# **Lab 1**

# **Hands-On Lab: Watson Discovery**

In this lab you will learn how to use Watson Discovery in order to extract insights from an unstructured publicly available dataset.

The core use case represented is the analysis of severe weather events in the United States from 2013 - 2018. Unstructured textual data such as the dataset that will be analyzed in this lab can reveal important insights that are otherwise missed by researchers lacking Natural Language Processing (NLP) capabilities. An insights engine such as Watson Discovery can quickly analyze freeform textual data by:

1. Ingesting data from a variety of formats (PDF, DOC, JSON, HTML) and performing NLP across the entire unstructured dataset.
2. Using a built-in cognitive search to query across extracted entities, concepts, keywords, relations and more in order to formulate new insights.

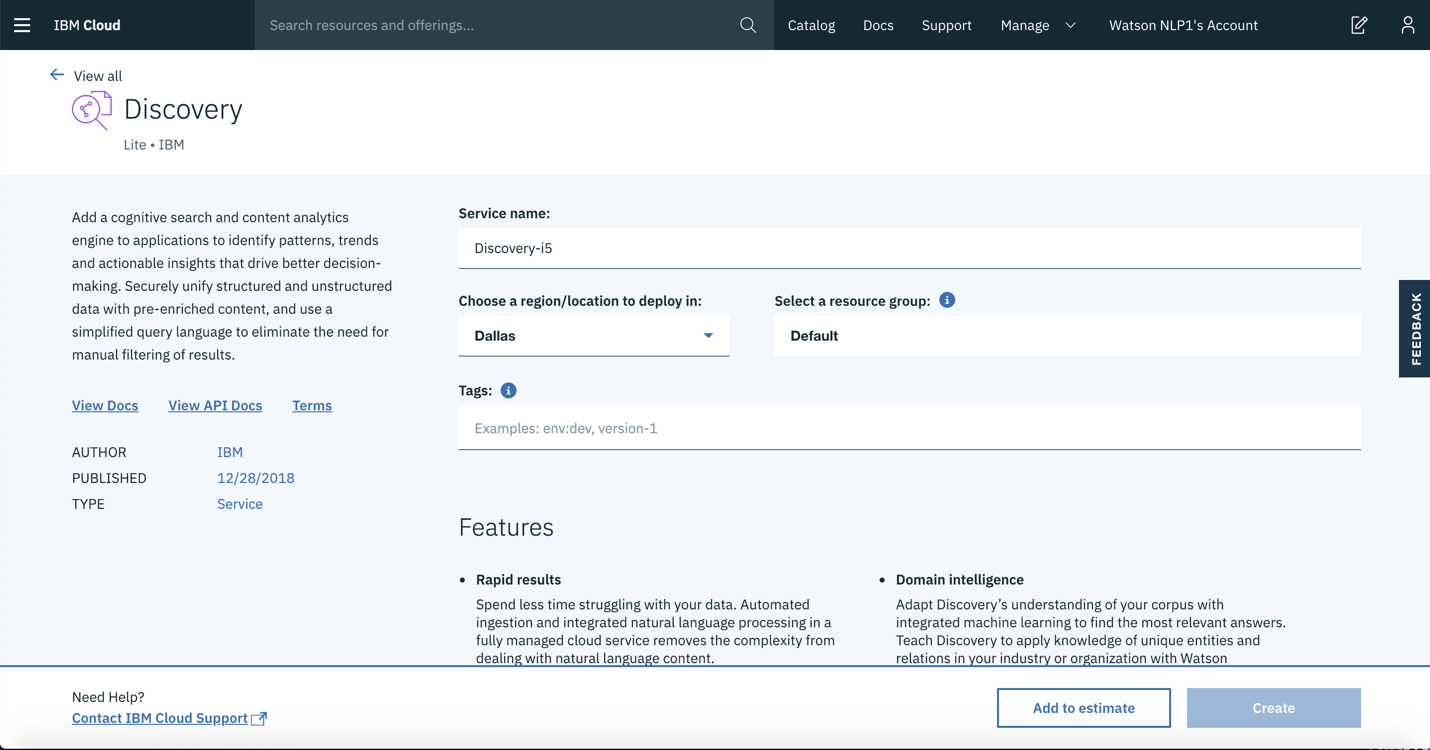
For this use case there are also a few key benefits to highlight with Discovery:

1. The ability to answer natural language questions
2. The ability to answer these questions directly from existing content instead of having to spend time and resources creating new content
3. The ability to train the service to improve how it both reads the content and finds answers to provide better results to the user

