IBM Training

Student Exercises

Lab-2: Create a knowledge management system and develop a COVID-19 vulnerability index Hands-On Lab

Table of Contents

Prerequisites *3*

Download the data files to the Desktop 3

Introduction *4*

Objectives *4*

Exercise 1: Create a Discovery collection *4*

Exercise 2: Upload the documents *8*

Entity Extraction 8

Relation extraction 8

Keyword extraction 9

Category classification 9

Concept tagging 9

Semantic Role extraction 9

Sentiment analysis 9

Emotion analysis 9

Exercise 3: Add the entity model from Knowledge Studio *12*

Exercise 4: Perform Custom Entity Extraction *17*

Exercise 5: Calculate the COVID-19 vulnerability index *19*

Exercise 7: Create a collection for a COVID-19 publication *26*

Exercise 8: Perform Smart Document Understanding *29*

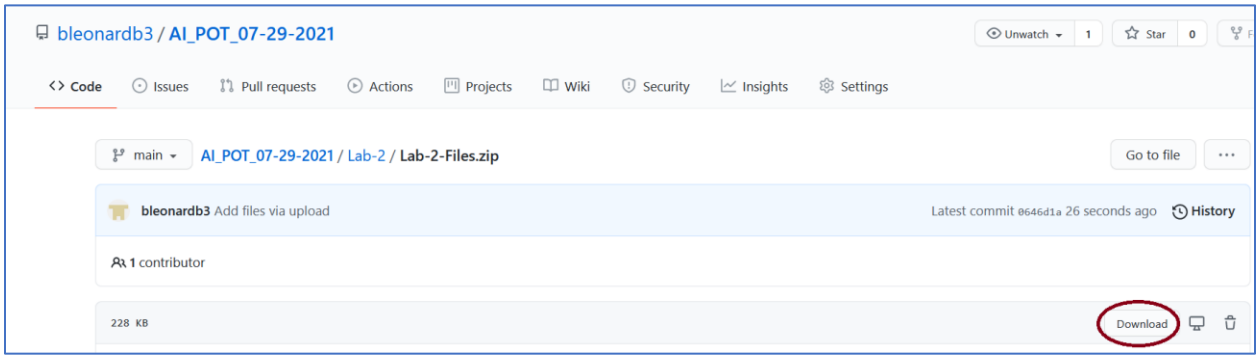
Exercise 9: Create and run Natural Language Queries *36*

Exercise 10: Improve accuracy with Relevancy Training *42*

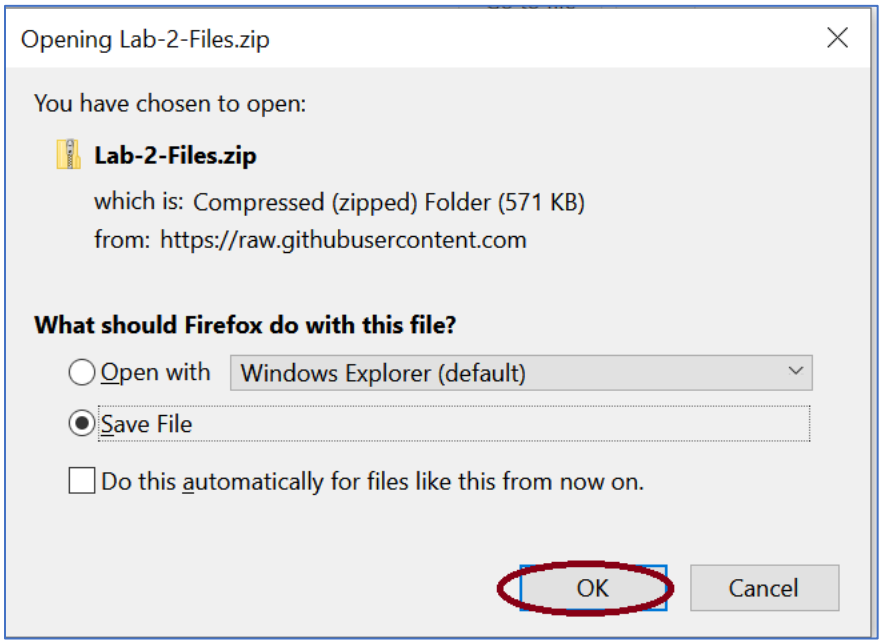
Prerequisites

Download the data files to the Desktop

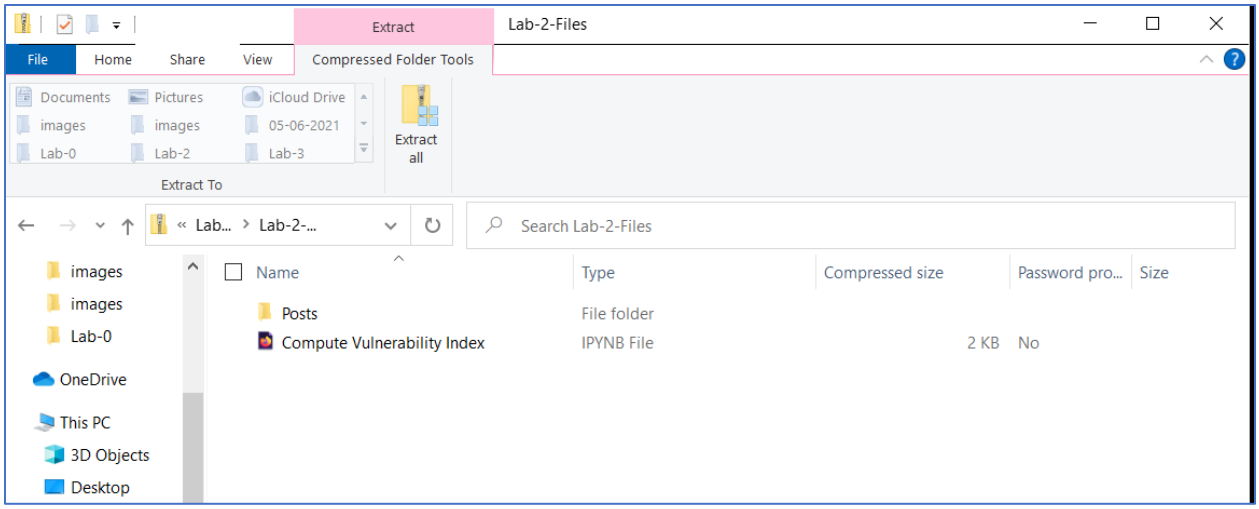
1. Click on https://github.com/bleonardb3/AI_POT_07-29-2021/blob/main/Lab-2/Lab-2-Files.zip
2. Click on the **Download** button.



3. Click **OK**.



4. Extract the file contents. You should have the one directory and one file extracted.



Introduction

In this lab you will create a knowledge management system (KMS), train the KMS to generate knowledge and analyze information to create a COVID-19 vulnerability index. IBM Watson Discovery will be used to develop and train the KMS.

Objectives

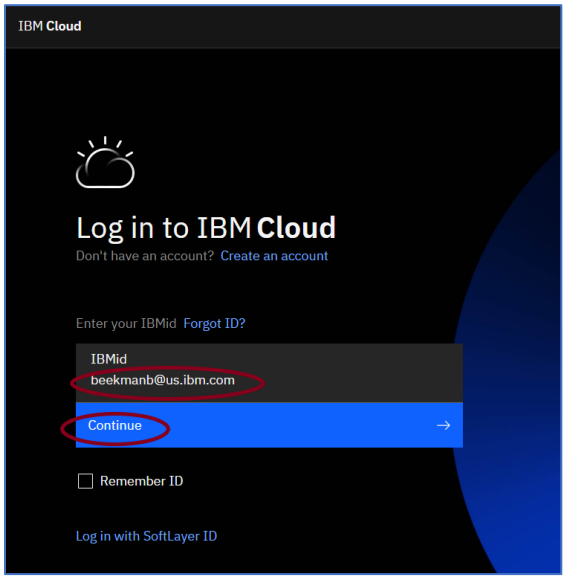
The goal of this lab is to familiarize the user with the Watson Discovery service. Watson Discovery is an enterprise AI search technology that leverages machine learning, including natural language processing (NLP), to retrieve specific answers to your questions and analyze trends and relationships buried in enterprise data. By integrating a machine learning annotator from Watson Knowledge Studio (which we created and deployed in Lab 1), Watson Discovery can be trained on the language of your domain to perform customized NLP. The Watson Discovery service can be deployed on any cloud or on-premises environment.

After completing this lab, you will be able to perform the following exercises with Discovery:

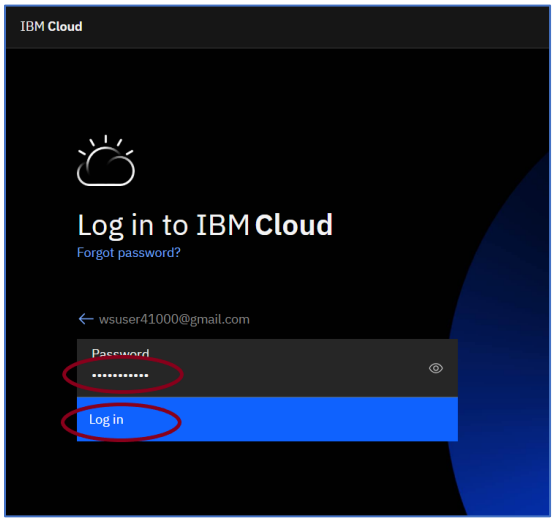
1. Create a Discovery collection
2. Upload the documents
3. Add the entity model from Knowledge Studio
4. Perform custom entity extraction
5. Retrieve the analyzed files using the Discovery API
6. Calculate the COVID-19 vulnerability index
7. Create a collection for a COVID-19 publication
8. Perform Smart Document Understanding
9. Create and run Natural Language Queries
10. Improve accuracy with Relevancy Training

Exercise 1: Create a Discovery collection

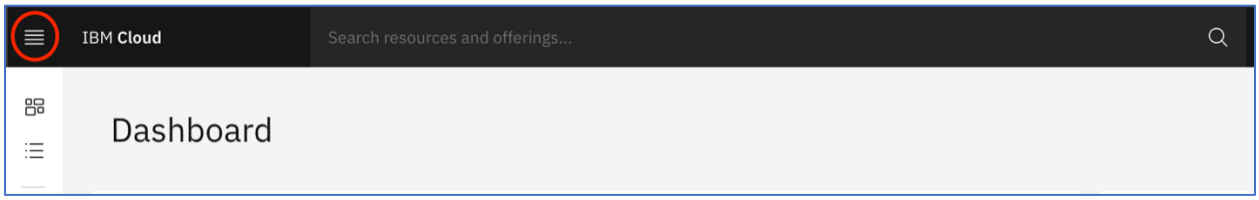
1. Log into your IBM Cloud account by typing **cloud.ibm.com** into the URL address bar of your Firefox or Chrome browser.
2. If you have been logged out, enter your **IBMid** and click **Continue.**



3. Enter your **Password** and click **Log in**.

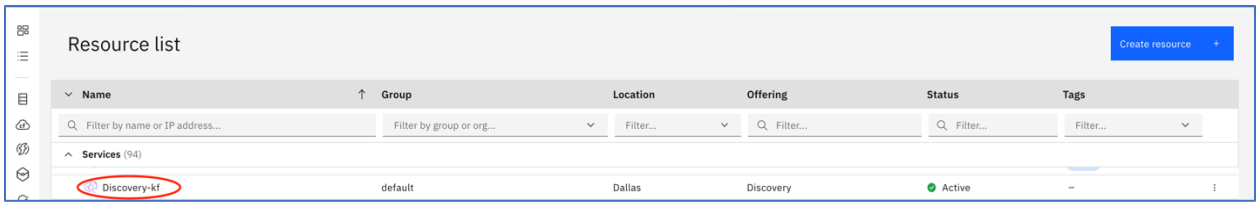


4. Click on the **icon.**

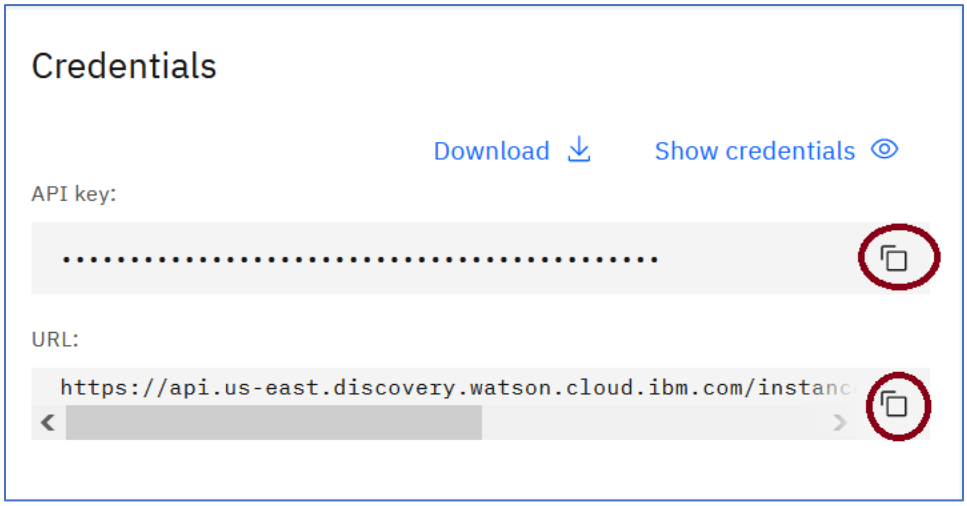


5. Select **Resource List** from the drop-down menu.

6. Under services, click on the **name of the Discovery instance** that you created in Lab 1.

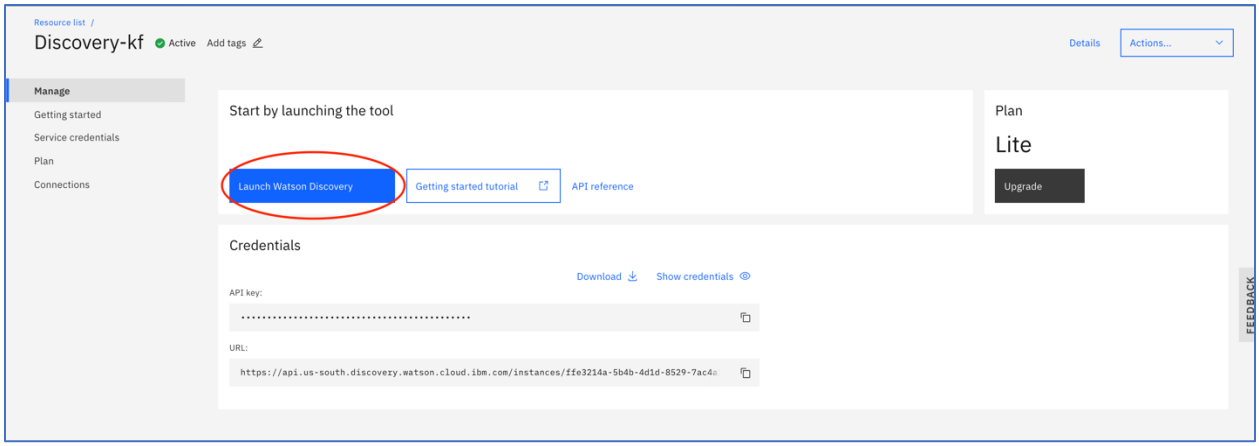


7. On the Manage screen, click on the copy icon adjacent to the **API key** and **URL** and paste these into a text editor (e.g Notepad) for later use. We will be using these credentials when calling the Discovery API later in the lab.

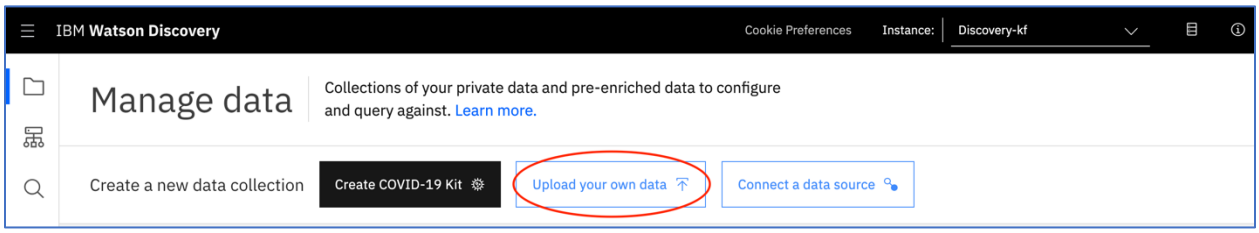


8. Click **Launch Watson Discovery** in order to start your Discovery instance.

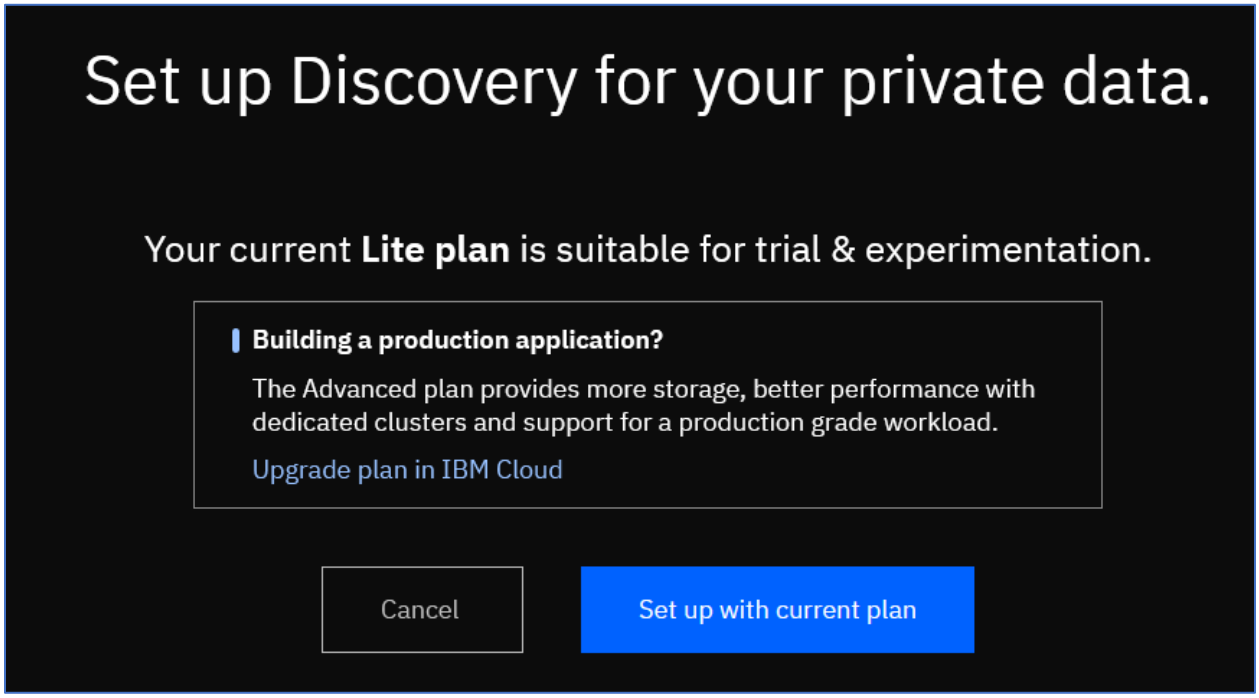
A collection is a grouping of your content within the environment. You must create at least one collection to be able to upload your content. Collections are comprised of your private data. But, Discovery News, a public data set that is pre-enriched with cognitive insights, is also included with Discovery. You can use it to query for insights; for example: news alerts, event detecting, and trending topics in the news; that you can integrate into your applications. See Watson Discovery News for more information. You cannot adjust the Discovery News configuration or add documents to this collection.



