CSV file is ingested, each row in the spreadsheet is stored as a separate document in the collection index. Each column becomes a root-level field in the document.

Follow these guidelines when you create a CSV file for use in the project:

- Add each record that you want to analyze as a row in the spreadsheet.

- Include a column for each significant data point.

- Specify column headers.

  The root-level field that is added to the document is given the column header name. If no header exists, hardcoded names, such as *column_0* and *column_1*, are applied to the columns. Specify column names to ensure that the resulting document fields have meaningful names.

- If you want to find trends over time, be sure that each record has some date information that can be used to plot the information on a timeline.

  Discovery recognizes the following date formats automatically:

  ```
  yyyy-MM-dd'T'HH:mm:ssZ
  yyyy-MM-dd'T'HH:mm:ssXXX
  yyyy-MM-dd'T'HH:mm:ss.SSSZ
  yyyy-MM-dd'T'HH:mm:ss.SSSX
  yyyy-MM-dd
  M/d/yy
  yyyyMMdd
  yyyy/MM/dd
  ```

  If you store dates in other formats, you can add the format to the list of supported formats.

  From the Discovery user interface, open the *Manage collection* page. Click your collection tile. From the *Manage fields* page for the collection, add a format to the **Date formats** field. Specify a date format that is supported by the Java SimpleDateFormat class.

  For example, if your records store only year values for dates, add `yyyy` to the supported date formats list. You can then set the data type for the field that contains a year value to *Date*, and reprocess your collection. As a result, an occurrence of `2019` in the date field is stored as `2019-01-01T05:00:00Z` in the index.

Sample CSV file

The following image shows an excerpt from a CSV file with data that is well suited for analysis with the Content Mining application. The data comes from 2010 traffic records that are published by the National Highway Traffic Safety Administration (NHTSA). Each record includes car make, model, and year information, the date of the traffic incident, and text from the driver's statement, along with other useful data points.

MAKETXT	MODELTXT	YEARTXT	CRASH	FAILDATE	FIRE	COMPDESC	CITY	STATE	DATEA	LDATE	MILES	CDESCR
TOYOTA	SIENNA	2010	N	20100101	N	ENGINE AND ENGINE COOLING	AVON	IN	20100101	20100101	3000	ENGINE SPEED CONTROL , IT DOESN'T GO UP SMOOTHLY FROM
FORD	EXPLORER	2002	N	20100101	N	LATCHES/LOCKS/LINKAGES	LOUISVILLE	KY	20100101	20100101	176000	2002 FORD EXPLORER DOOR LOCKS WILL NOT FUNCTION PROPI
FORD	FREESTAR	2005	N	20100101	N	POWER TRAIN:AUTOMATIC TRANSMISSION	NILES	MI	20100101	20100101	55000	WE WERE IN MY WIFE'S 2005 FORD FREESTAR DRIVING HOME FR
MERCEDES BENZ	E430	2000	N	20100101	N	AIR BAGS:FRONTAL	VIRGINIA BEACH	VA	20100101	20100101		ON E-CLASS MERCEDES, PASSENGER SEAT HAS FUNCTION TO C
CHEVROLET	IMPALA	2007	N	20100101	N	POWER TRAIN:AUTOMATIC TRANSMISSION	TEMPE	AZ	20100101	20100101	40000	TRANSMISSION "SLIPS" THEN ENGAGES HARD. HAS PROGRESS
JEEP	LIBERTY	2002	N	20100102	N	WHEELS	AF	UT	20100102	20100102		FRONT LUG NUTS LOOSEN ON 2002 JEEP LIBERTY. THIRD TIME (
JEEP	GRAND CHEROKEE	2002	N	20100101	N	ELECTRICAL SYSTEM:IGNITION	MINNEAPOLIS	MN	20100101	20100101	98400	KEY WON'T TURN IN IGNITION, STEERING WHEEL LOCKED, CAR
JEEP	GRAND CHEROKEE	2002	N	20100101	N	STEERING	MINNEAPOLIS	MN	20100102	20100102	98400	KEY WON'T TURN IN IGNITION, STEERING WHEEL LOCKED, CAR
CHEVROLET	HHR	2007	N	20100102	N	SERVICE BRAKES, HYDRAULIC	MOORESVILLE	NC	20100102	20100102	30000	WITH 61,000 MILES ON MY 2007 CHEVY HHR, I AM HAVING TO RE

Figure 1. Sample CSV file

For more information about the sample data, see https://www.nhtsa.gov/data/traffic-records.

Creating a report

If you discover insights as you analyze your data, you can save and share them with others by creating a report. A report consists of snapshots and notes about the analysis.

Take a snapshot

1. Click the camera icon from the dashboard toolbar.

   > **Note:** You can also take a snapshot of the document preview. When you select one or more documents, only the selected documents are stored and displayed in the report. When no selection is made, all documents in the current page are stored and displayed in the report.

   the snapshot are displayed in the *Report* pane, which is a temporary store for snapshots.

   This store is cleared when the browser is refreshed or another collection is opened.

3. From the menu icon of the snapshot's thumbnail, you can enter comments, or delete the snapshot. You can also edit comments later.

4. Choose thumbnails that you want to add to a new report, and then click **Create**.

Create a report

To create a report, complete the following steps:

1. From the *Report* pane, click **Create**.

2. On the *Basic* tab, name and date the report.

3. **Optional**: On the *Comments* tab, edit the title of the analysis result, and enter a comment.

4. Review the preview on the *Preview* tab.

5. When you're done editing, click **Save**.

Your report is added to the *Report* tab on the application launch page. From the *Actions* menu, you can copy the link for your report to share it with others.

Exporting data

If you discover insights as you analyze your data, you can export the data to share with others or analyze further in another business insights tool, for example.

You can export your data as a CSV file or you can generate a separate JSON file for each record.

To export your data, complete the following steps:

1. Submit a search to find the documents of interest.

2. Click **Show documents** to open the *Documents* view, and then click the *Export* icon ⬆ in the toolbar.

3. Complete the appropriate steps for the format in which you want to export the data.

   - If you want to export the data in JSON format, complete the following steps:

     1. Choose **Export JSON** to generate one JSON file for each record.

     2. **Optional**: You can change the following values:
        - Name. The file is named `export_document_{today's_date}` by default.
        - Encoding. `UTF-8` is used by default.
        - Choose whether to include fields and facets. They are excluded by default.

   - If you want to export the data in CSV format, complete the following steps:

     1. **Optional**: To customize the CSV output, choose **Export CSV with advanced options**.

        You can define the format of the following elements:
        - Text content field: This is the main body field (or fields, if you configured more than one field with analyzable text). You can choose to exclude it from the export. It is exported as a column for a fact table by default.
        - All other fields: You can choose to export them as columns for fact tables or export them as dimension tables.

They are excluded from the export by default.

- Facets: You can choose to export the facets as separate CSV files that can be used as dimension tables. They are excluded from the export by default.

After customizing the CSV format, click **Save**, and then click the *Export* icon from the toolbar again.

> 🔖 **Note:** If you use the same web browser to export data in CSV format again later, your saved settings are applied automatically.

2. Choose **Export CSV**.

3. **Optional**: You can change the following values:

   - Name. The file is named `export_document_{today's_date}` by default.
   - Encoding. `UTF-8` is used by default.
   - Date and time format. `Unix epoch time` is used by default.

4. Click **Export**.