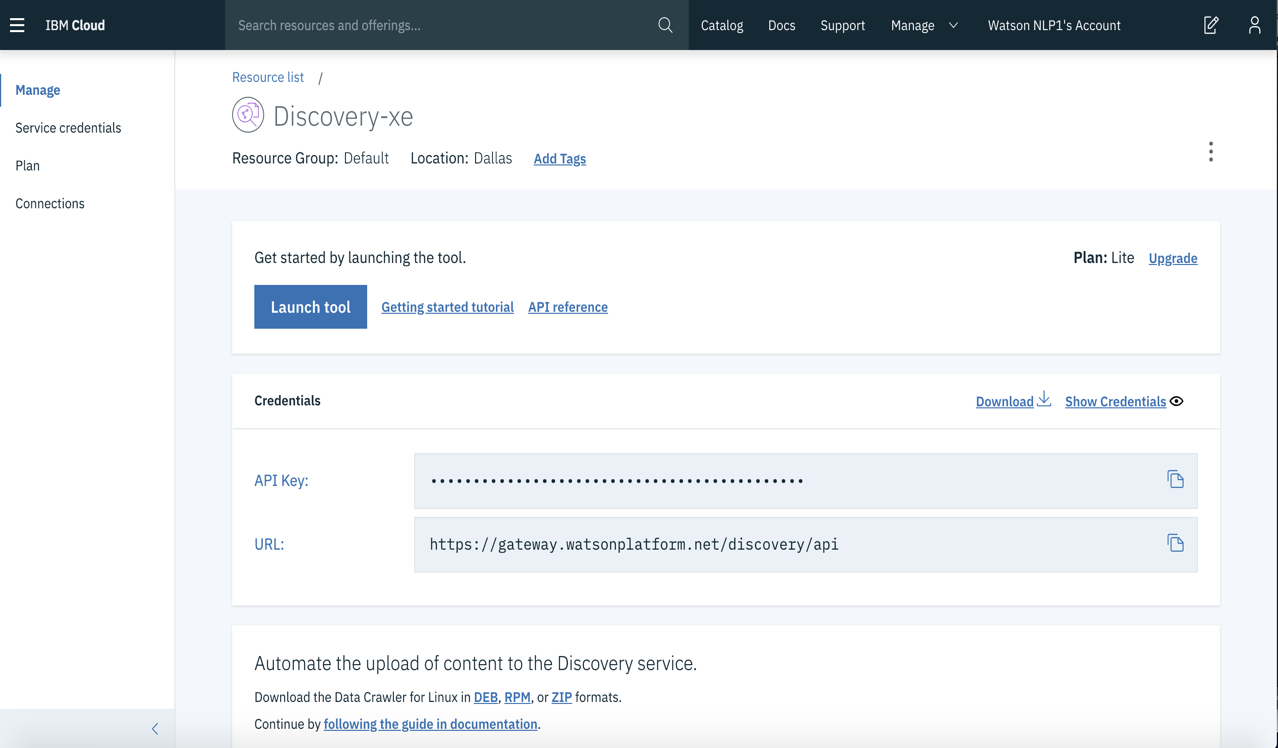
**Prerequisites:**

* IBM Cloud Account (Use <https://console.bluemix.net> to sign up for a new account before starting the labs)

**Step 1: Create and Launch Discovery Instance**

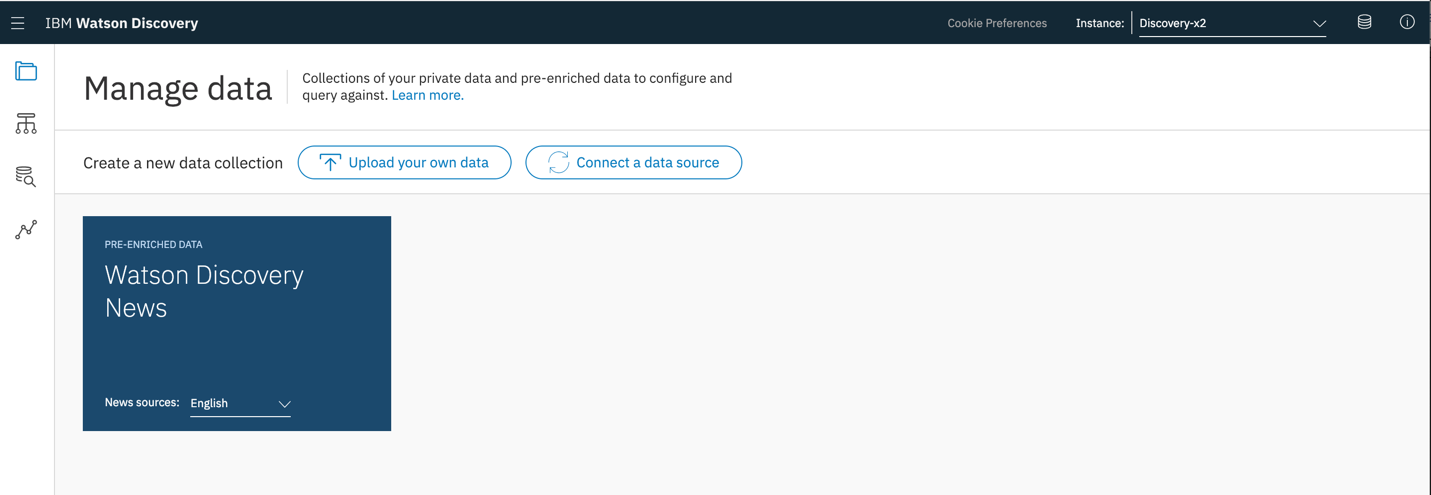
****

1. Create an instance of the Discovery service and get your credentials:
   * Log in to your IBM Cloud account.
   * Go to the [Discovery](https://cloud.ibm.com/catalog/services/discovery) page in the IBM Cloud Catalog.
   * Give your service a unique name (or keep the generated name), select the **Dallas** region and **Default** resource group.
   * Click **Create**.
2. Select your instance and launch Discovery:
   * On the [Resource list](https://cloud.ibm.com/resources) page, double click your Discovery service under **Services**