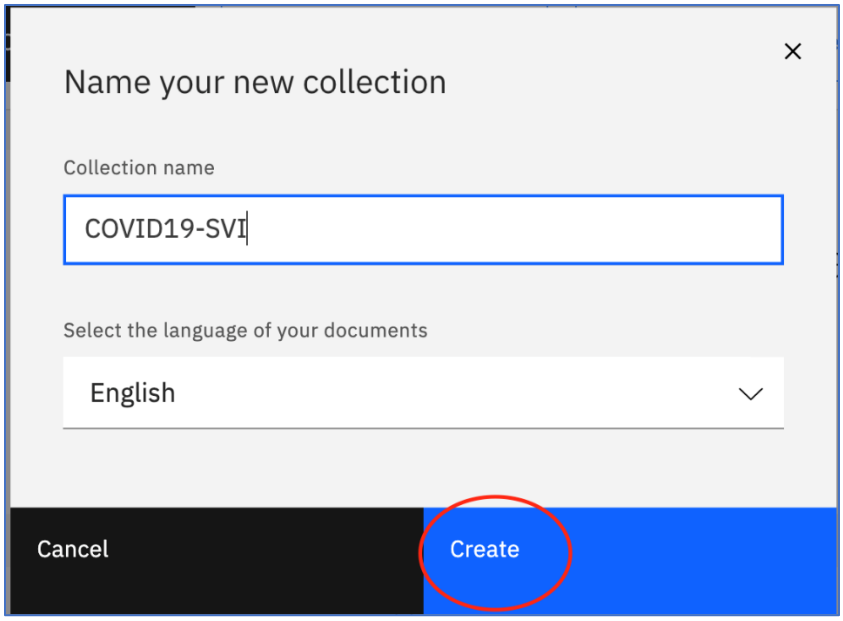
9. On the Manage data screen, click **Upload your own data** to create a new collection.



10. Click **Set up with current plan**.



11. Give your collection a name of **COVID19-SVI** and click **Create**.



Exercise 2: Upload the documents

Now that we have created a collection, we can upload all of our social media posts to this collection. In Watson Discovery, a collection stores all the relevant documents (preferably in the same file format) and is subsequently used to perform passage retrieval and content mining on the analyzed data set.

Discovery enriches (adds cognitive metadata to) the `text` field of your ingested documents with semantic information collected by these four Watson functions - Entity Extraction, Sentiment Analysis, Category Classification, and Concept Tagging. There is a total of nine Watson enrichments available; the others are Keyword Extraction, Relation Extraction, Emotion Analysis, Element Classification, and Semantic Role Extraction. Each enrichment is briefly described below.

Entity Extraction

Returns items such as persons, places, and organizations that are present in the input text. Entity extraction adds semantic knowledge to content to help understand the subject and context of the text that is being analyzed. The entity extraction techniques are based on sophisticated statistical algorithms and natural language processing technology and are unique in the industry with their support for multilingual analysis and context-sensitive disambiguation. You can also create and add a custom entity model with IBM Watson™ Knowledge Studio as was one done in Lab-1.

Relation extraction

Recognizes when two entities are related and identifies the type of relation. You can also create and add a custom relation model with IBM Watson™ Knowledge Studio.

Keyword extraction

Important topics in your content that are typically used when indexing data, generating tag clouds, or when searching. Discovery automatically identifies supported languages in your input content, and then identifies and ranks keywords in that content.

Category classification

Categorizes input text, HTML, or web-based content into a hierarchical taxonomy up to five levels deep. Deeper levels allow you to classify content into more accurate and useful subsegments.

Concept tagging

Identifies concepts with which the input text is associated, based on other concepts and entities that are present in that text. Concept tagging understands how concepts relate and can identify concepts that are not directly referenced in the text. For example, if an article mentions CERN and the Higgs boson, the Concepts API functions identifies Large Hadron Collider as a concept even if that term is not mentioned explicitly in the page. Concept tagging enables higher level analysis of input content than just basic keyword identification.

Semantic Role extraction

Identifies subject, action, and object relations within sentences in the input content. Relation information can be used to automatically identify buying signals, key events, and other important actions.

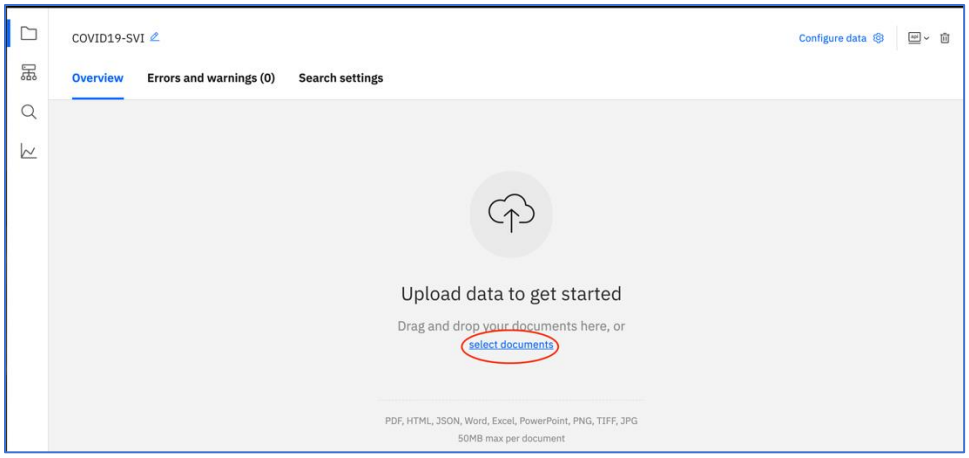
Sentiment analysis

Identifies attitude, opinions, or feelings in the content that is being analyzed. Discovery can calculate overall sentiment within a document, sentiment for user-specified targets, entity-level sentiment, quotation-level sentiment, directional-sentiment, and keyword-level sentiment. The combination of these capabilities supports a variety of use cases ranging from social media monitoring to trend analysis.

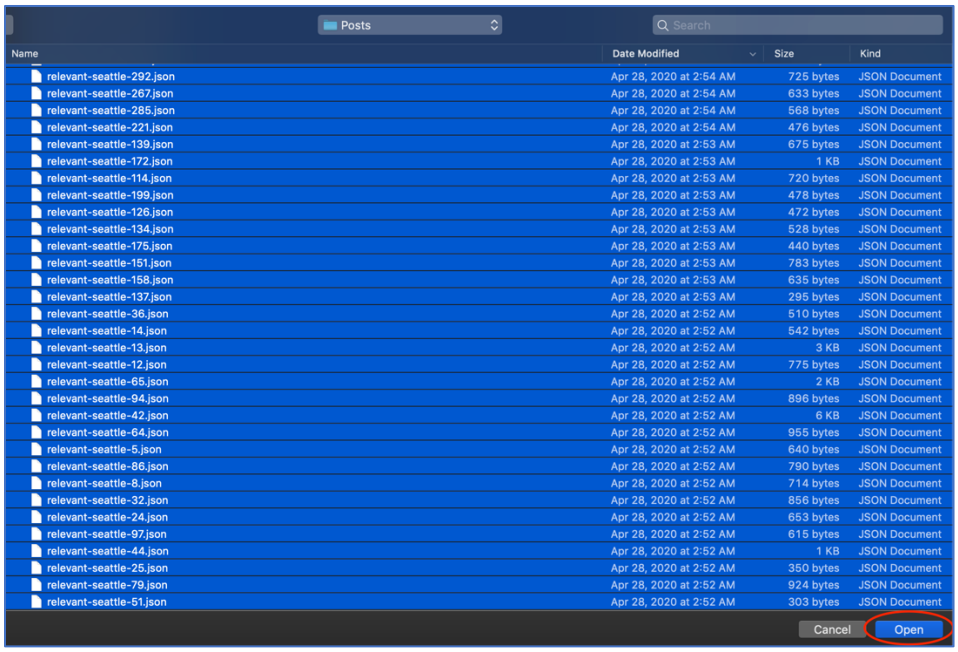
Emotion analysis

Detects anger, disgust, fear, joy, and sadness implied in English text. Emotion Analysis can detect emotions that are associated with targeted phrases, entities, or keywords, or it can analyze the overall emotional tone of your content.

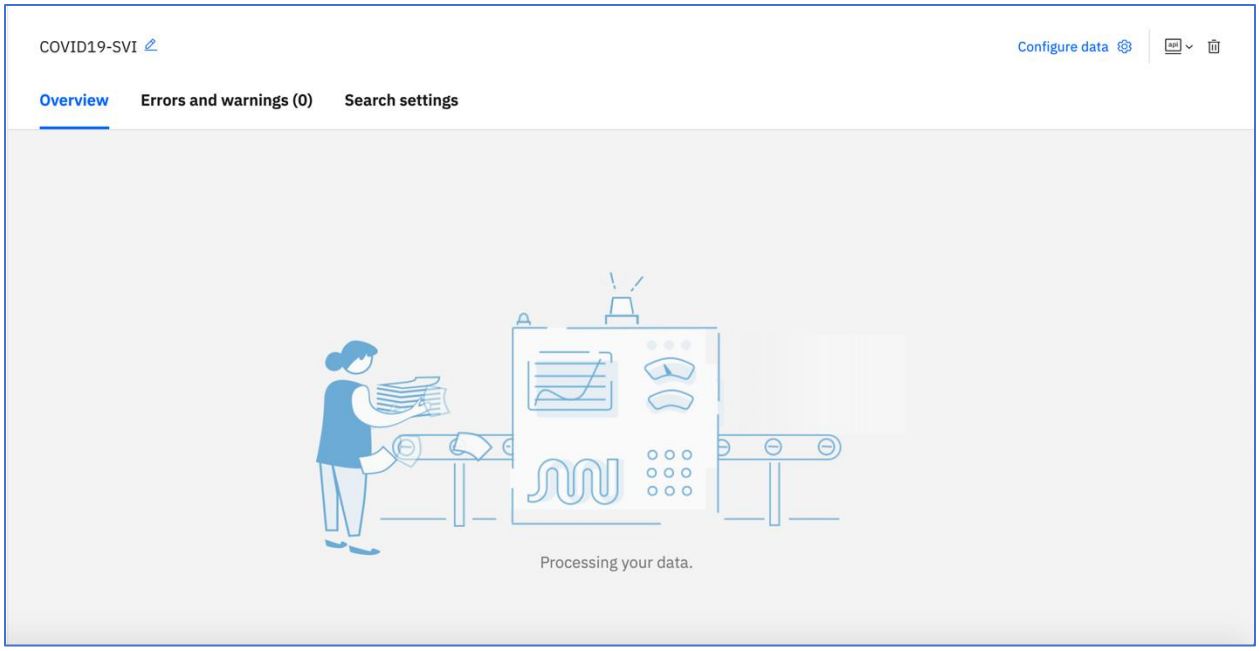
1. Click **select documents**.



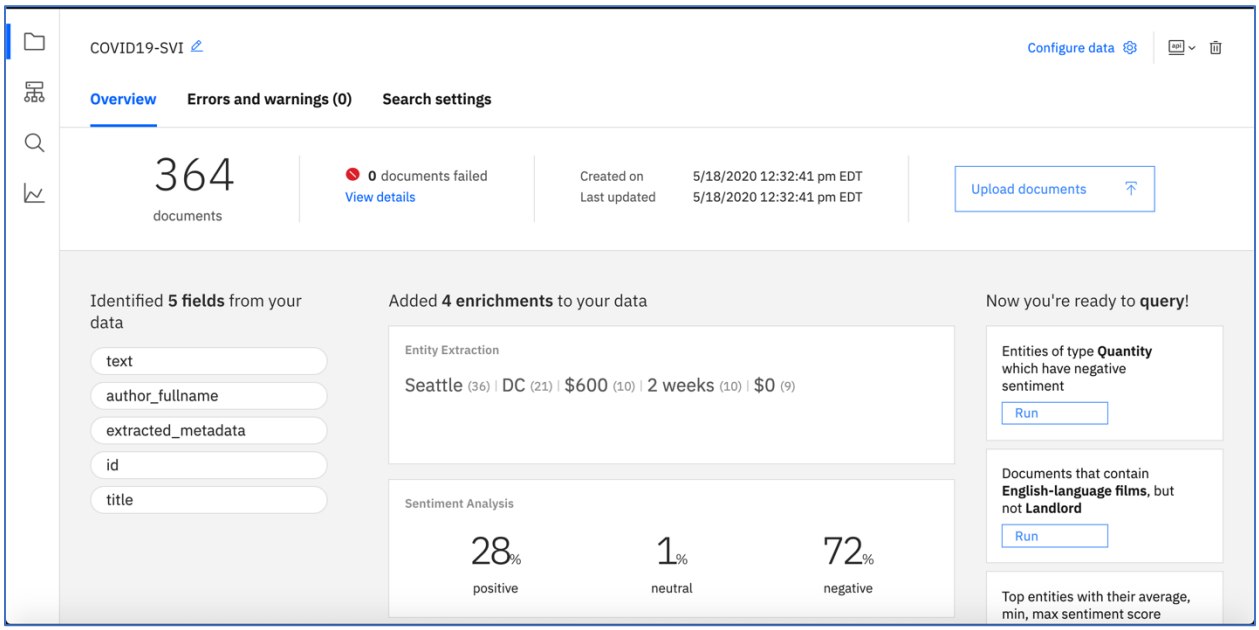
2. Navigate to the folder where you extracted the zip file downloaded from GitHub. Double-click on the **Posts** folder, **shift select all of the files** (except the README.md) in the folder and click **Open**.



3. It will take approximately 6 minutes to ingest all of the social media posts into the collection. During this ingestion process, Discovery additionally applies out-of-box enrichments, described above, such as entity extraction, concept tagging, category classification, and sentiment analysis to all the documents.



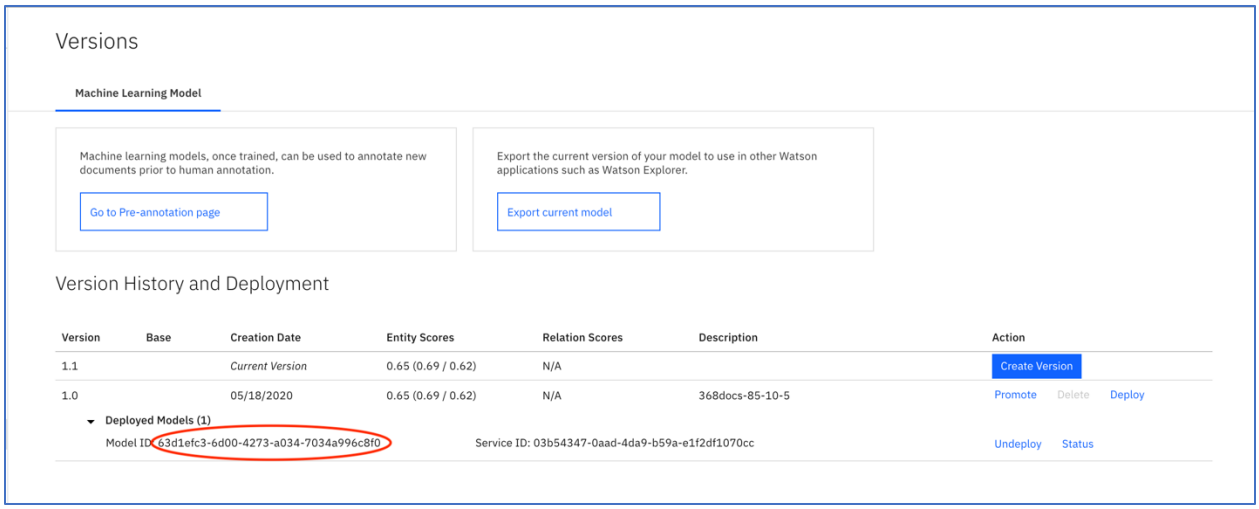
4. As soon as all the documents have been successfully ingested, you should see the following screen:



On this overview screen, you should be able to see the total number of posts in the collection, the number of fields identified per post, the top entities extracted, the overall sentiment of the documents and some sample queries that can be applied to the collection.

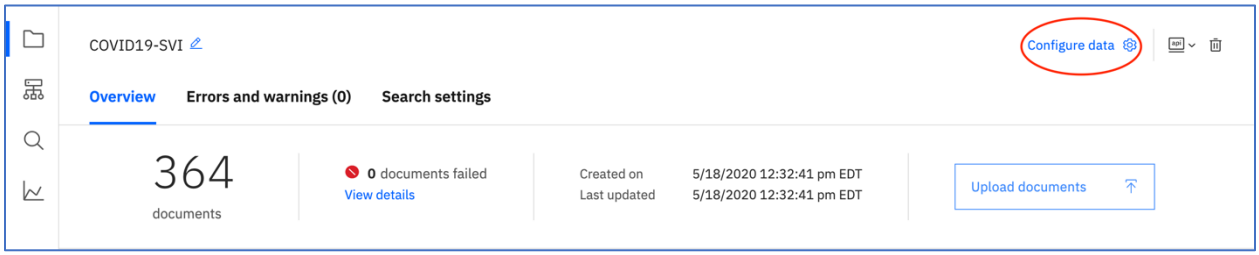
We will now configure the dataset to only apply entity extraction with the machine learning annotator from Watson Knowledge Studio. Hopefully, you saved the Model ID at the conclusion of Lab 1.

If you didn't copy it or misplaced the Model ID number, this would be a good opportunity to revisit your COVID19-Vulnerability workspace in Watson Knowledge Studio and copy the Model ID underneath Deployed Models on the Versions page.