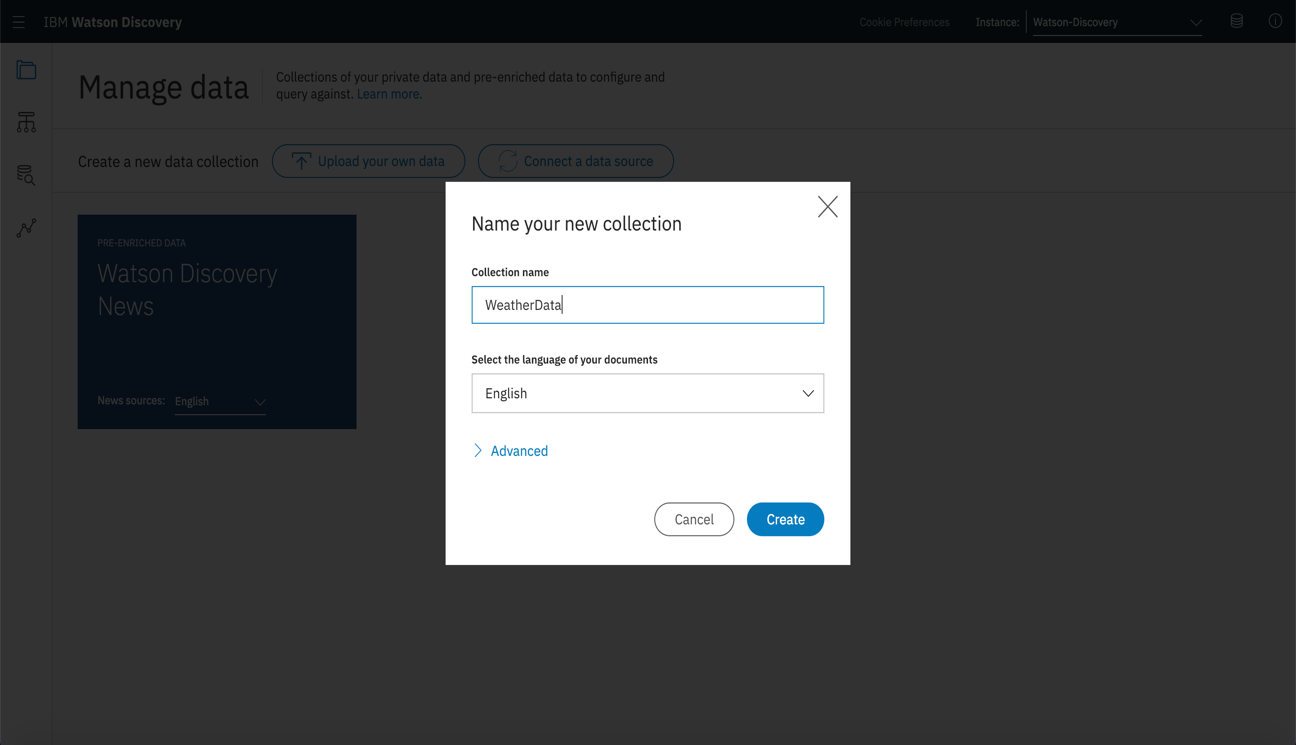


* + Click **Launch tool**

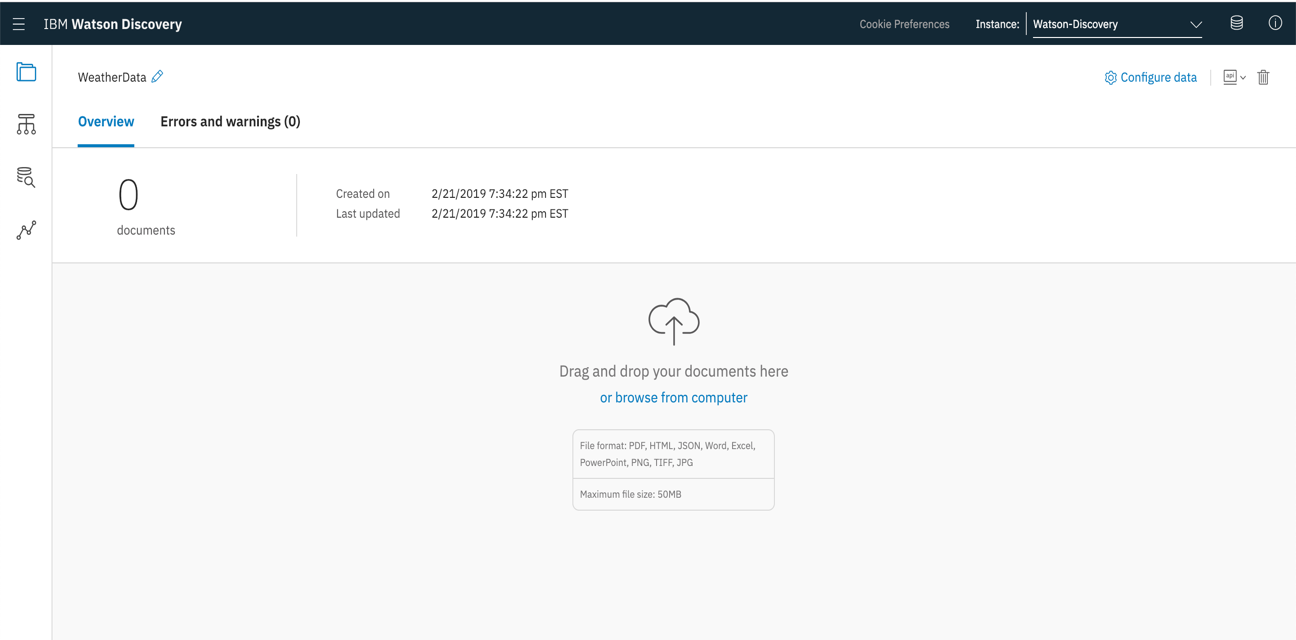
**Step 2: Create and Configure Data Collection**

****

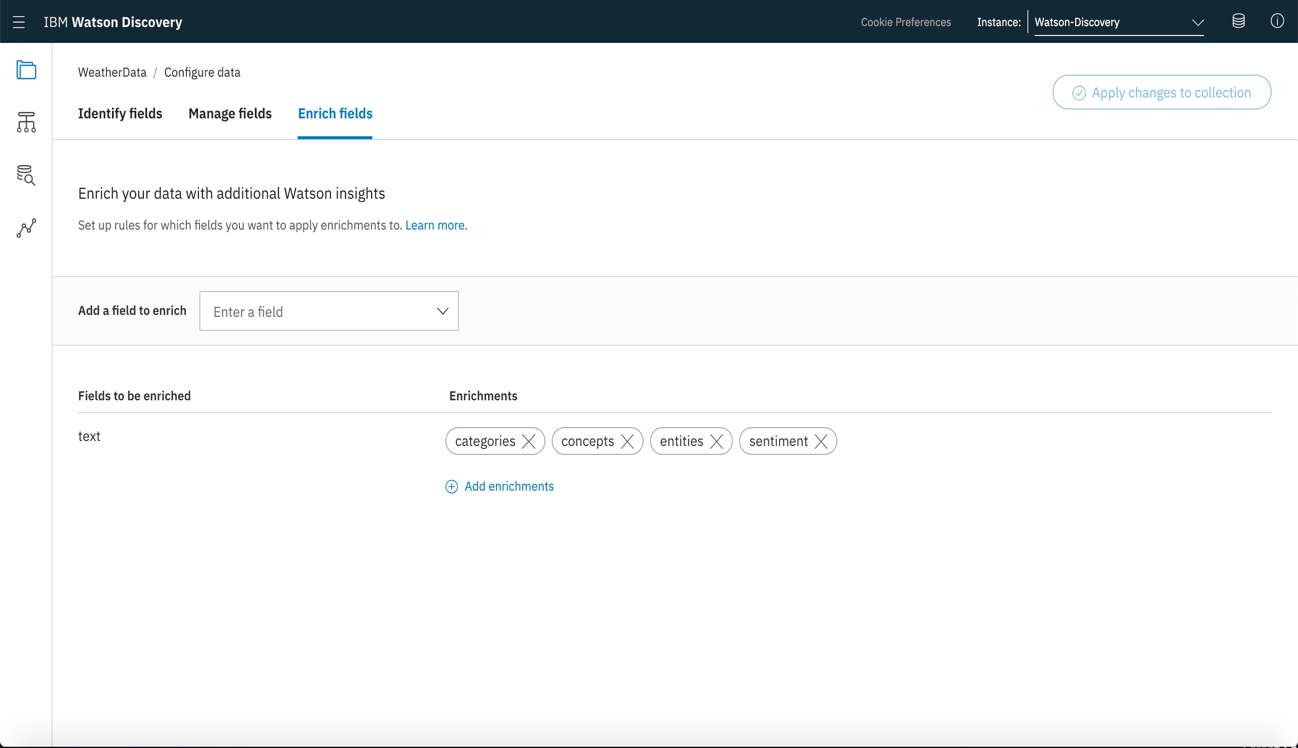
1. On the Discovery home page, click **Upload your own data** in order to start uploading your data
2. Give your collection a unique name (e.g. WeatherData), keep the language of the documents as **English** and click **Create**