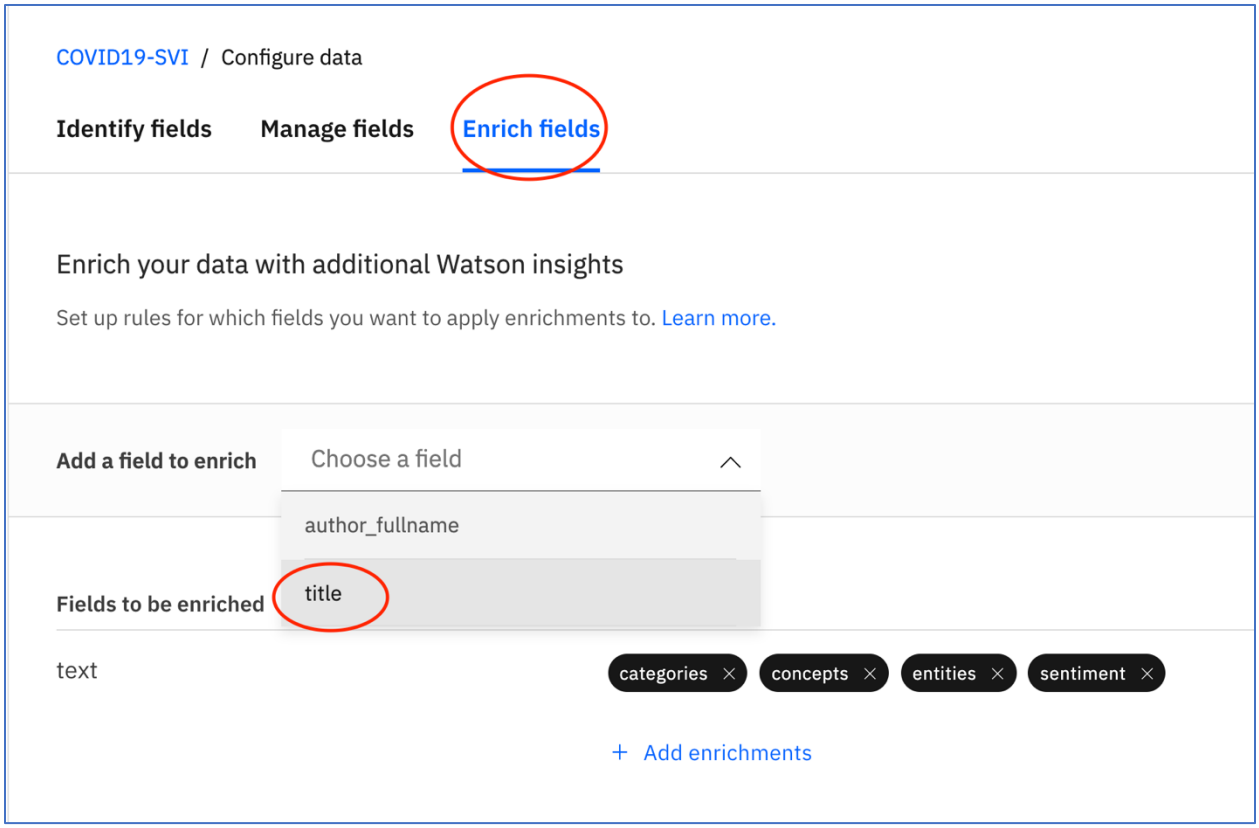


Exercise 3: Add the entity model from Knowledge Studio

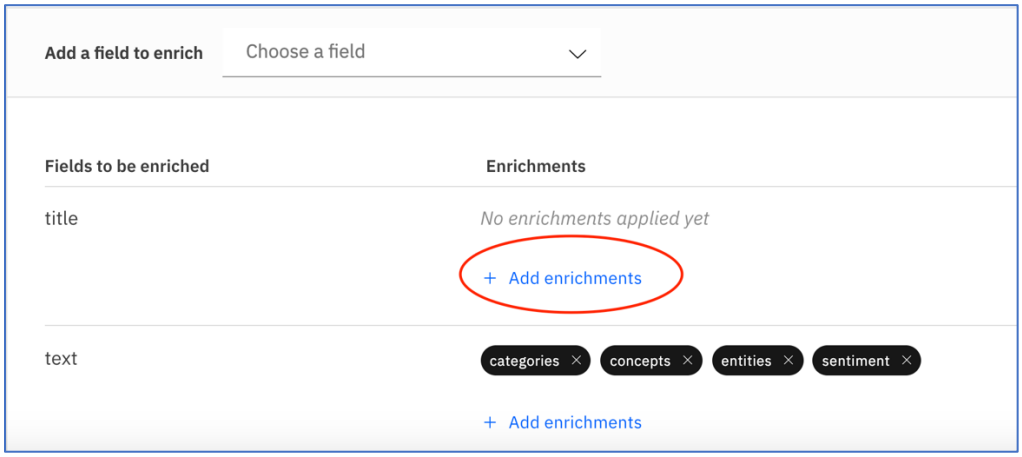
1. Click **Configure Data**.



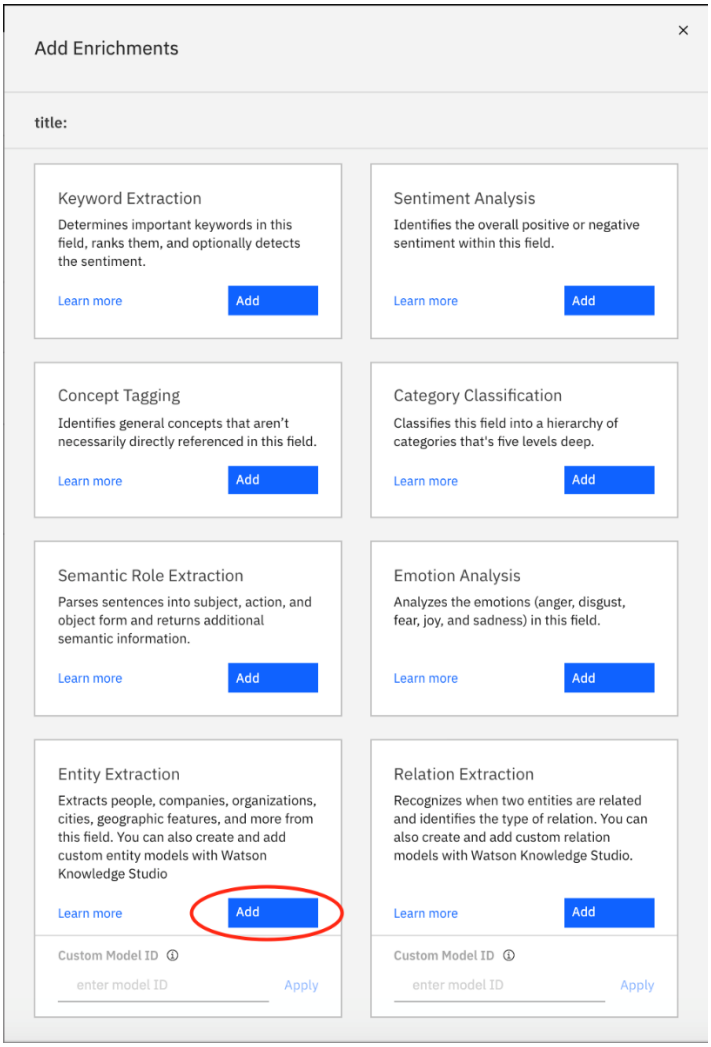
2. On the Configure Data screen, click the **Enrich fields** tab. Here we can specify the sections of our files that will be subjected to the NLP enrichments (in our case, we will only be selecting entity extraction). Click the drop-down menu next to Add a field to enrich and select **title**.



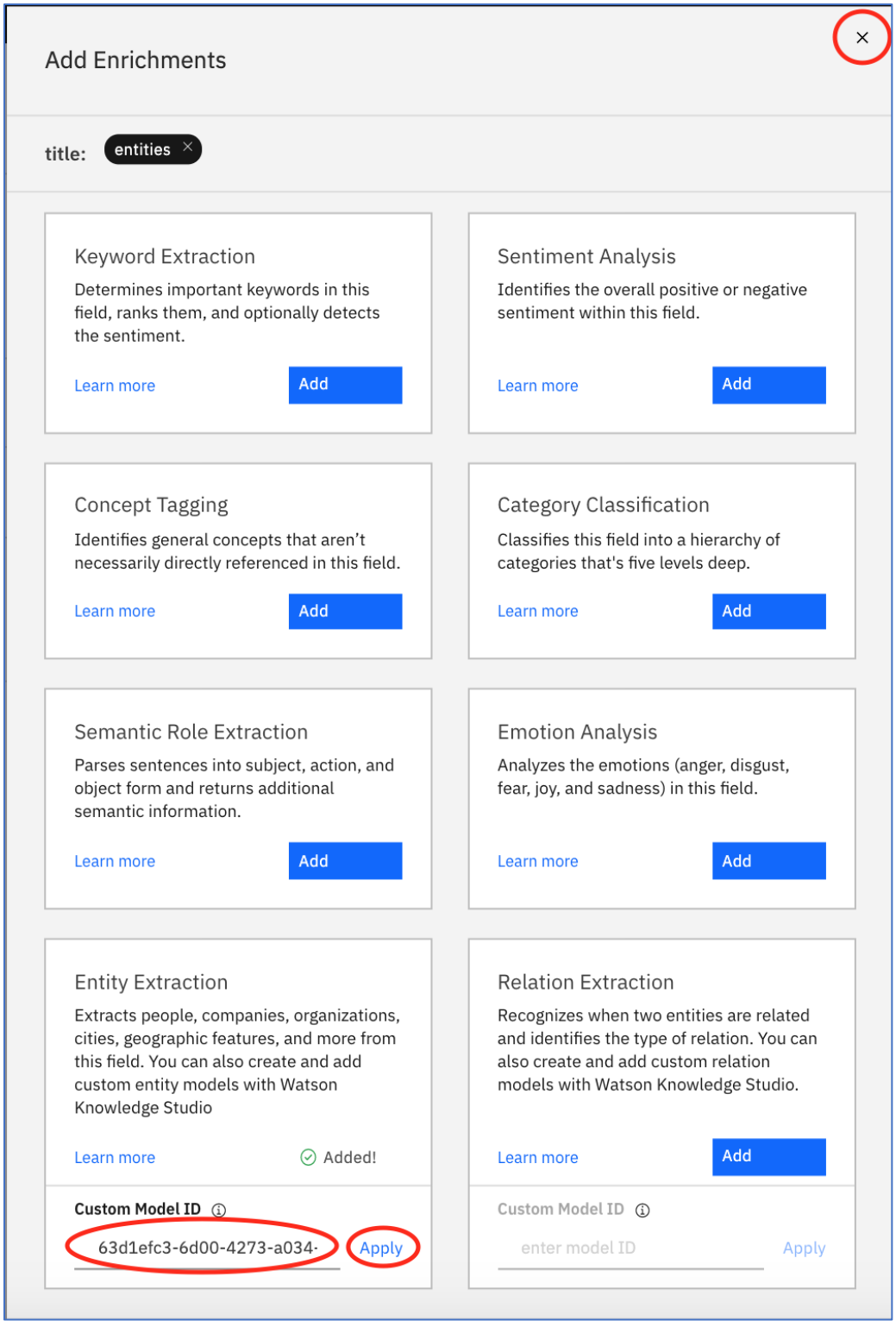
3. To the title field of each post, we will apply entity extraction. Click **+Add enrichments** on the title row.



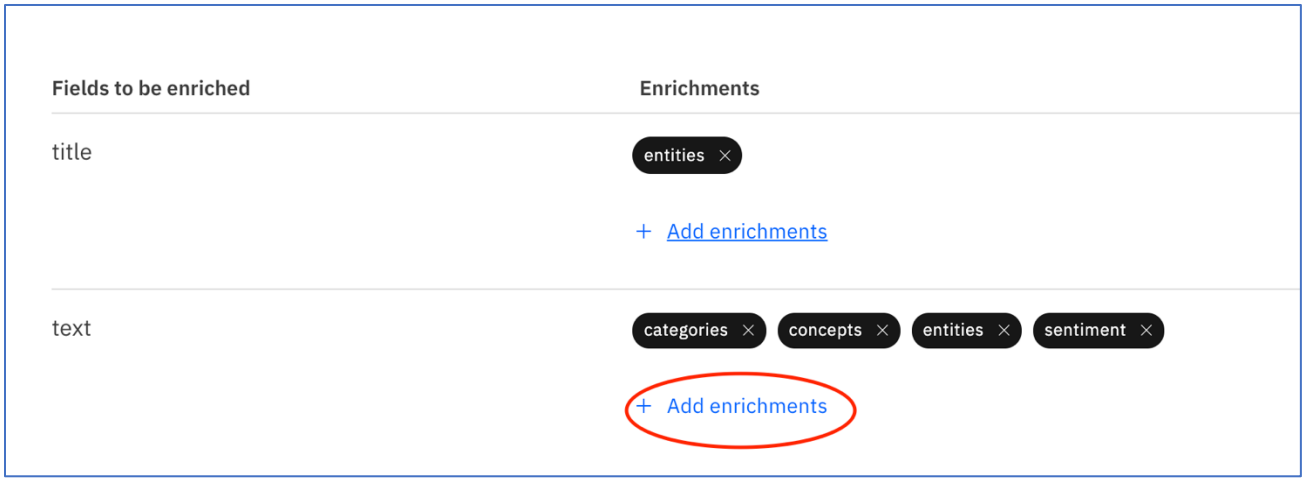
4. On the Add Enrichments pop-up screen, click the **Add** button inside of the **Entity Extraction** card.



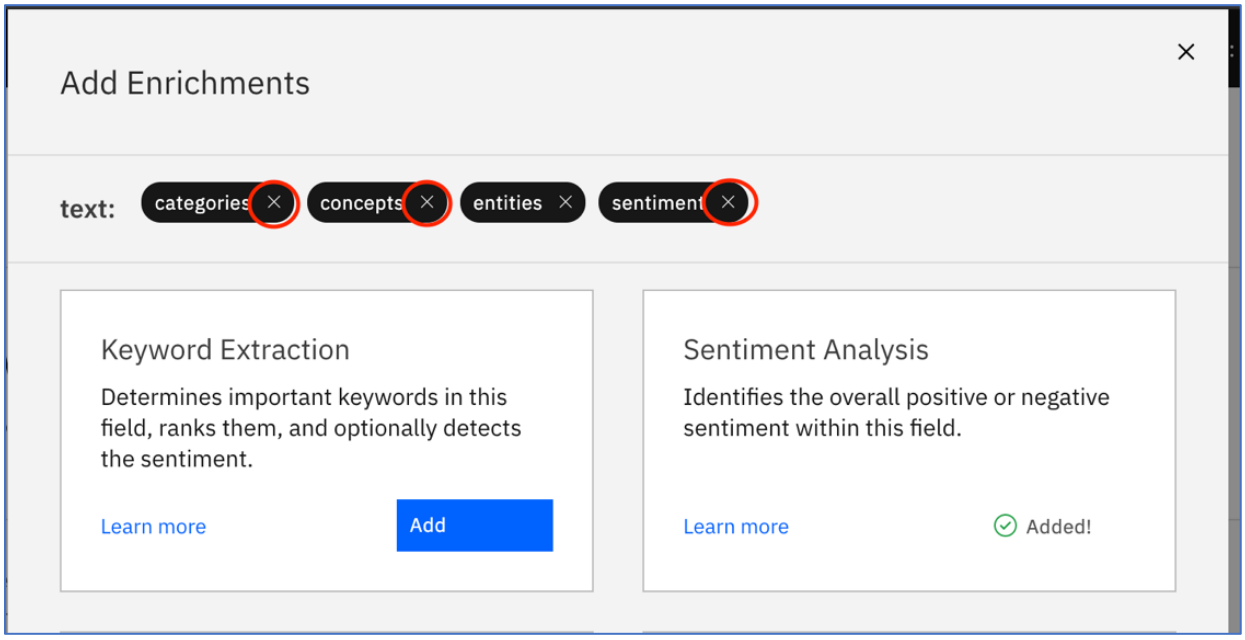
5. Paste the Model ID number for the machine learning annotator from Lab 1 underneath **Custom Model ID** and click **Apply**. Then click the **x** in the top right corner of the pop-up screen to save your changes.



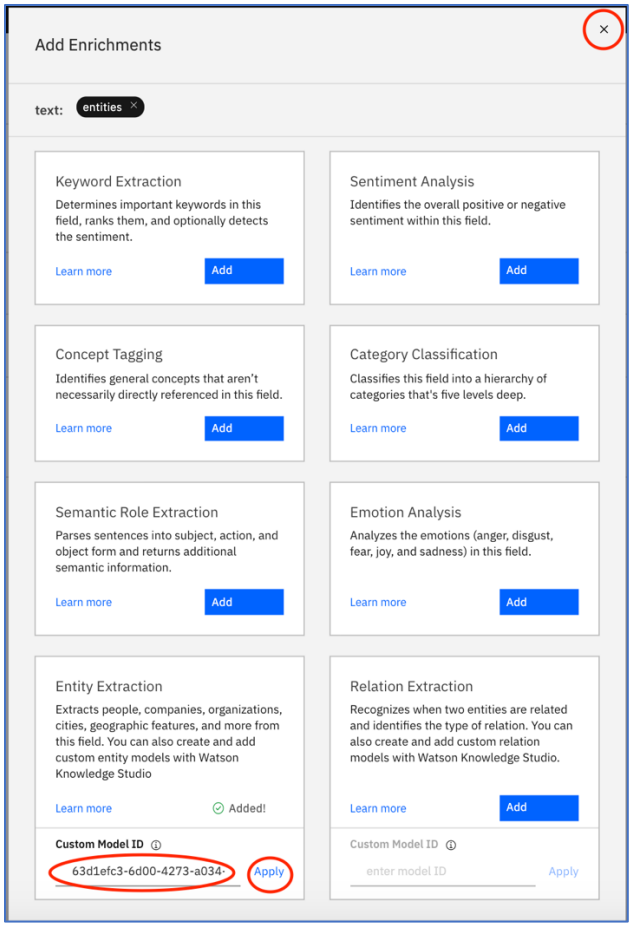
6. Now let's do the same thing for the text field of each post. Click on **+ Add enrichments** on the text row.



7. Remove the NLP enrichments of categories, concepts and sentiment by clicking on each **x** next to **categories**, **concepts** and **sentiment**.

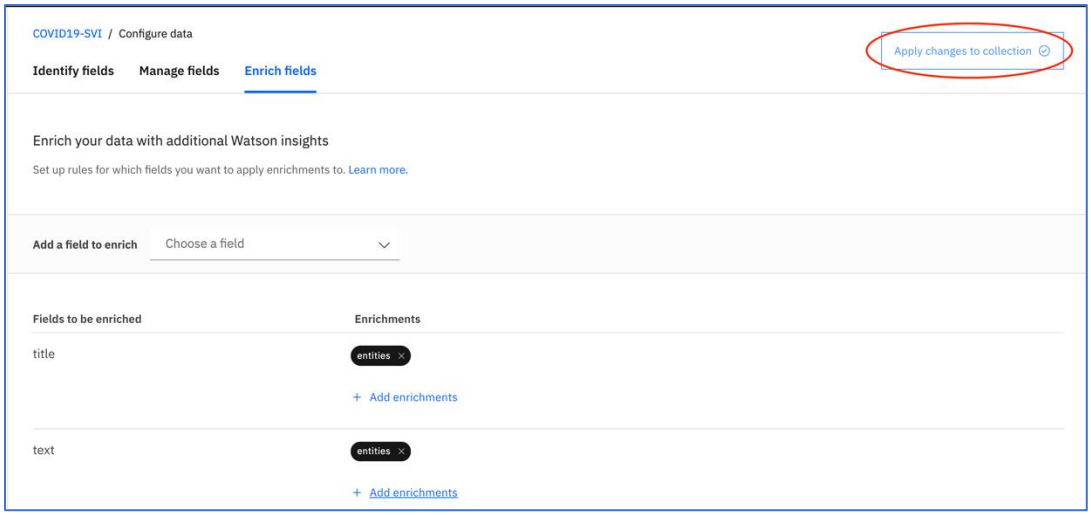


8. Scroll down to the bottom of the Add Enrichments pop-up screen and paste the Model ID number for the ML annotator underneath **Custom Model ID**, click **Apply** and then click **X** to exit this screen.

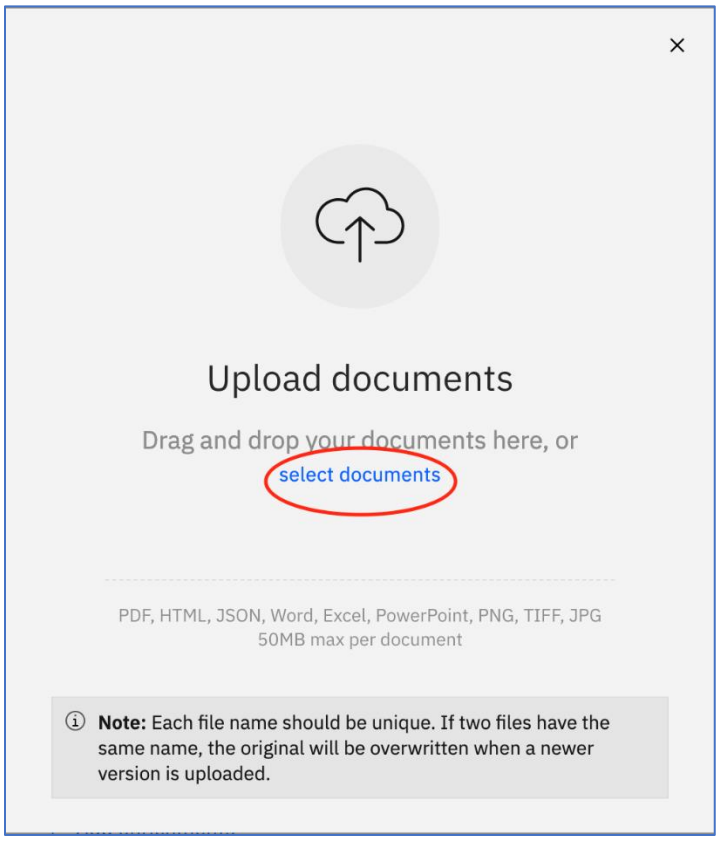


Exercise 4: Perform Custom Entity Extraction

1. Now that we have specified that entity extraction will occur on the title and text fields of each document using our ML annotator, click **Apply changes to collection**.

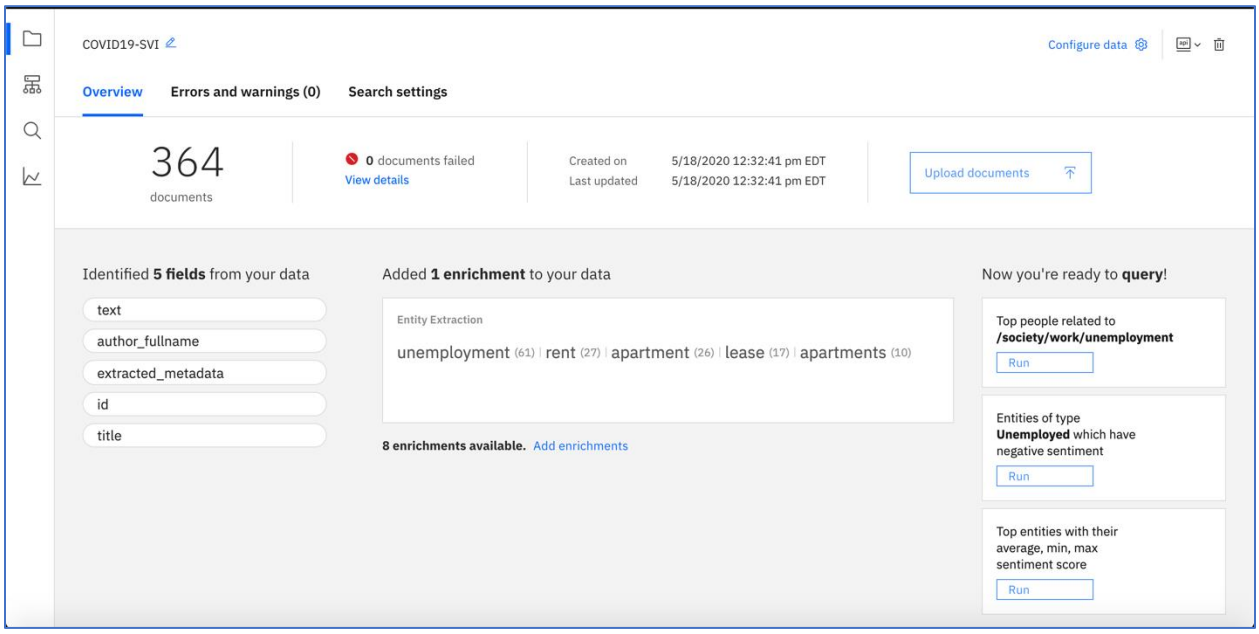


2. We will need to tell Discovery that we are extracting entities from all the documents by selecting our documents again. Click **select documents**.



3. Navigate to the folder where you extracted the zip file downloaded from GitHub. Double-click on the **Posts** folder. **Shift select** all of the documents in the Posts folder and click **Open**.

It will take approximately 6 minutes for entity extraction to occur on all the documents in the collection. When it is complete, you should see the following screen:

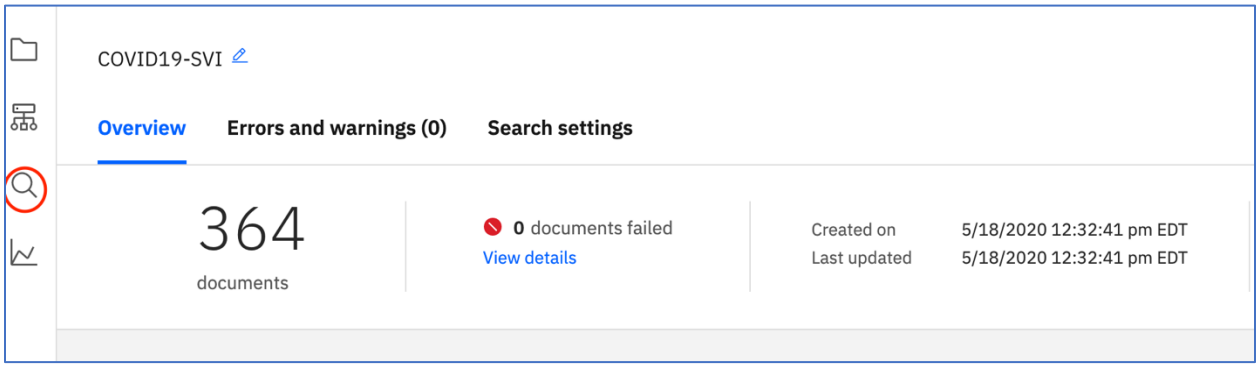


While we are able to see the same number of documents and fields per post, we are now only seeing one enrichment applied to our dataset. Entity extraction has been successfully performed on our dataset and we can now retrieve the output of this process.

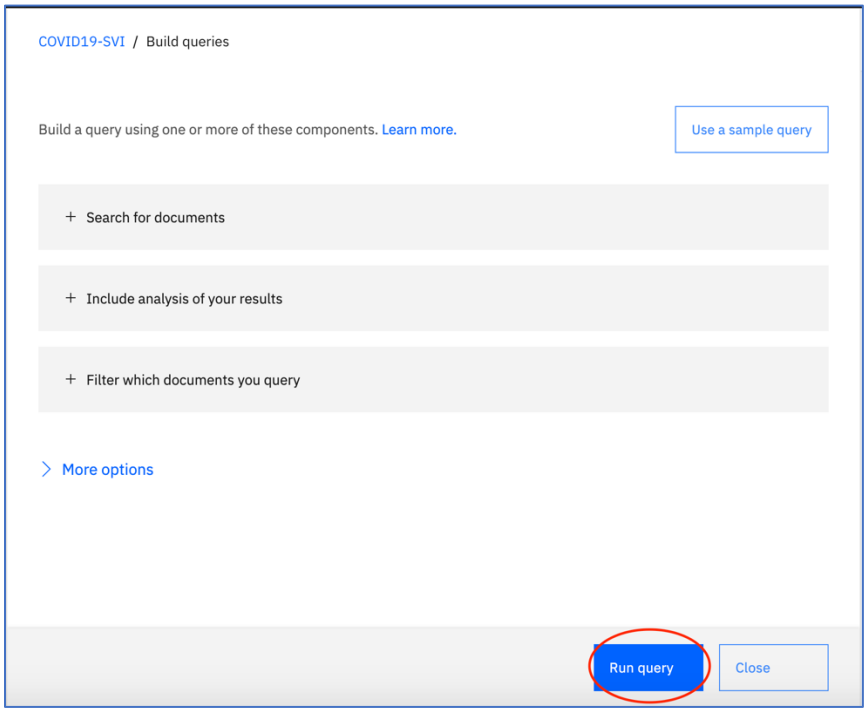
Exercise 5: Calculate the COVID-19 vulnerability index

Let’s start by viewing the entity extraction output inside the Discovery UI.

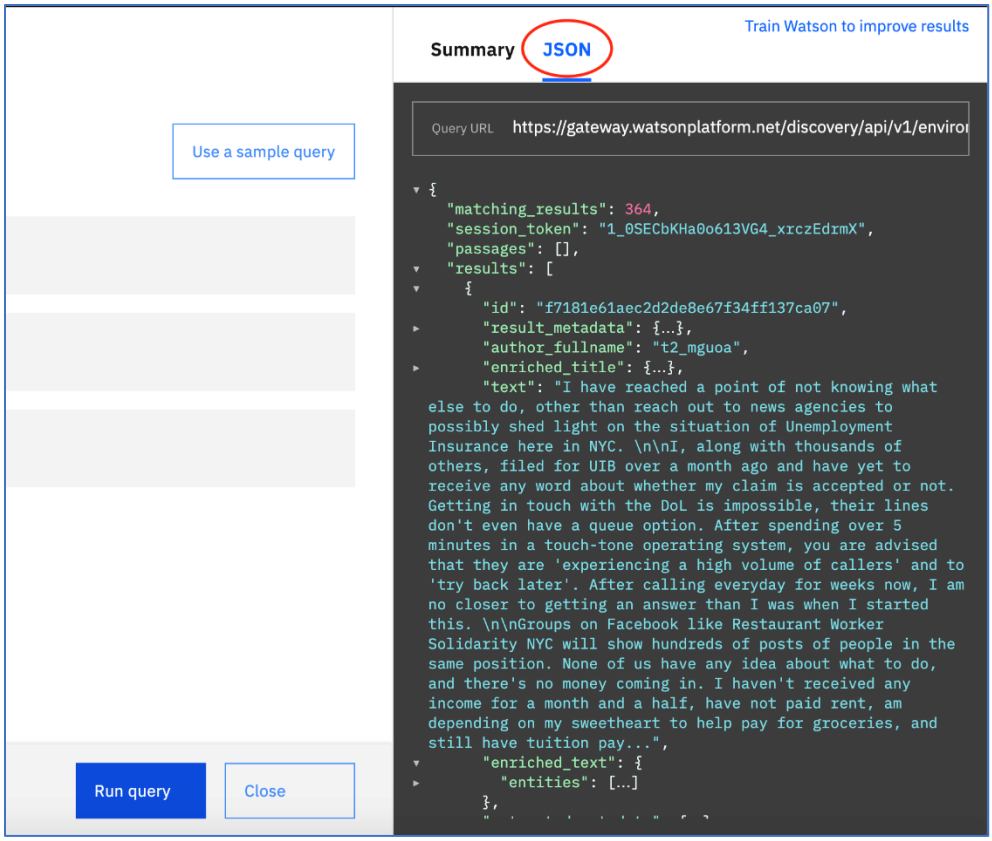
1. Click on the **magnifying glass icon** on the left-hand side of the screen.



2. This will take us to the Build Queries page where we can query using structured Discovery Language Queries or Natural Language Queries. Since we are only interested in viewing the output of the entity extraction process, click **Run Query**.



3. This will return a massive JSON file on the right-hand side of the screen consisting of all the documents in our collection after entity extraction.

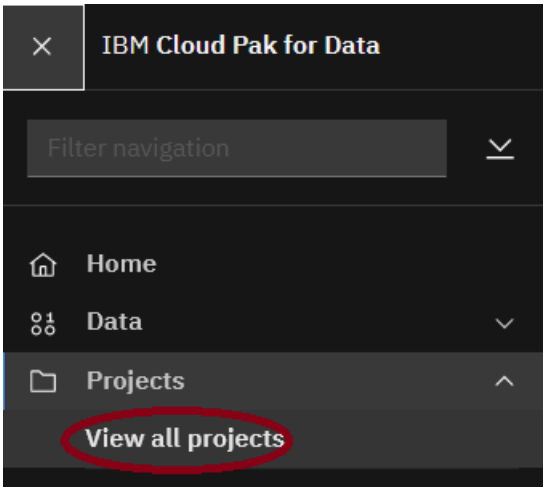


Scrolling down this right panel, we can see that the output consists of all the documents with their original text as well as the extracted entities. We will be using the extracted entities for each document to compute the Social Vulnerability Index (SVI) for COVID-19. To access the extracted entities, we will use the Discovery API in a Jupyter notebook to compute each city SVI. The greater the computed SVI value, the more vulnerable to COVID-19 the city is estimated to be. We will use the Watson Studio service to create and run the Jupyter notebook to calculate the SVI.

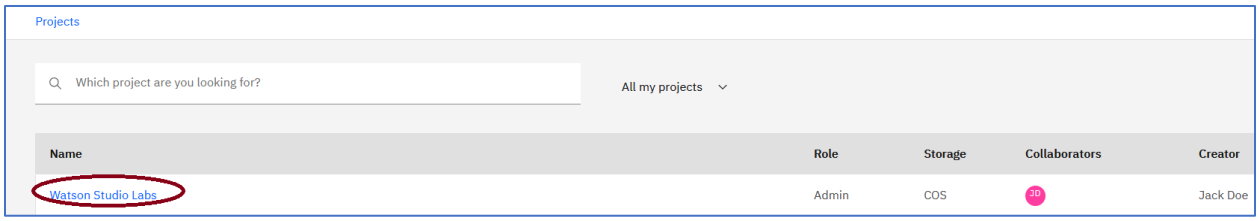
4. Click on the **IBM Cloud Pak for Data** tab to navigate to Watson Studio. If you had closed this tab, you can enter http://dataplatform.cloud.ibm.com in the browser to navigate to Watson Studio.

5. Click on the hamburger icon

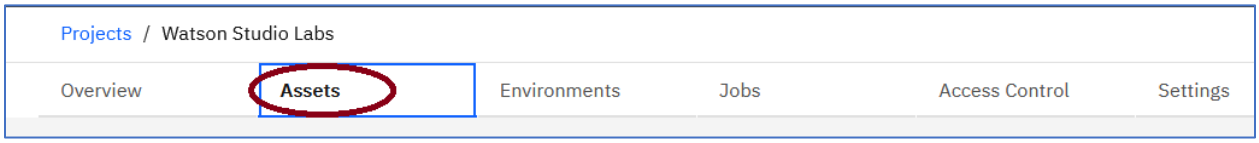
6. Click on **View all projects**.



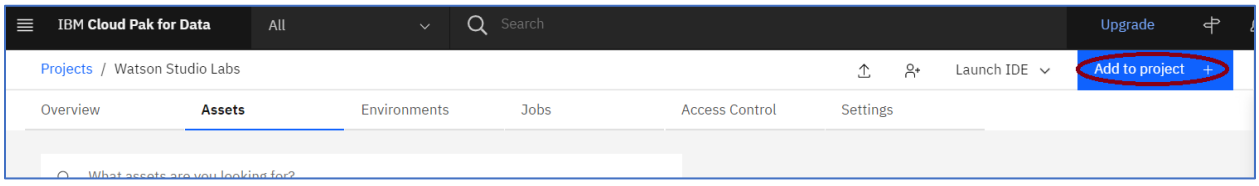
7. Click on **Watson Studio Labs**.



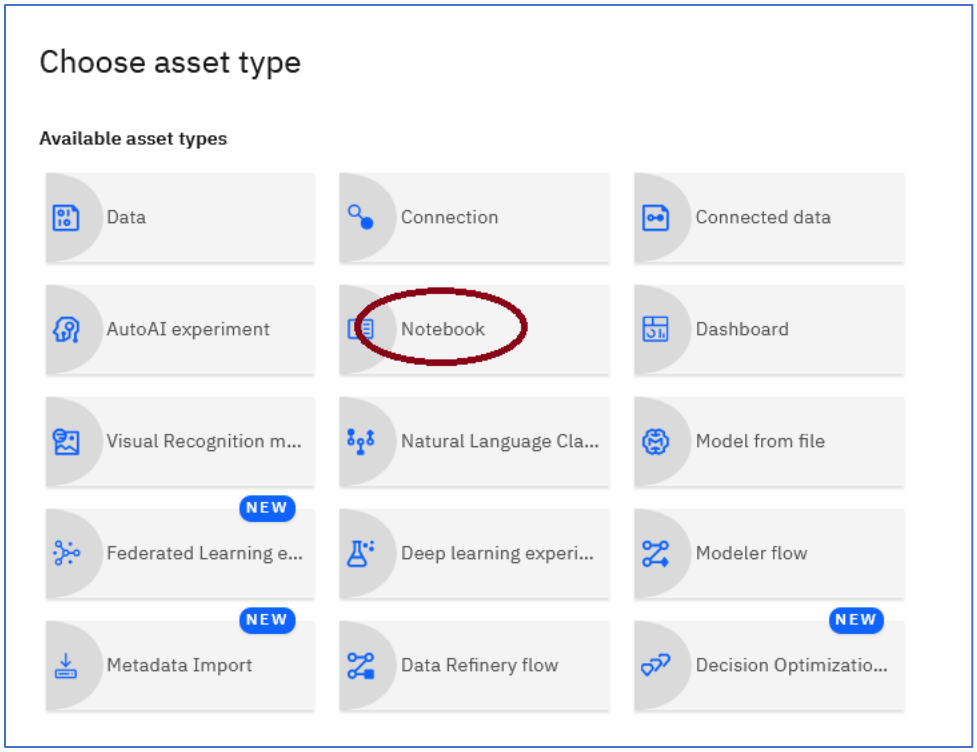
8. Click on the **Assets** tab.