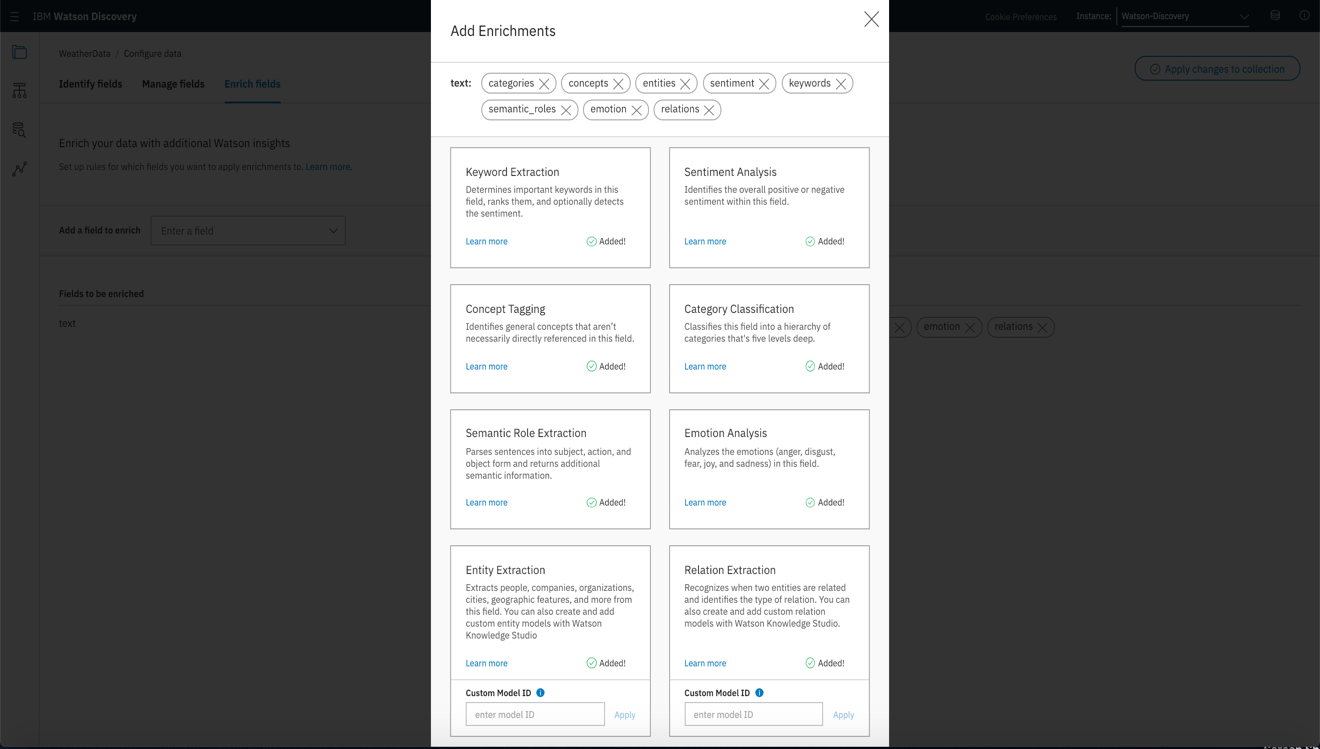
1. Before we can start uploading our data to the collection, we must first configure the collection to enrich our dataset. Click **Configure data**



1. We will be using Discovery to analyze the “text” field in our json dataset. Update the list of enrichments for our dataset by clicking **Add enrichments**



1. Select the following enrichments by clicking the **Add** button: **Keyword Extraction**, **Semantic Role Extraction**, **Emotion Analysis**, **Relation Extraction.** As a default, Sentiment Analysis, Concept Tagging, Category Classification, Entity Extraction have already been added.