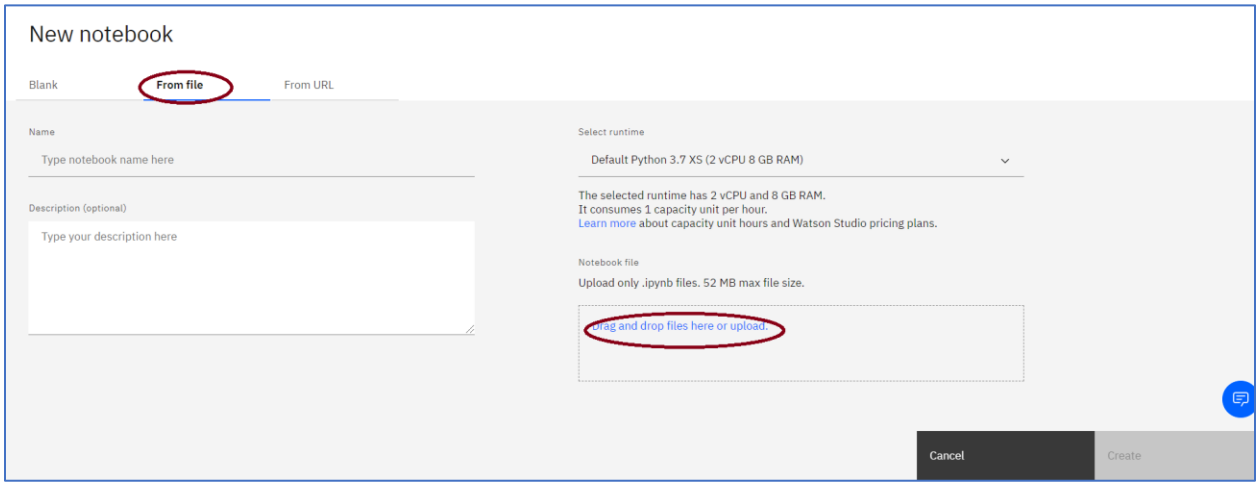
9. Click on **Add to project**.



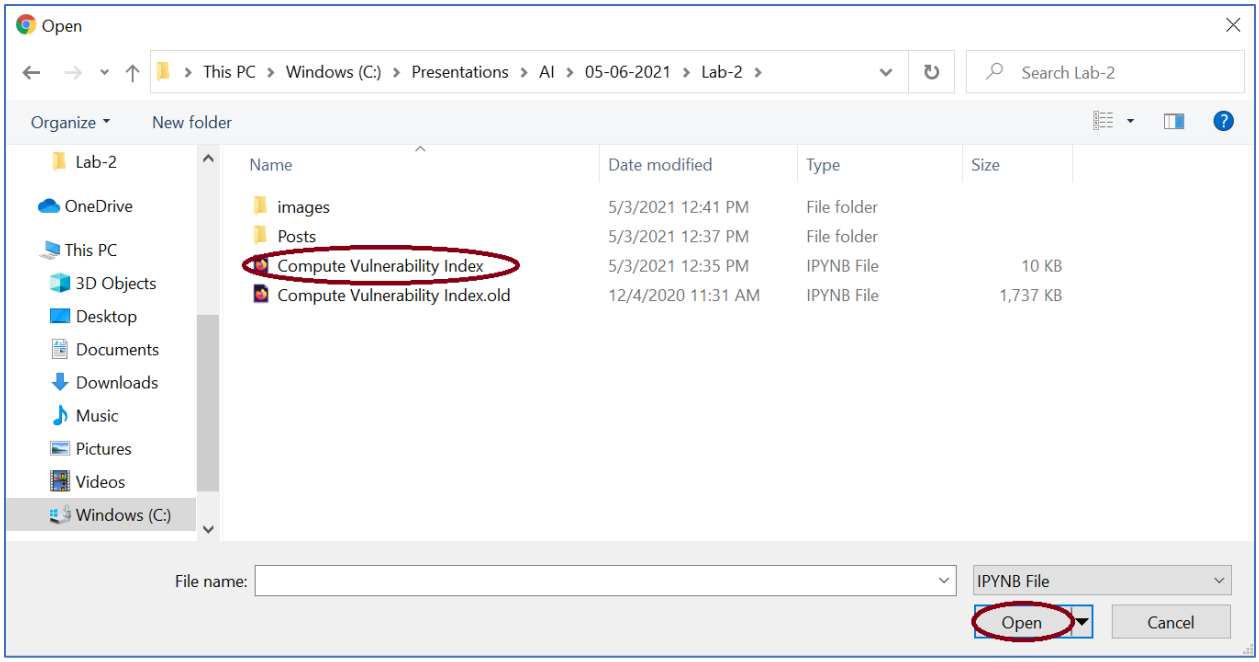
10. Click **Notebook**



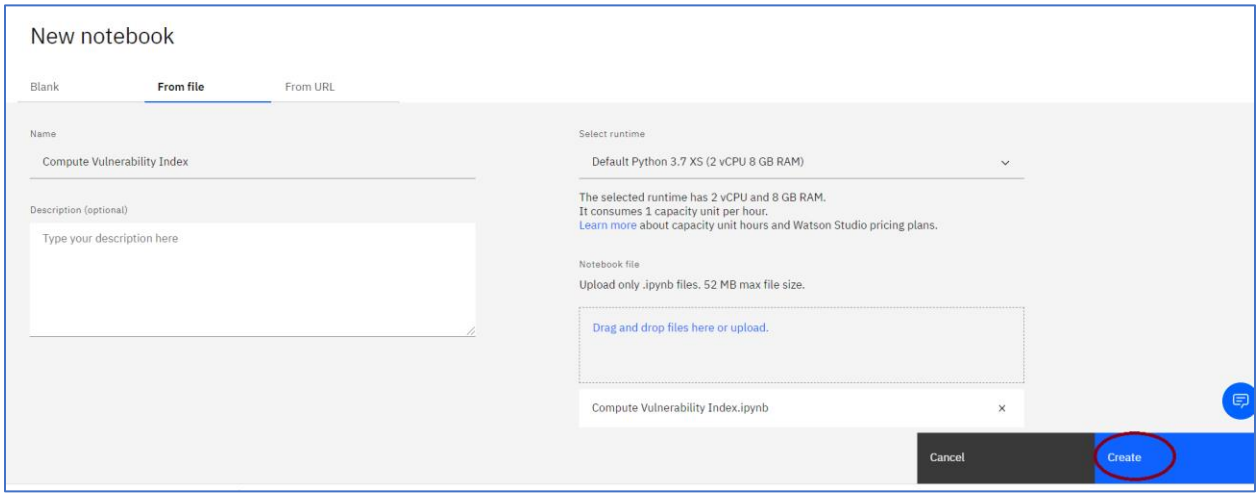
11. Click on the **File** tab, and then click on **Drag or drop files here or upload.**



12. Navigate to the folder where you extracted the zip file downloaded from GitHub. Click on the file **Compute Vulnerability Index** and then click **Open**.



13. Click **Create**.

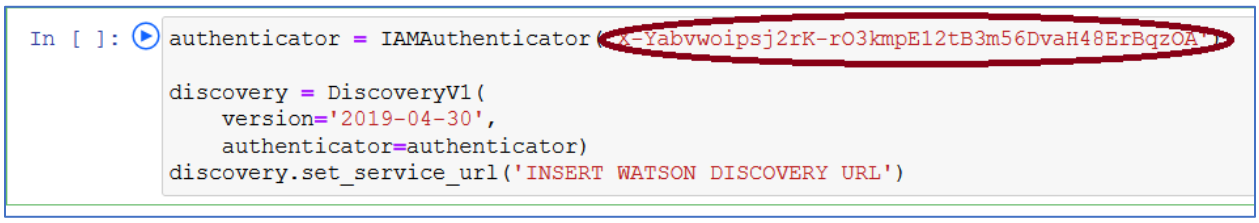


14. Before executing the notebook, please replace the INSERT WATSON DISCOVERY API KEY in the Notebook (see below) with the Watson Discovery API key that you copied earlier. Copy the Discovery API key from the text editor, then highlight the INSERT WATSON DISCOVERY API KEY as shown below

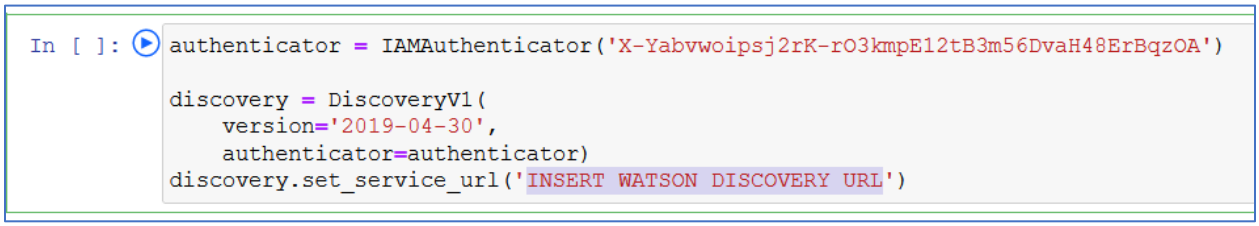
```
In [ ]: authenticator = IAMAuthenticator('INSERT WATSON DISCOVERY API KEY')

        discovery = DiscoveryV1(
            version='2019-04-30',
            authenticator=authenticator)
        discovery.set_service_url('INSERT WATSON DISCOVERY URL')
```

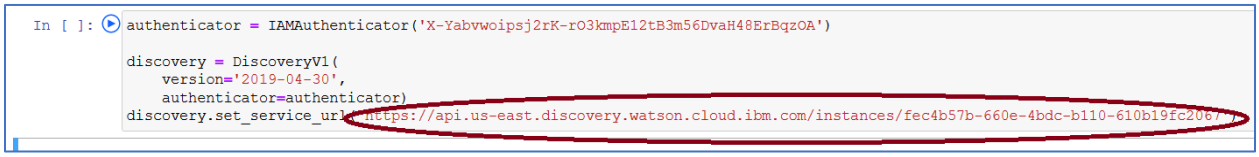
15. Then paste the API key. Note, your API key will be different.



16. Similarly, please replace the INSERT WATSON DISCOVERY URL in the Notebook (see below) with the Watson Discovery URL that you copied earlier. Copy the Watson Discovery URL from the text editor, then highlight the INSERT WATSON DISCOVERY URL as shown below



17. Then paste the URL. Note, the url will depend on the region.



The next step will be to execute the cells in the notebook. For those not familiar with Jupyter notebooks, a Jupyter notebook consists of a series of cells. These cells are of 2 types (1) documentation cells containing markdown, and (2) code cells (denoted by a bracket on the left of the cell) where you write Python code, R, or Scala code depending on the type of notebook. Code cells can be run by putting the cursor in the code cell and pressing **<Shift><Enter>** on the keyboard. Alternatively, you can execute the cells by clicking on the **Run icon** on the menu bar that will run the current cell (where the cursor is located) and then select the cell below. In this way, repeatedly clicking on **Run** executes all the cells in the notebook. When a code cell is executed the brackets on the left change to an asterisk '*' to indicate the code cell is executing. When completed, a sequence number appears. The output, if any, is displayed below the code cell.

18. Execute each of the notebook cells in order (either by typing in **<Shift><Enter>** or using the **Run** menu option). Read the notebook comments to gain an understanding of the code that is executing. **When all the cells in the notebook have been successfully executed, please return to this document, and continue with the below steps.**

The SVI calculation is notional. We have summed the count of entity mentions into different categories and provided a weighting factor for those categories to calculate a notional SVI. As we can see, since NY-

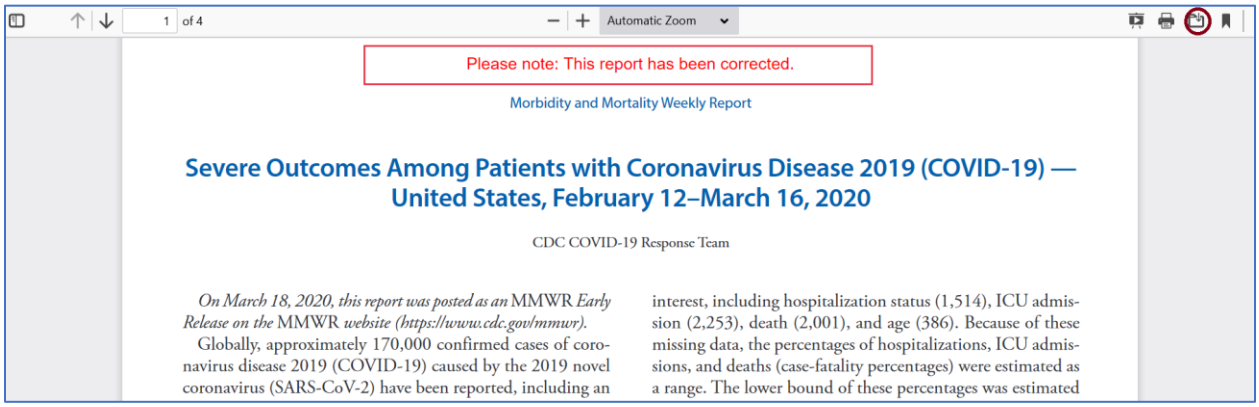
Discovery has the largest SVI, New York City was calculated to be the city most socially vulnerable to COVID-19.

Exercise 7: Create a collection for a COVID-19 publication

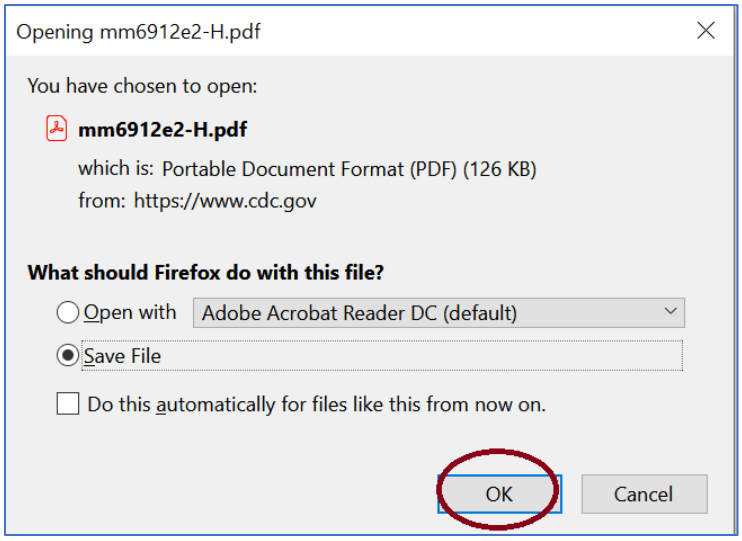
Although we have just used Watson Discovery to ingest a collection of social media data, perform custom entity extraction and use the analyzed files to compute a vulnerability index for different U.S. cities, we have not shown you how to search through a journal publication to answer natural language questions and retrieve relevant passages…yet. In this exercise, we're going to first create a brand new collection for a COVID-19 related journal publication.

1. Let's find our COVID-19 journal article by visiting https://www.cdc.gov/mmwr/volumes/69/wr/pdfs/mm6912e2-H.pdf

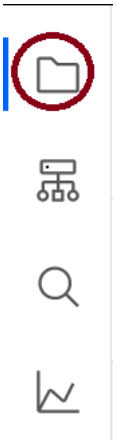
2. Download the file to your workstation. Open up the downloaded pdf file to verify that you have the complete article, which spans 4 pages. We will be uploading this article to a brand new collection in Watson Discovery.

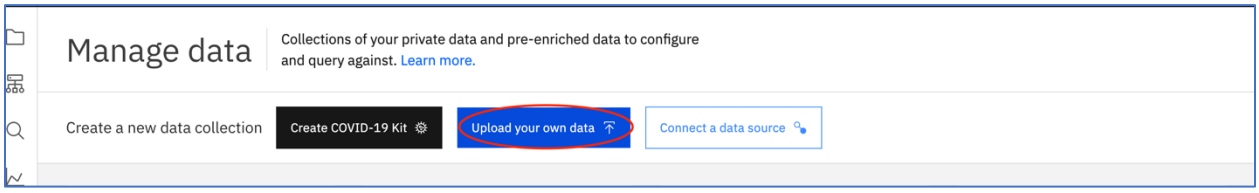


3. Click OK.

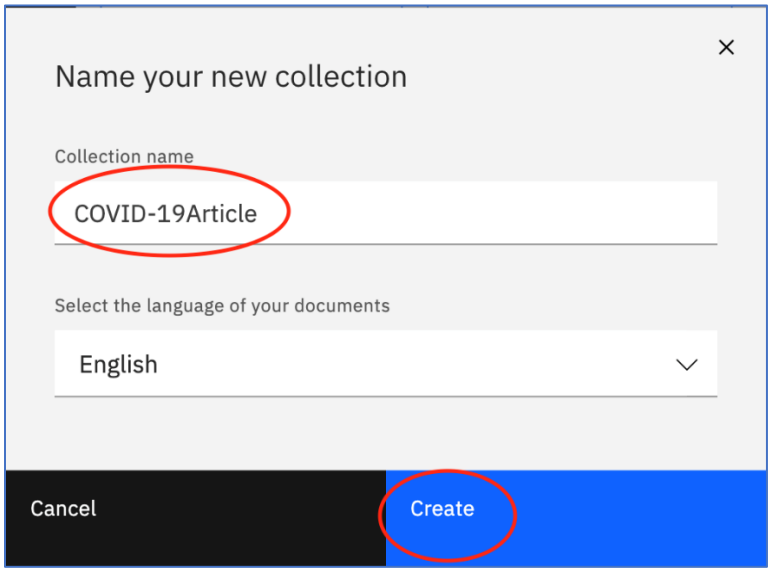


4. Open up your Discovery instance (you may have to log in again) and click **Upload your own data**. If you are not at the right panel, click on the Document icon.

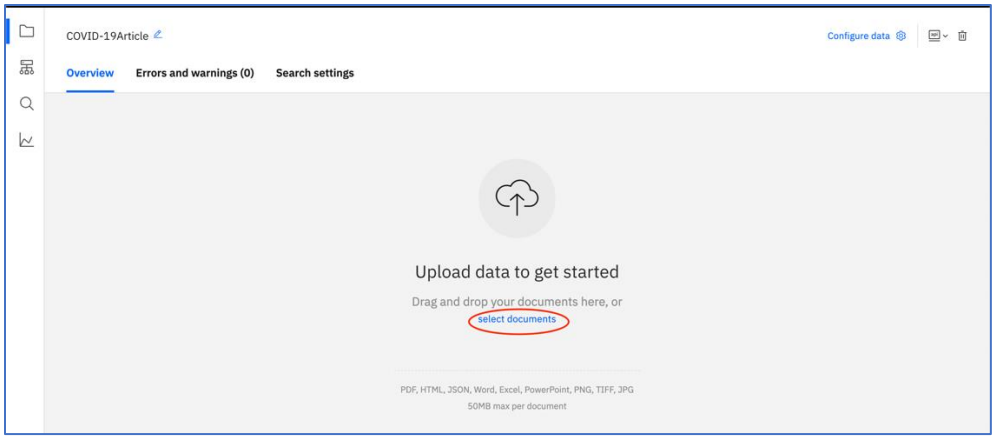




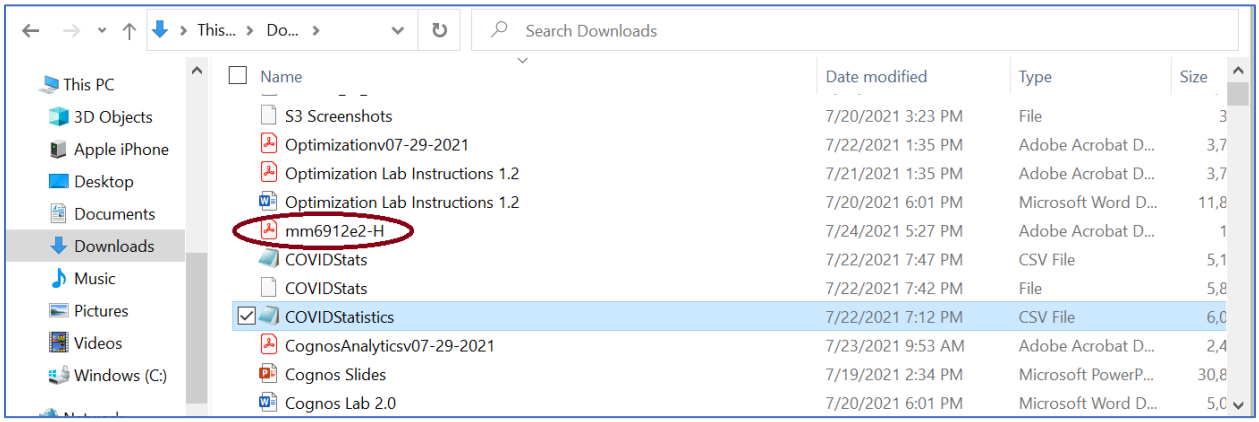
5. Give the new collection a name of **COVID-19 Article** and click **Create**.



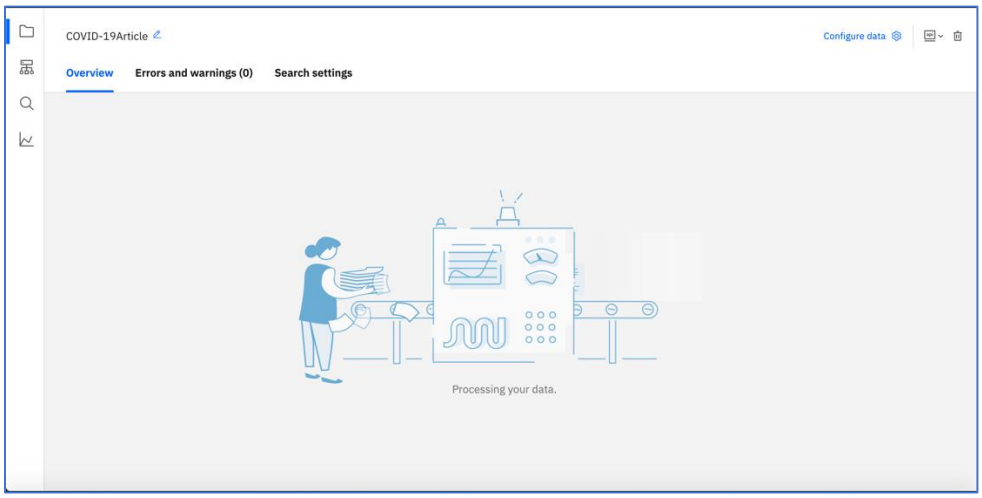
6. Inside the new collection, click **select documents.**



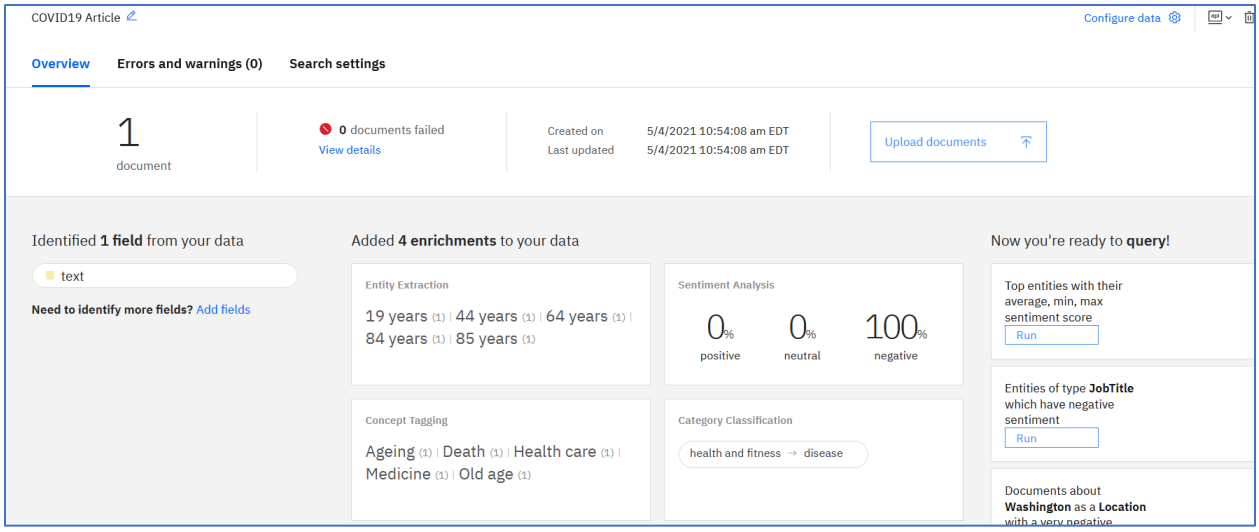
7. Select the pdf article inside your Download folder and click **Open**.



8. Wait approximately 13 minutes for the document to be completely uploaded to the new collection.



9. After the document is successfully uploaded, you should see the following screen:

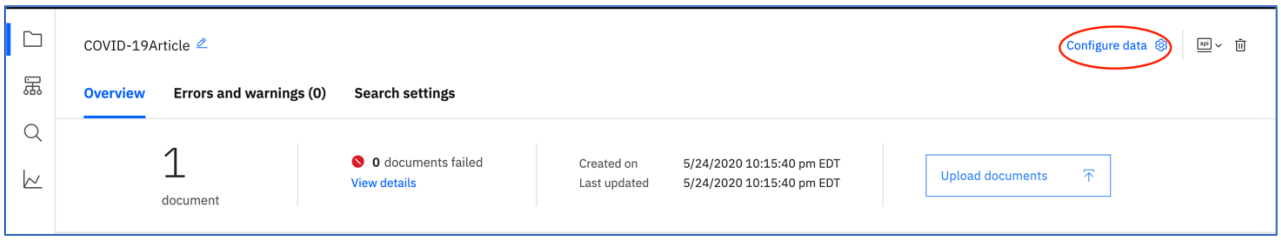


The default NLP enrichments of entity extraction, concept tagging, sentiment analysis and category classification have already been applied to this journal article. We will not be making any changes to these enrichments and we will instead teach Watson to understand the underlying structure of the article in the next section.

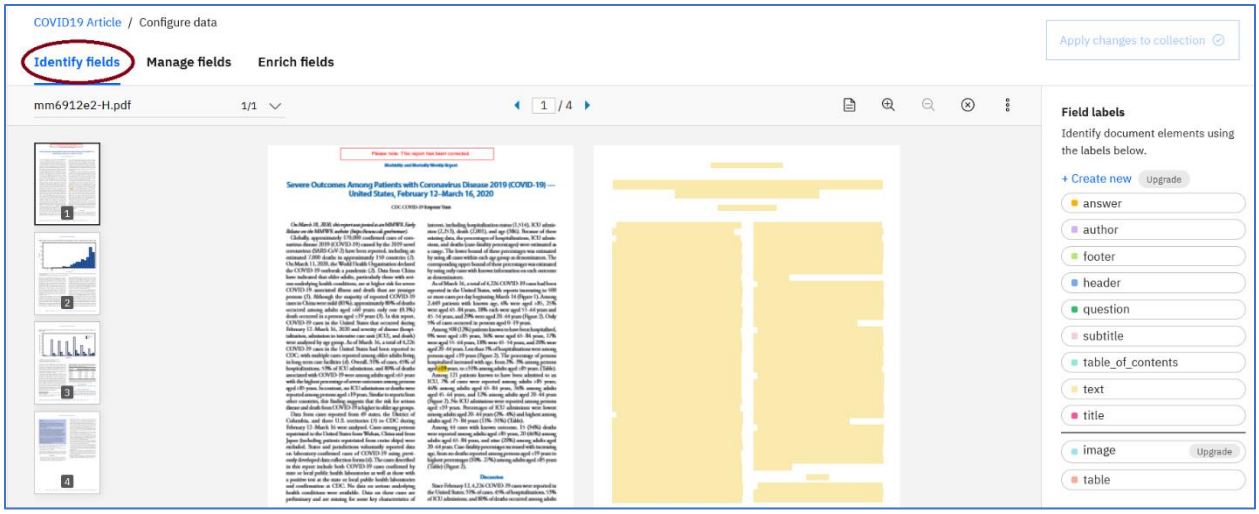
Exercise 8: Perform Smart Document Understanding

Before we can search through the pages of our document for answers to specific questions, we must first train Watson to understand the underlying structure and format of our entire document. This is made possible through a capability known as Smart Document Understanding (SDU). We can access SDU within our Discovery instance.

1. Click **Configure Data**.

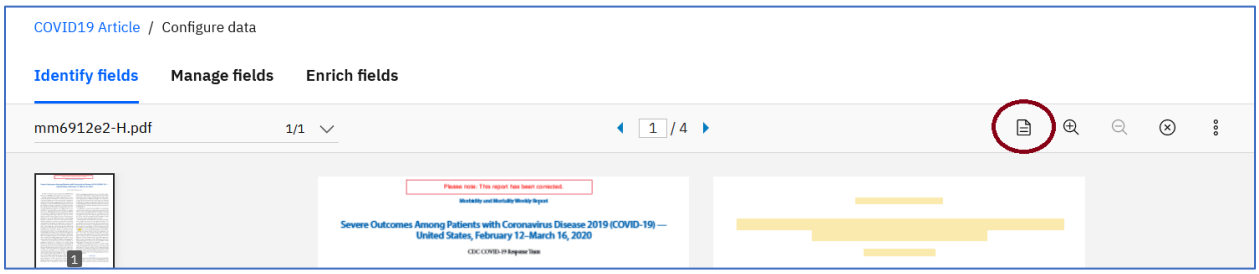


2. On the **Identify fields** page, you should be able to see a page by page preview of the uploaded article. This is where we can access SDU.

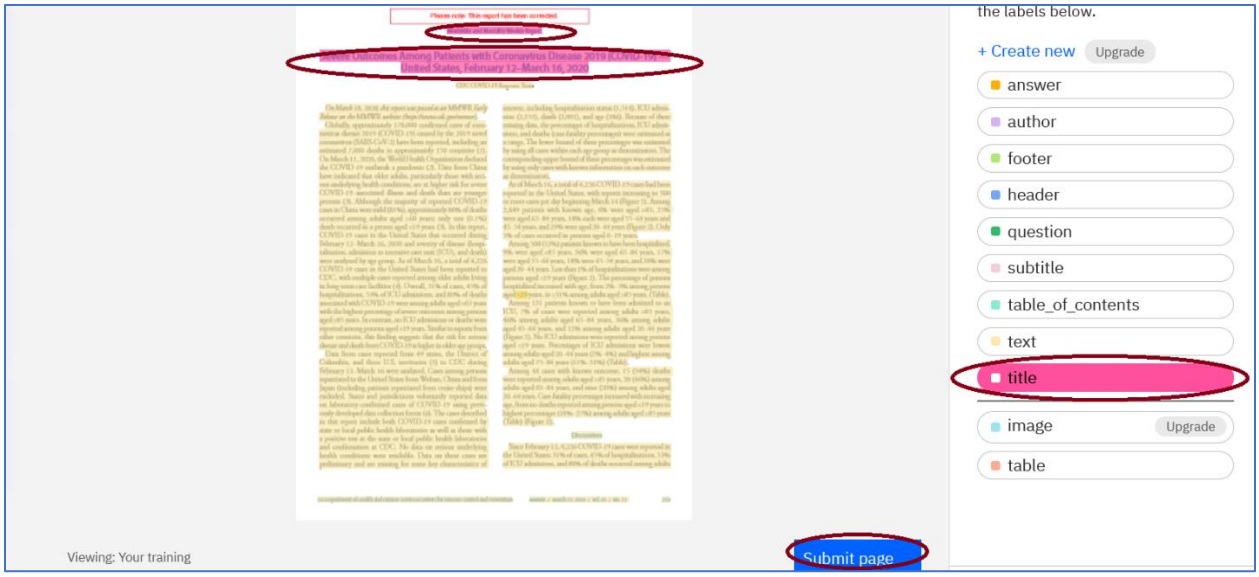


We are now going to use SDU on each page of our document and mark each section with the appropriate field on the right hand side in order to train Watson to understand the structure and format.

3. Click on the **small book icon** near the center of the screen to revert to a single page view.



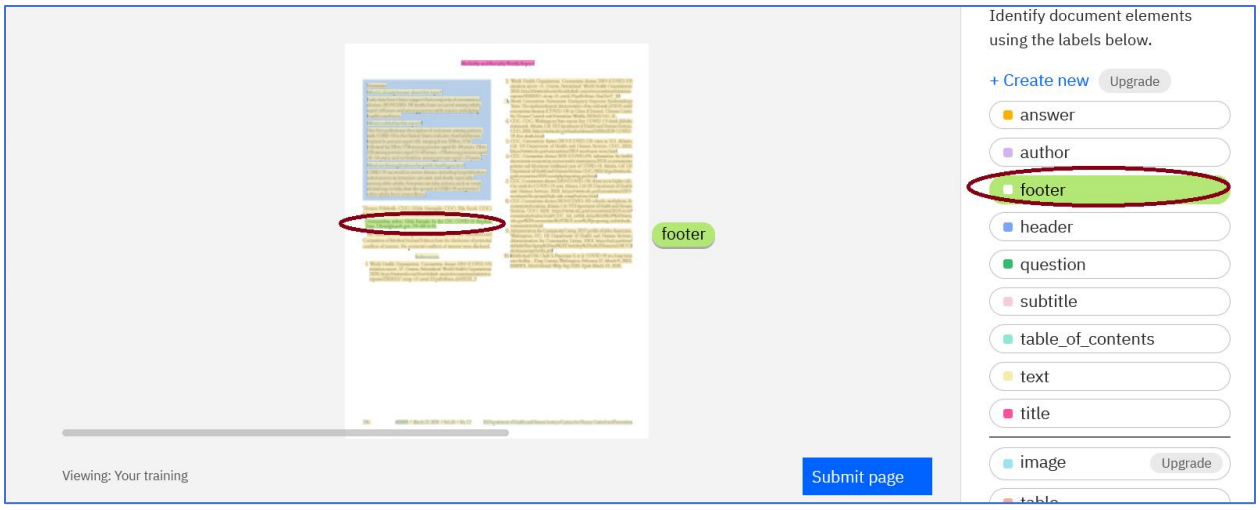
4. On the first page of the document, we are going to click on title underneath Field labels on the right side of the screen and then use it to highlight “Morbidity and Mortality Weekly Report” and “Severe Outcomes Among Patients with Coronavirus Disease….” Use the zoom in icon as needed. To highlight a phrase with a field label, simply select the field label first and then click and drag it over the desired phrase. You should be able to see the following labeled page. Click **Submit** to advance to the next page.



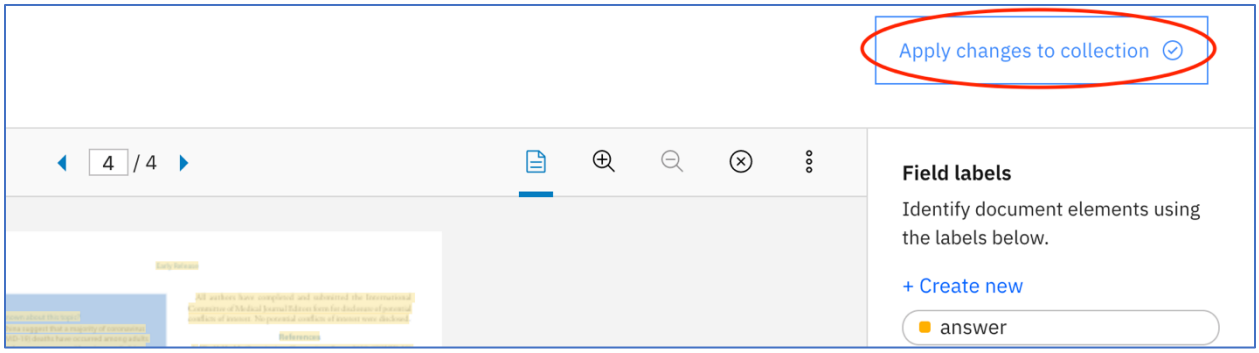
5. Label the text on the second page of the document with the text field label and click Submit page.

6. Label the third page using the text and footer fields. Click the **Submit page** button to move on to the last page.

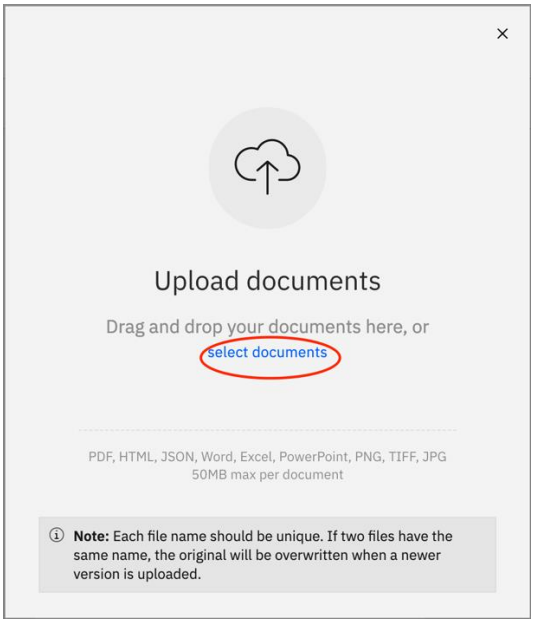
7. Label the fourth and last page so that it resembles the following labeled page. Click **Submit Page**.