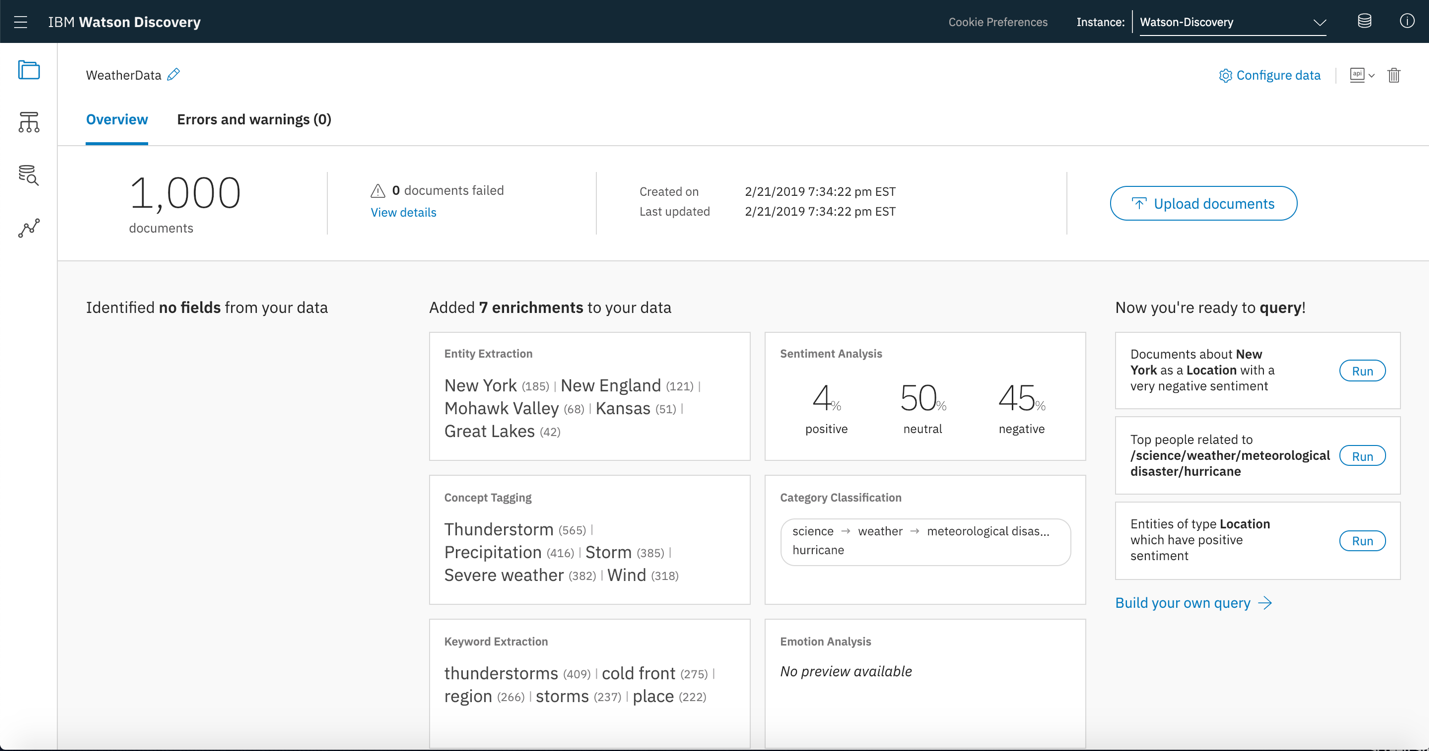


1. Click **Done** to save the list of enrichments.

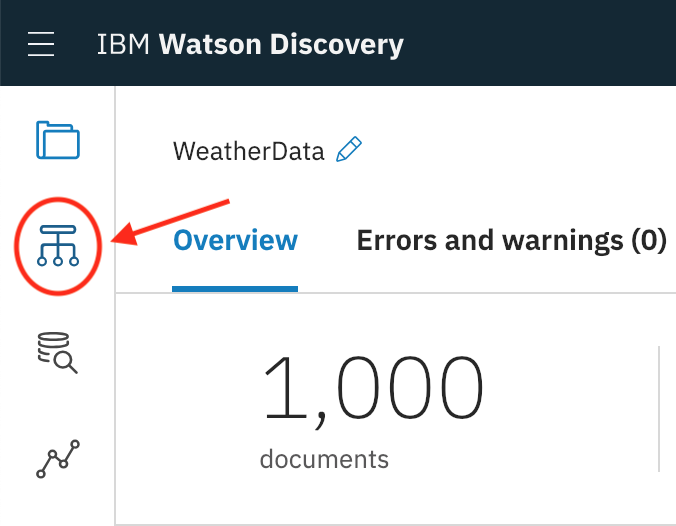
**Step 3: Download the Data and Upload to Discovery**

1. Go to <https://github.com/cerebralace/WatsonPoT/tree/master/Lab1>
2. Download the dataset by clicking on **WeatherData.zip** and then clicking **Download.**
3. Locate the downloaded zip file on your computer and unzip it to get the WeatherData folder. This folder contains data for all severe weather events tracked by NOAA in the United States from 2013 – 2018.
4. We can now upload the entire dataset into our Discovery collection. On the WeatherData collection page, click **Browse from Computer**
5. Select and click the WeatherData folder and shift select all of the files in the folder. Upon clicking **Open**, this will upload and enrich all 1000 JSON files within Discovery. This process should take at least 15 - 20 minutes to complete.