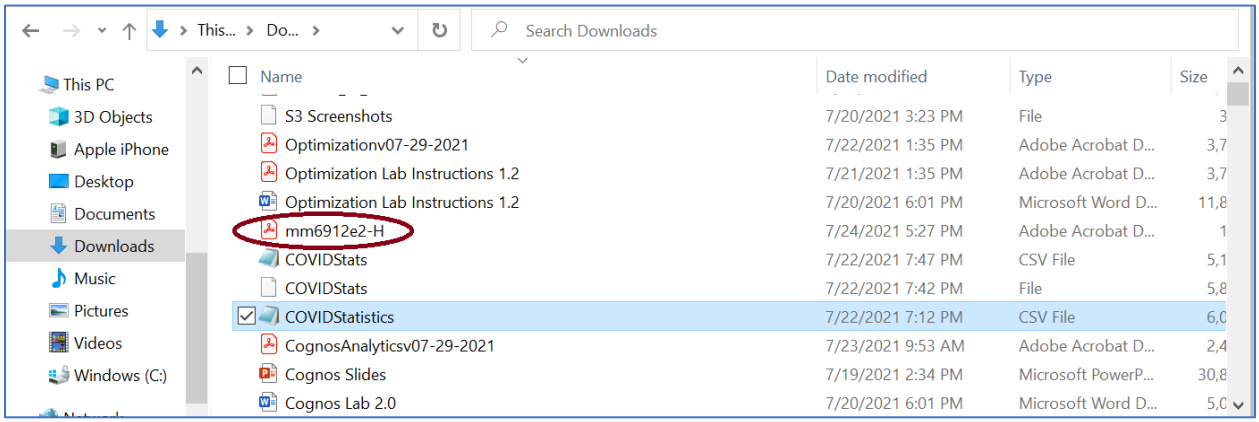
8. Click the **Apply changes to collection** button on the top right corner of the screen to save the labelled pages.



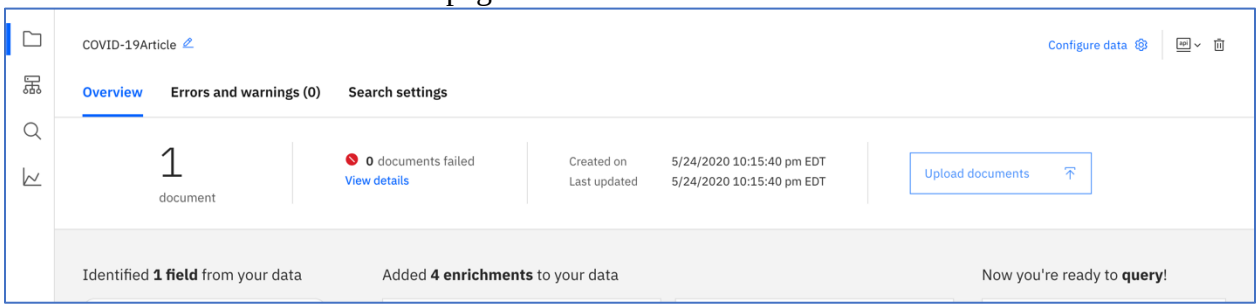
9. You will be asked to select your document again in the Upload documents screen. Click **select documents**.



10. Select the pdf article in the Download folder and click **Open**.



11. In a few seconds, the changes will be applied to the collection and you will be taken back to the Collection overview page.

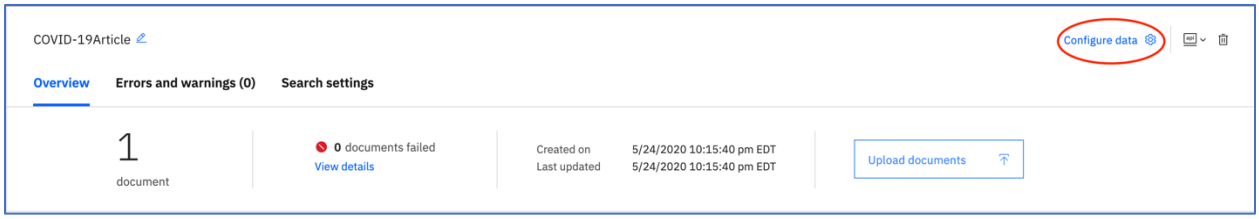


We have successfully used SDU to train Watson on the structure and format of each page of our PDF document. This will enable us to create queries to search through our document in the next exercise and retrieve relevant passages in order to answer natural language questions.

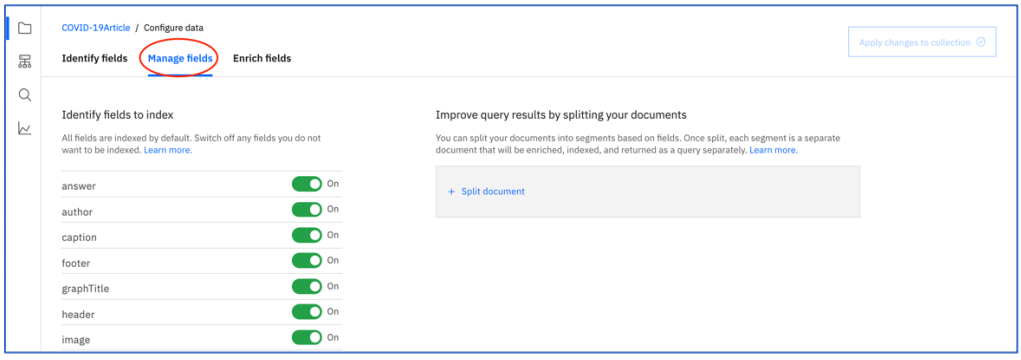
A final note about SDU: when we have multiple documents in a collection to train, we can label one document and save its corresponding SDU model to use on other similarly-formatted documents.

Before we start creating queries, let's make one change to the configuration of the document.

12. Click on **Configure data**.

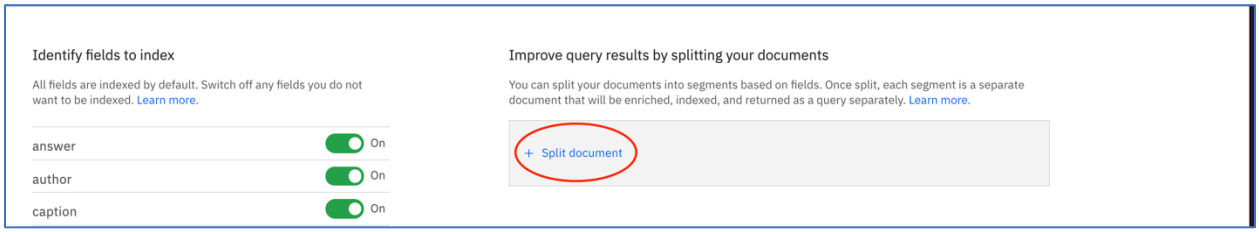


13. Click on **Manage fields.**

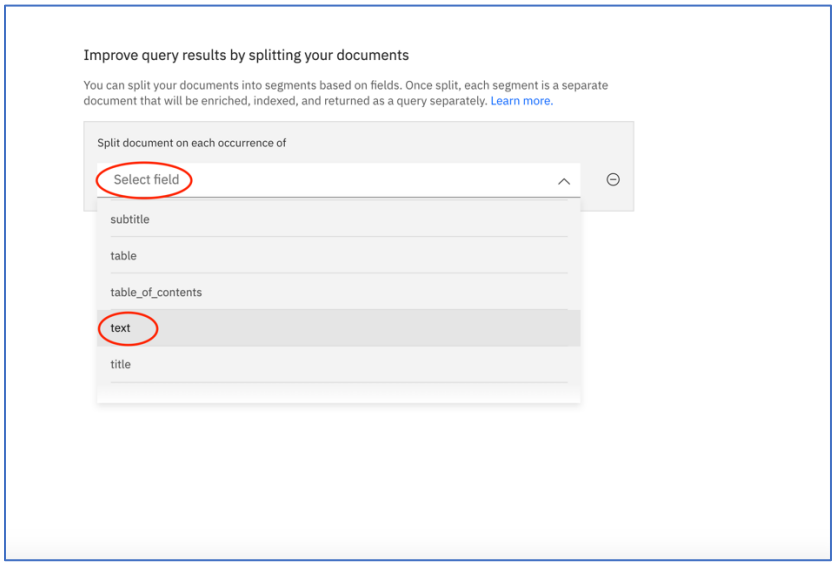


14. In order to improve the results of the queries that we will be creating in the next exercise, we can split the journal article on each occurrence of a specific field. In our case, we will split the article on each occurrence of the text field, which approximately represents each

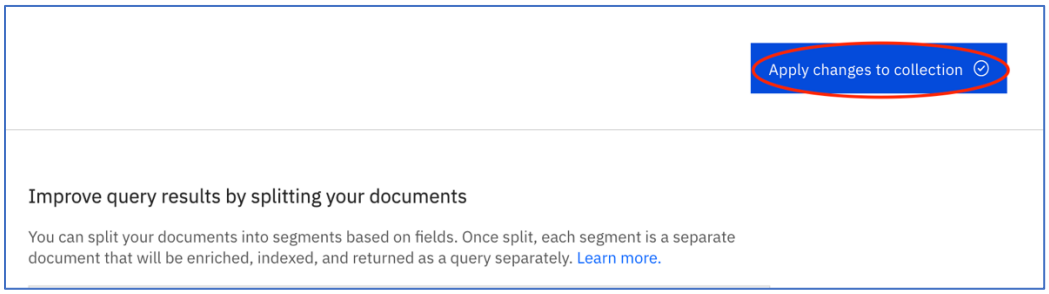
paragraph in the document. Once the split is complete, each paragraph in the article will now be a separate document that will be enriched, indexed and returned as a query result. Click **+ Split document**.



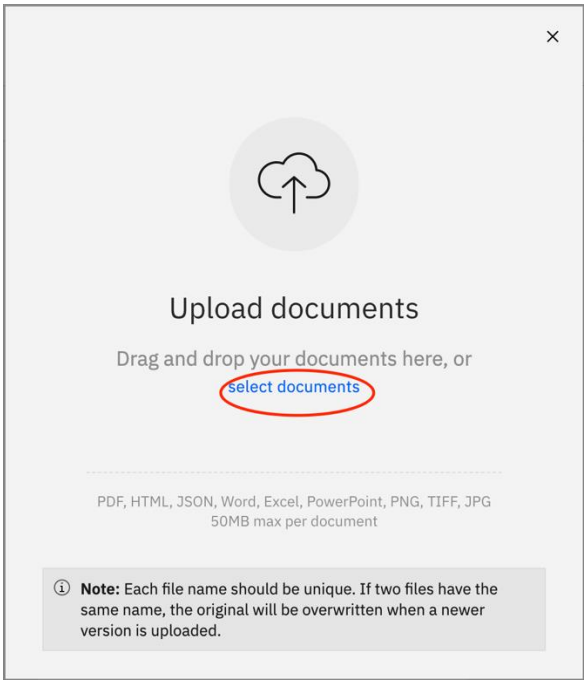
15. Click **Select field** and choose **text**.



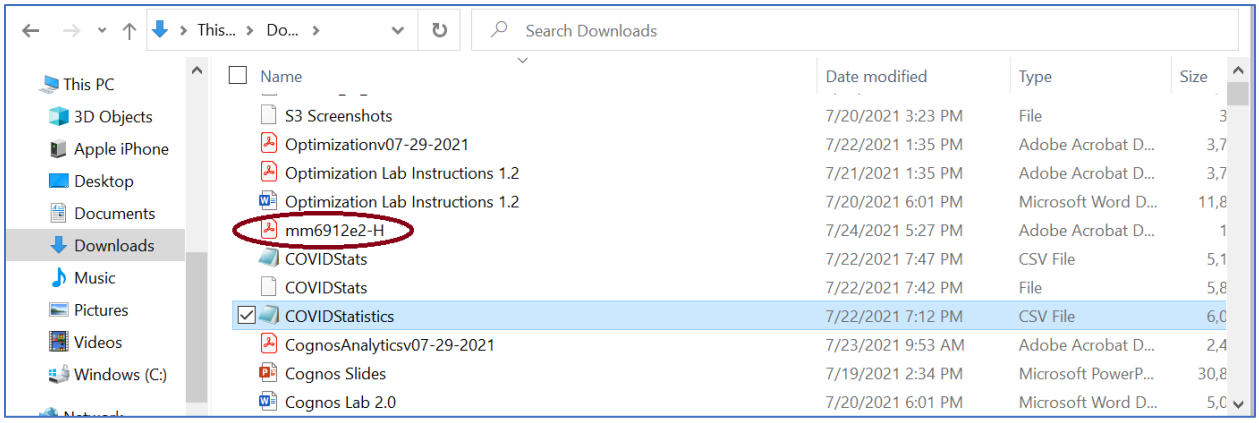
16. Click **Apply changes to collection** for the journal article to be split on each occurrence of the text field.



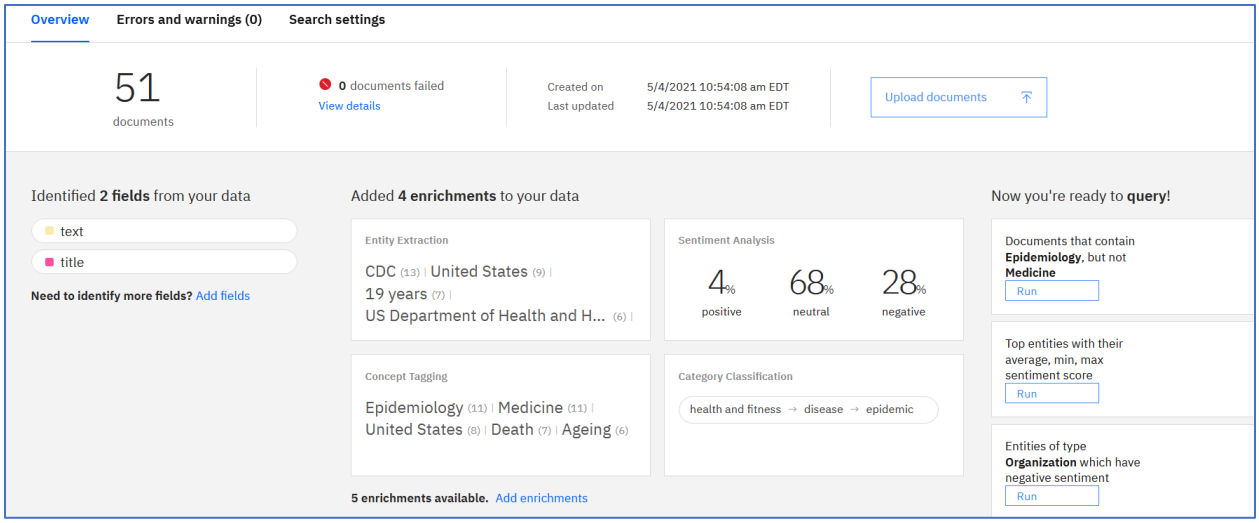
17. You will be prompted to select the journal article. Click **select documents**.



18. Select the journal article and click **Open**.



19. You will be taken back to the overview page where you'll see that the original document has been split into 50 or 51 documents. Do not proceed to the next exercise until you see the additional documents (wait approximately a minute and refresh the page if needed).

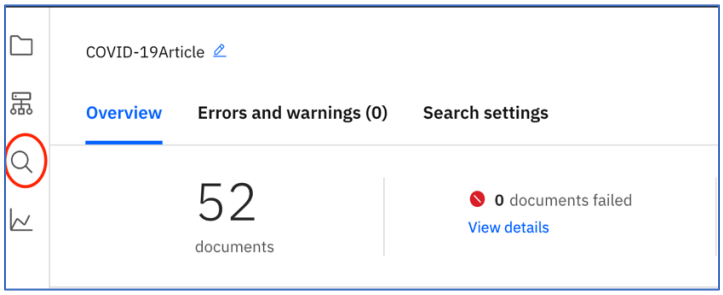


Exercise 9: Create and run Natural Language Queries

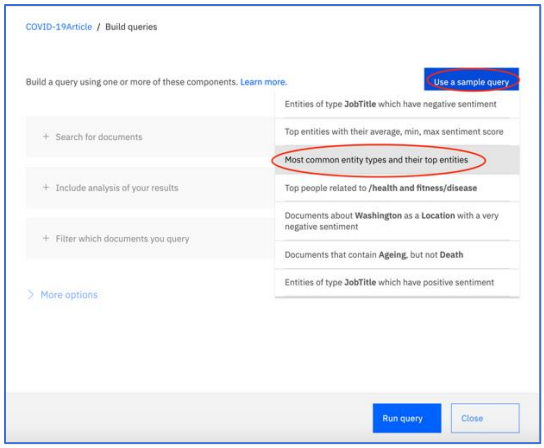
There are two types of queries that we can create inside of Watson Discovery to search through documents – structured queries and natural language queries. Let's start by running a few structured queries on our uploaded document.

Given the NLP enrichments that were applied to our document by default – entity extraction, sentiment analysis, concept tagging and category classification – we can use any combination of these enrichments to produce structured queries. Let's start by creating a structured query using the entity enrichment.

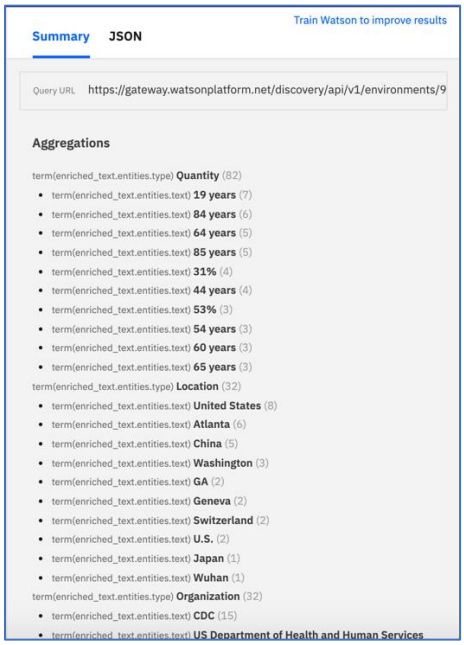
1. Let's navigate to the Query page by clicking on the **magnifying glass icon** on the left-hand side of the screen.