**Step 4: Overview and Document View**

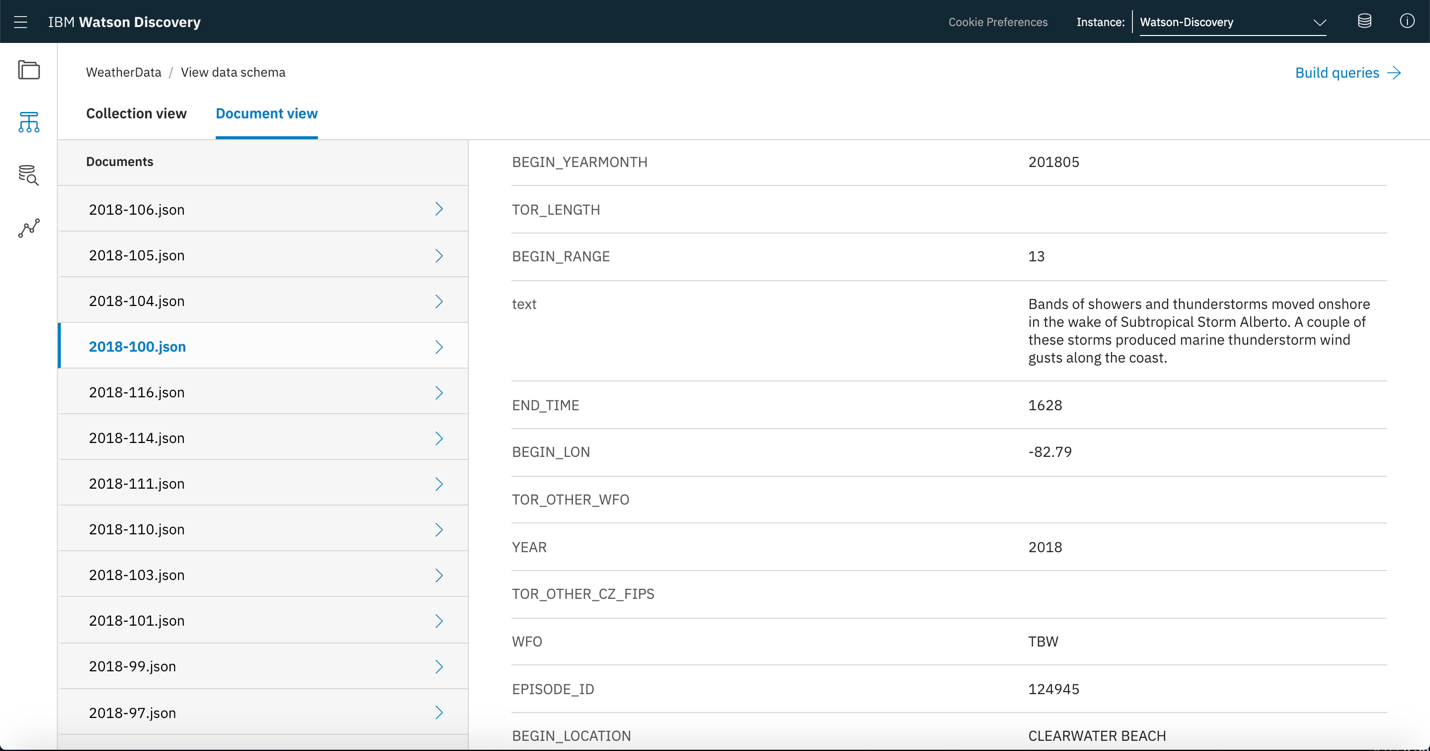
****

1. As soon as all of the JSON documents are uploaded and enriched, you should see the overview page in your collection pictured above. Since we are working with the Lite edition of Watson Discovery, we can only upload a maximum of 1000 documents. However with Advanced editions of Discovery, we can upload up to 10 million documents per month.
2. The Overview page shows the highlights from the enrichments including the top entities, concepts, overall sentiment across all documents, category classification and most common keywords. We can also see options to run sample queries on the right side of the page. We will be building and running our own queries in Step 5.
3. Now click on the data schema icon on the left toolbar in order to access the Document view.



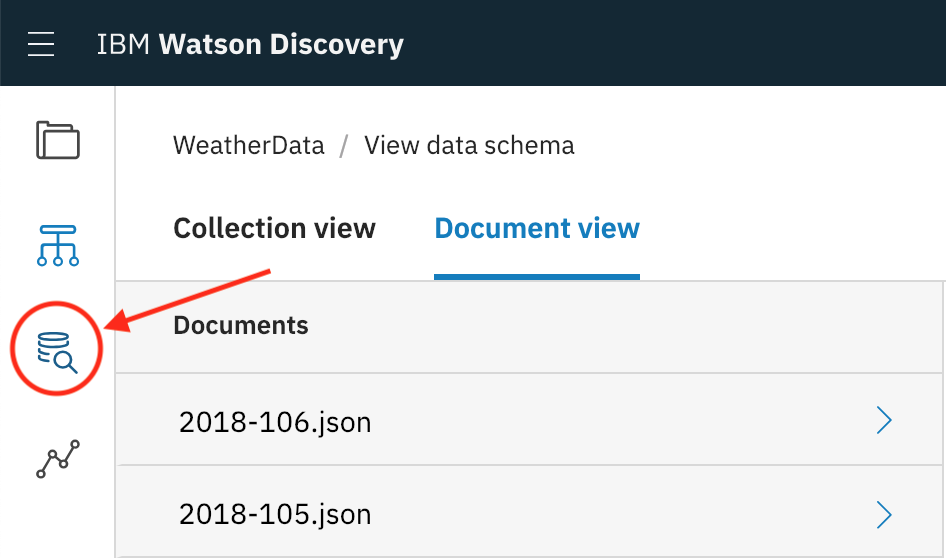
1. In the Document view, you can see a subset of the uploaded and enriched documents. Click through the various documents in order to notice that each document not only contains the original metadata fields such as “END\_DAY”, “DAMAGE\_PROPERTY”, “EVENT\_TYPE”, etc. but also newly enriched fields such as “enriched\_text.entities.text”, “enriched\_text.sentiment.document.score”, “enriched\_text.concepts.text” and more.

We will be using a combination of the original fields and the enriched fields in order to build our queries in the next step.

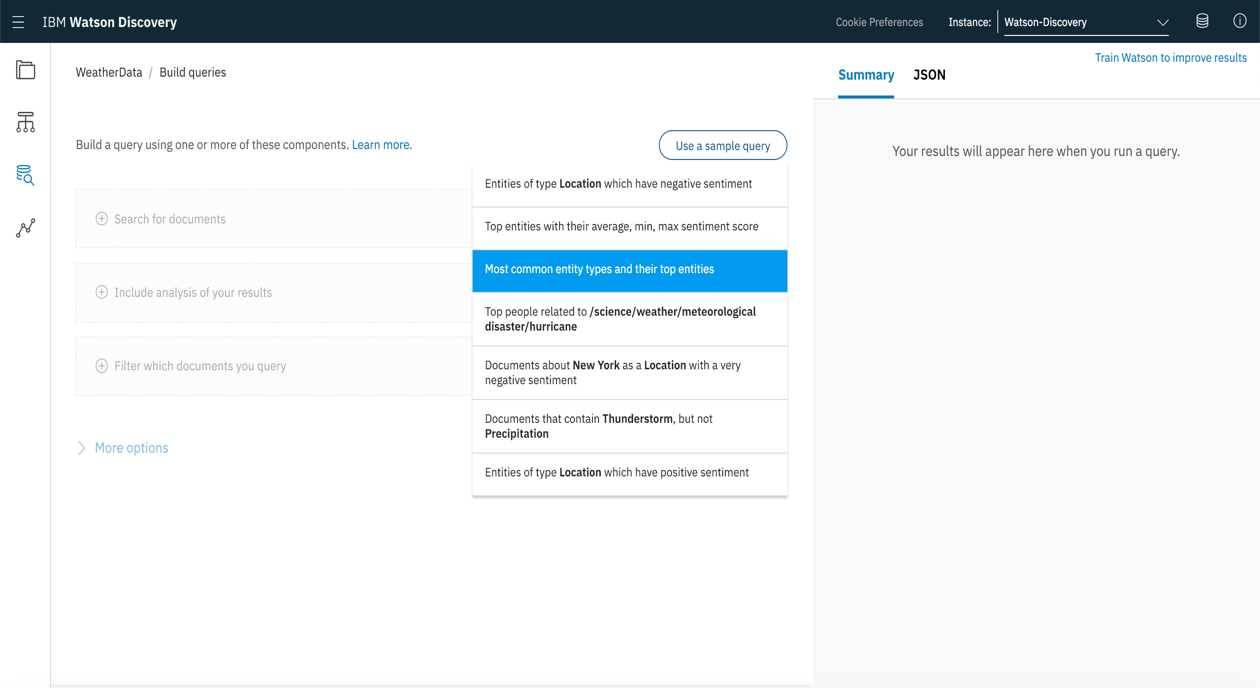


**Step 5: Build custom queries**

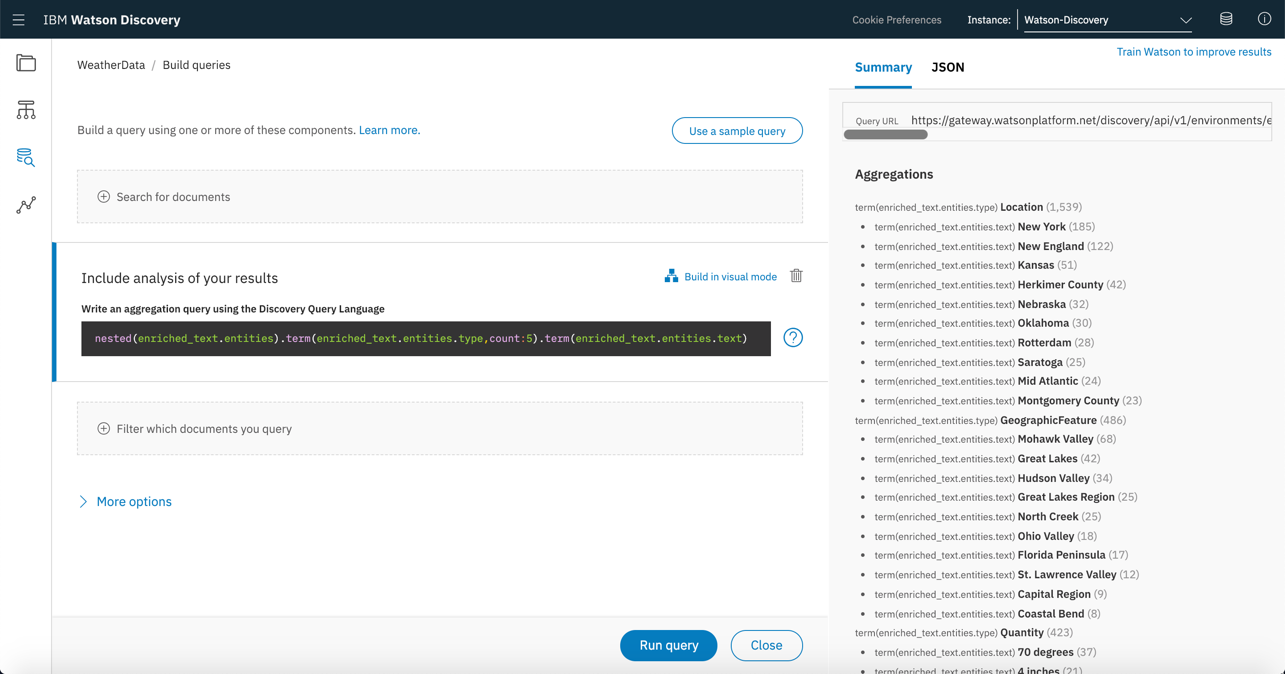
1. Click on the Queries icon on the left toolbar in order to access the **Build Queries** page.



1. On the Build queries page, we can build custom queries in order to test the cognitive search capability of Watson Discovery. Let’s start with the sample created queries.
2. Click **Use a sample query** and select **Most common entity types and their top entities**



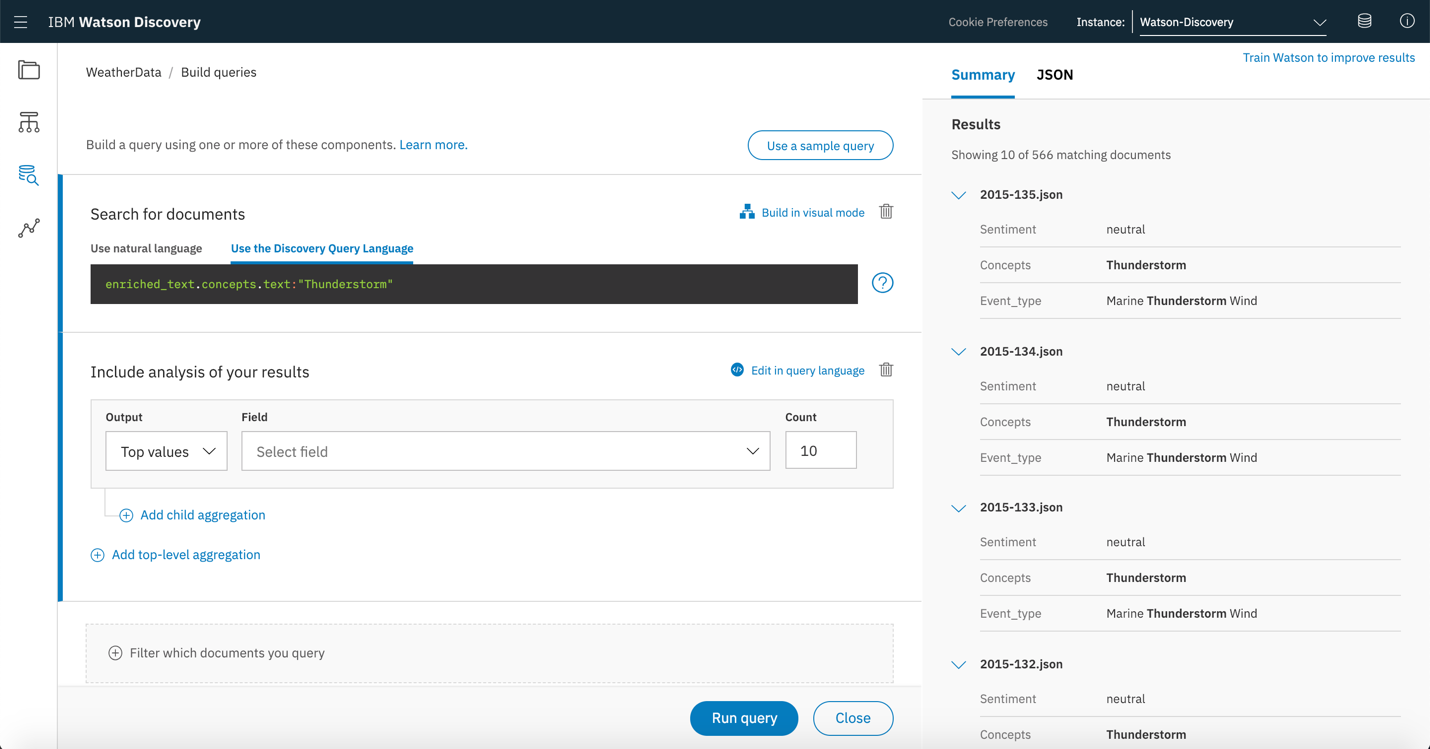
1. This sample query returns the aggregated query created using the Discovery Query Language underneath **Include analysis of your results** as well as the results of the query on the right side of the page. The results show the top five entity types across all documents as well as specific entities in each category. For example, the most common entity type was Location and this included specific entities such as New York, New England, Kansas, etc.



1. Let’s build a basic query. After clicking the **Trash icon** next to **Include analysis of your results**, click **Search for documents** and then select **Use the Discovery Query Language:**

* Click the **Field** drop-down and select **enriched\_text.concepts.text**, for the **Operator** choose **contains**, then enter the **Value** of **Thunderstorm.** The query *enriched\_text.concepts.text:Thunderstorm* will display under the Visual Query Builder.
* Alternately, you can also click **Edit in query language**, then **Use the Discovery Query Language**. Enter *enriched\_text.concepts.text:”Thunderstorm”* into the **Enter query here** field.

1. Click **Run query**. This query should return relevant passages that mention the Thunderstorm concept and a list of 566 matching documents. In order to retrieve the complete text of these documents, you can always click on the JSON link at the top of the right side of the page.

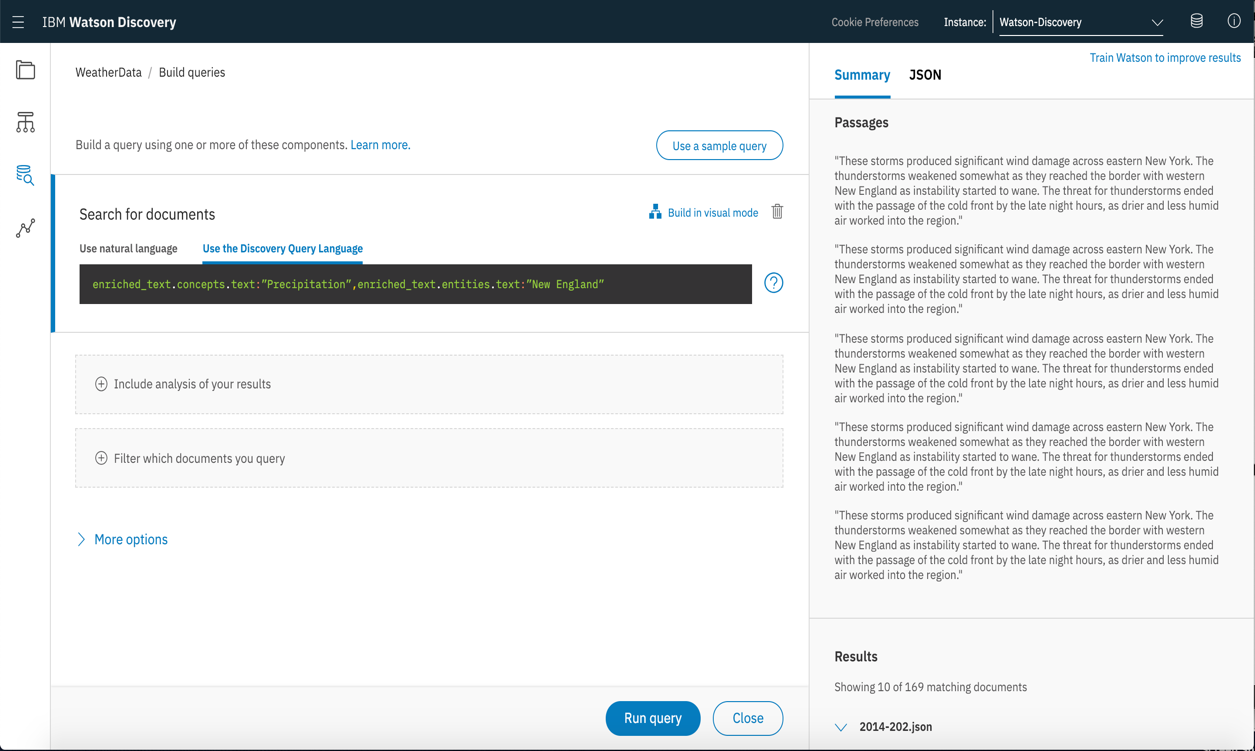


1. Now let’s build a more complex query. In order to return all documents that mention the concept of precipitation in New England, click **Search for Documents**, **Use the Discovery Query Language** then:

* Click the **Field** drop-down and choose **enriched\_text.concepts.text**, for the **Operator** choose **contains**, then enter the **Value** of **Precipitation**. Click **Add rule**, then for the **Field** choose **enriched\_text.entities.text**, for the **Operator** choose **contains**, then enter the **Value** of **New England**.

The query **enriched\_text.concepts.text:Precipitation,enriched\_text.entities.text:New England** will display under the **Visual Query Builder**.

* Alternately, you can also click **Edit in query language, Use the Discovery Query Language** and type the following query: *enriched\_text.concepts.text:”Precipitation”,enriched\_text.entities.text:”New England”*
* Click **Run query** to see the results of the query on the right.



1. Finally, let’s create a combined query. You can combine query parameters together to build more targeted queries. Let’s try using both the filter and query parameters to return documents about tornados in 2018. The filter parameter will narrow down the results to only documents from 2018, and then the query parameter will return all results about tornados, in order of relevance.

* Clear your previous queries by clicking on the Trash icon and then select **Filter which documents you query**
* Click the **Field** drop-down and choose **YEAR**, for the **Operator** choose **contains**, then enter the **Value** of **2018**. The query *YEAR:2018* will narrow down the documents to only those from 2018.
* Under **Search for Documents**, click **Use the Discovery Query Language**. Then click the **Field** drop-down and choose **enriched\_text.concepts.text**, for the **Operator** choose **contains**, then enter the **Value** of tornados.

The query *enriched\_text.concepts.text:”tornados”* will return all documents that include the concept of tornados, and those documents will be ranked in order of relevance.

* Click **More options**, then **Fields to return** and choose **Specify**. Select **text**. This will return only the text of the relevant articles and exclude everything else.
* Click **Run query** to see the results of the combined query on the right.