2. We will be running a sample query to retrieve the most common entity types identified in our document and their top entities. Click on **Use a sample query** and select **Most common entity types and their top entities.**



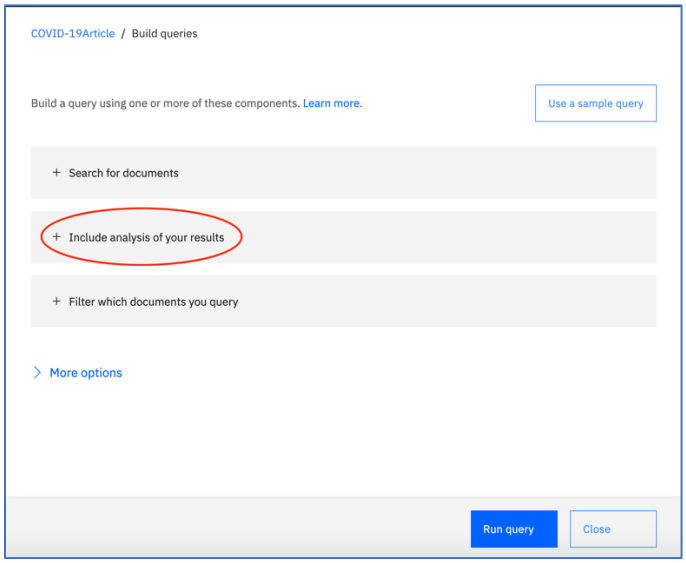
3. You should immediately see the results of this query on the right-hand side of the screen:



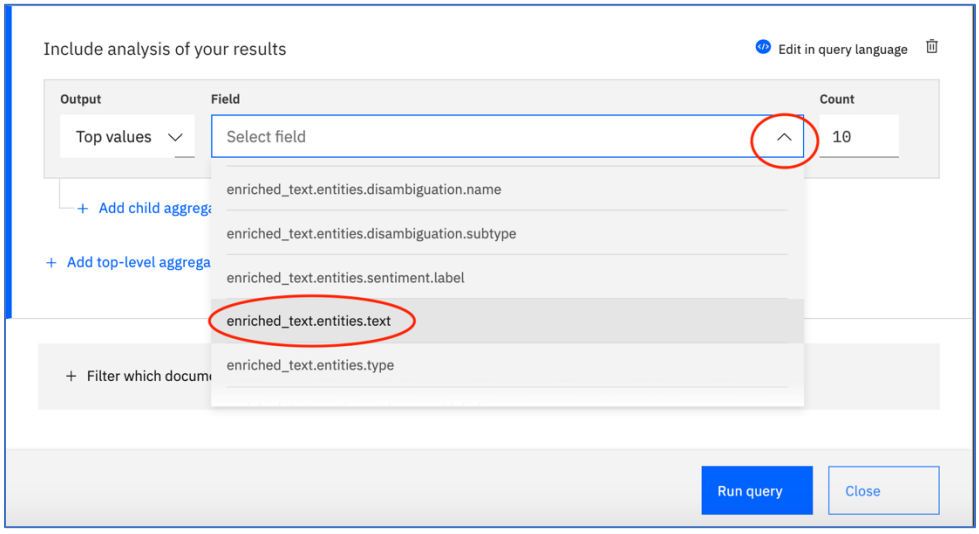
As you can see, the most entity types in our PDF article are Quantity, Location, Organization, Person and EmailAddress and we can also see the top entities pertaining to each entity type.

4. Now instead of using a sample query, let's build our own structured query to determine the top entities with their average sentiment score.
   Click the **trash icon** next to Include analysis of your results

5. Click on **+ Include analysis of your results**

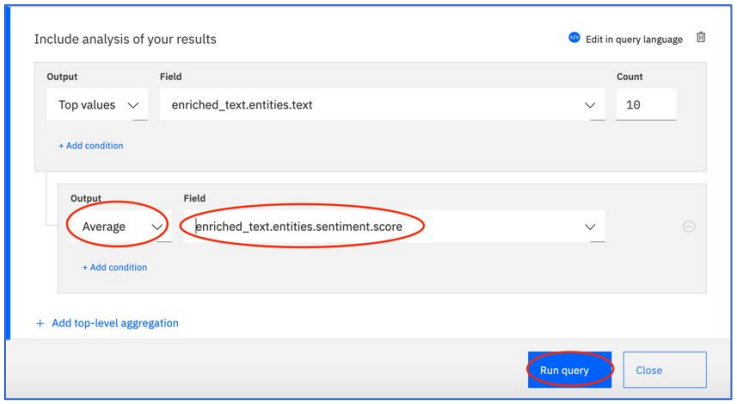


6. From the **Field** drop-down, select **enriched_text.entities.text**



7. Click **+ Add child aggregation**

8. Under **Output**, select **Average** and under **Field**, select **enriched_text.entities.sentiment.score.** Click **Run query** to view the results.

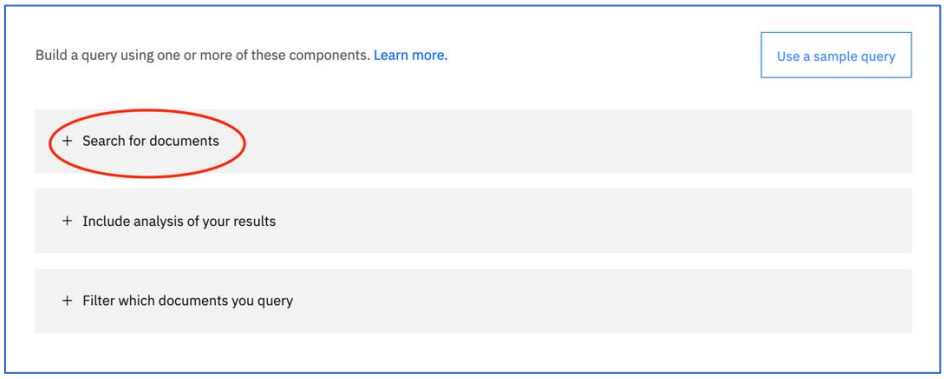


9. You should now be able to see the top entities in the document with their associated average sentiment scores.

While it is certainly useful to run queries to learn more about the NLP enrichments that were identified in the document, we haven't actually performed a detailed search through the content of the document. Moreover, it would be convenient if we could ask colloquial questions about our document without using the Discovery Query Language. Fortunately, we are able to do this with Natural Language Queries.

10. Click on the trash icon next to Include analysis of your results to clear the previous query.

11. Click on **+ Search for documents**.



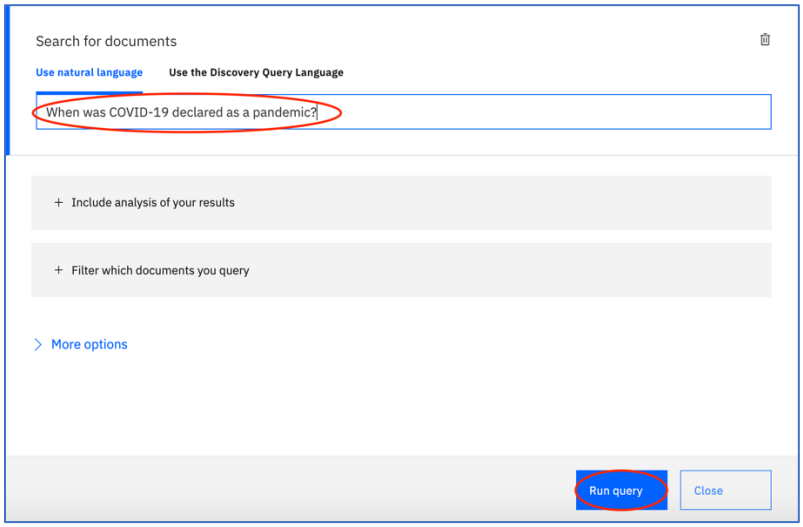
Here we can create natural language queries to ask specific questions about our document. After a quick glance at our article, we can ask the following questions:

- When was COVID-19 declared as a pandemic?
- Which age group in the United States has suffered the highest percentage of severe outcomes?
- How can we protect older adults from COVID-19?
- What was the percentage of fatalities among people that are less than 19 years old?

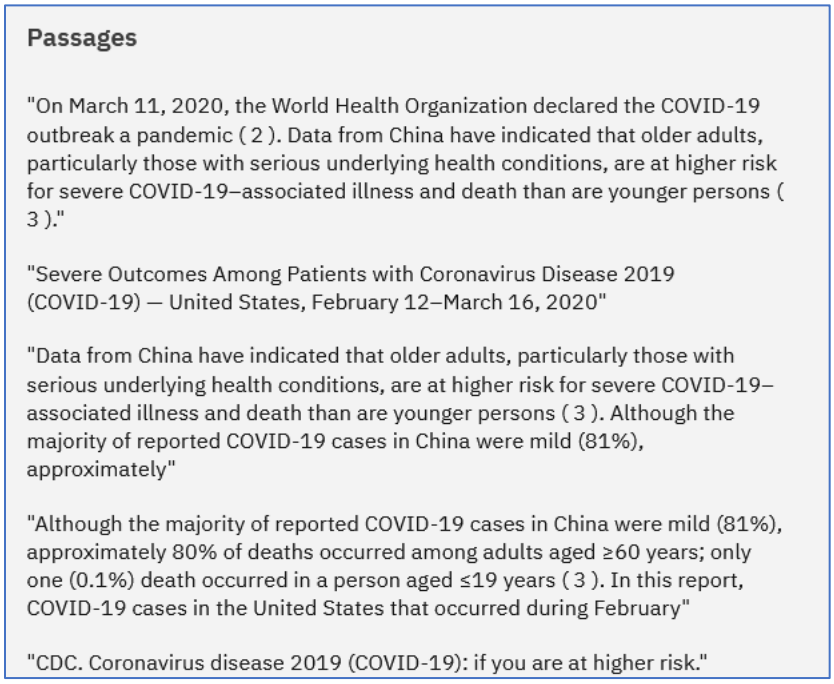
Let's create natural language queries with each of these questions.

12. Creating a new natural language query is as simple as typing in a question you would like to ask about the document. As soon as the natural language query is created, Watson uses this query to retrieve relevant passages from the document in an attempt to answer the question. Since this Discovery collection has applied NLP enrichments to only the text field of the document, the passages retrieved to answer each query will originate from the body text of the document (and not from the title or footer fields).

Underneath **Use natural language**, type in **When was COVID-19 declared as a pandemic?** Then click the **Run query** button.



13. You should be able to see 5 passages on the right side of the screen that were retrieved to answer this question.



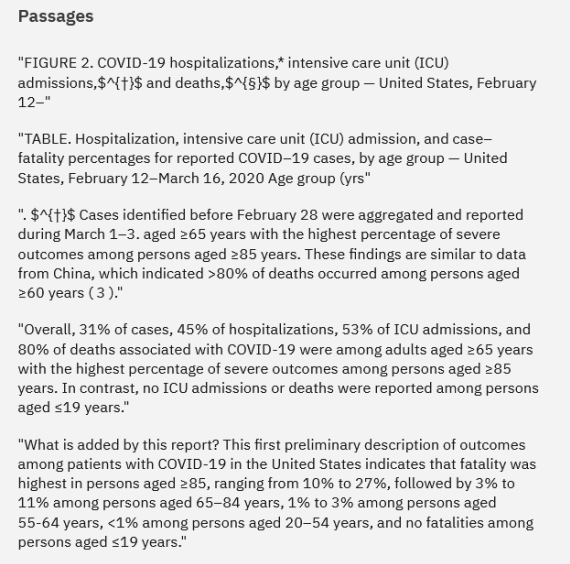
14. Repeat steps 12 and 13 to generate and run the following natural language queries:

    Which age group in the United States has suffered the highest percentage of severe outcomes?
    How can we protect older adults from COVID-19?

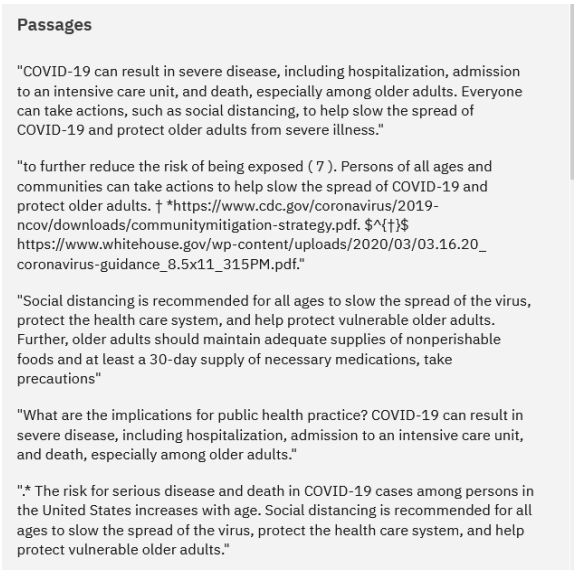
What was the percentage of fatalities among people that are less than 19 years old?

15. You should see the following passages returned for the queries:

Which group had highest % of severe outcomes?



How can we protect older adults from COVID19?



What is % among people less than 19?

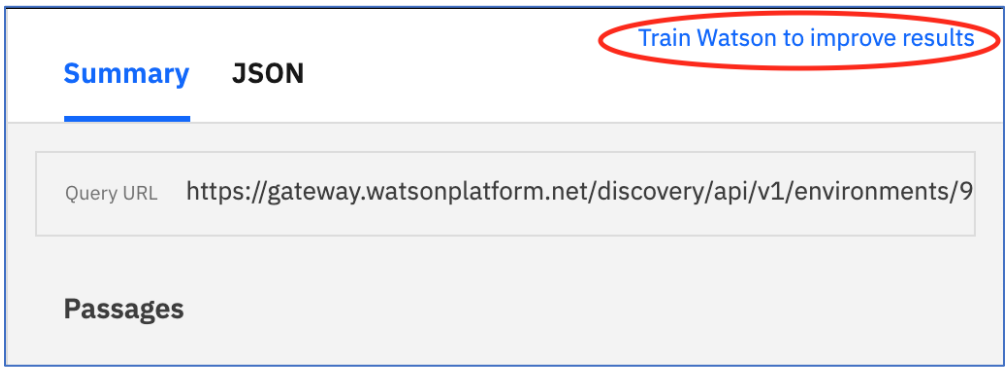
The results are pretty good, but can be improved by using relevancy training, which we will do in the next Exercise.

Exercise 10: Improve accuracy with Relevancy Training

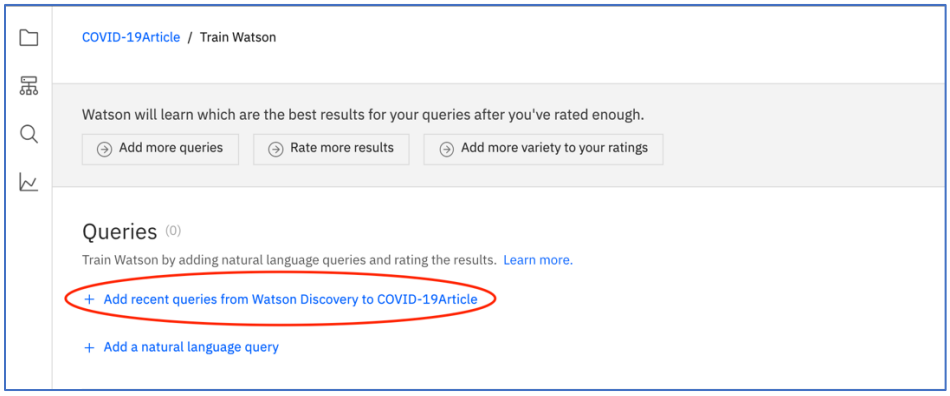
Now that we have been able to search through our document by answering specific content-related questions, we can proceed to improving the accuracy of the responses retrieved by

Watson using another capability of Watson Discovery known as Relevancy Training. Relevancy Training allows us to train Watson to improve passage retrieval results. Let's do this for one of our four natural language queries from the previous exercise.

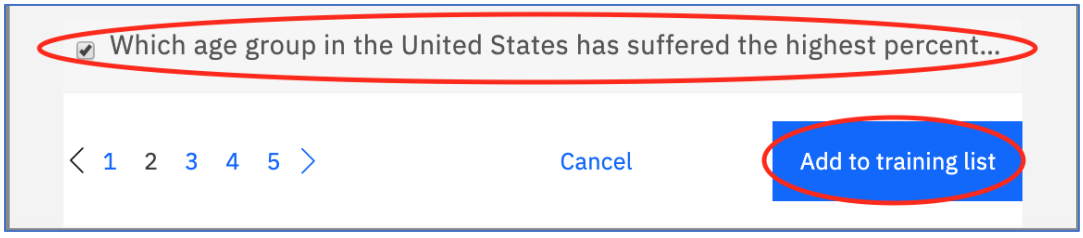
1. Click on **Train Watson to improve results** in the top right corner of the screen.



2. Now let's add all of the natural language queries that we created in the previous exercise. Click **+ Add recent queries from Watson Discovery to COVID-19Article**

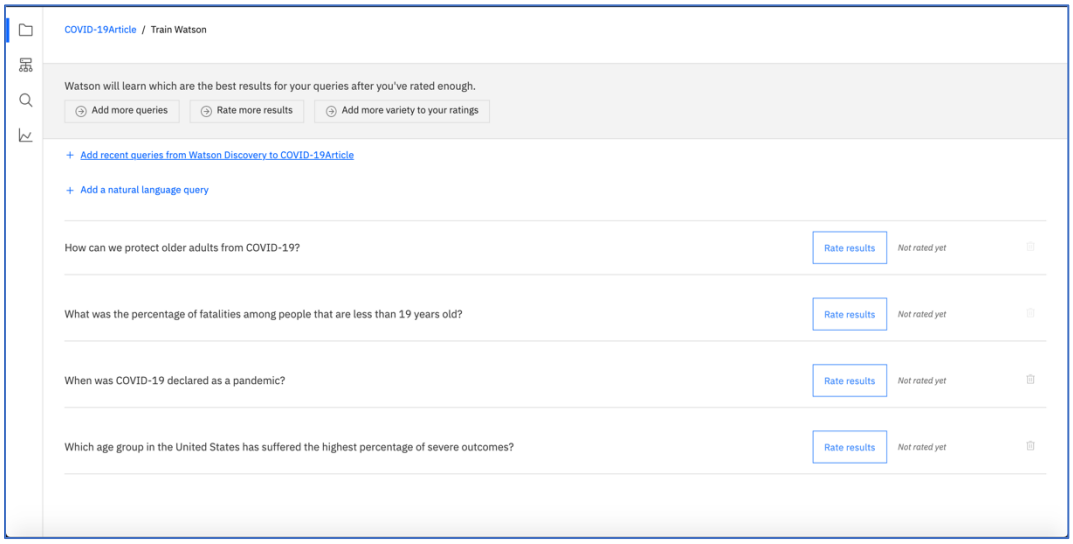


3. Search through the list of recent natural language queries and select each of the four queries from the previous exercise. After selecting all 4 queries, click **Add to training list**.

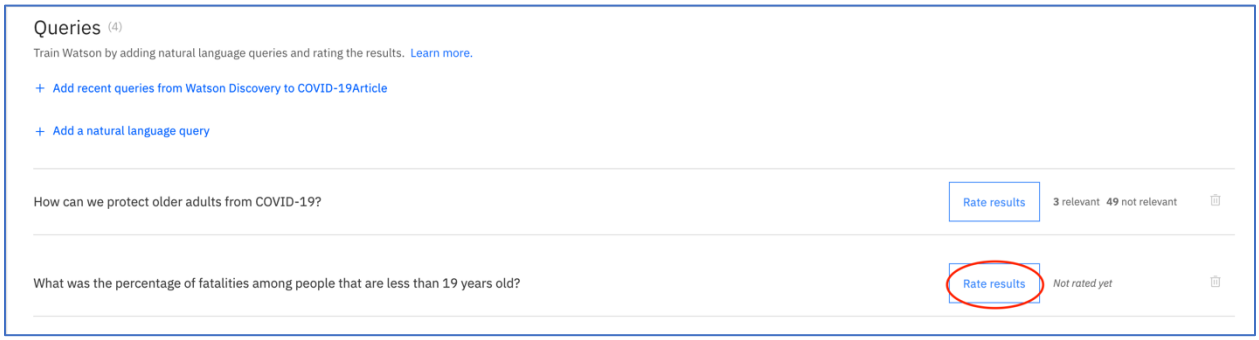


4. After adding all 4 queries to the training list, close the pop-up screen by clicking the X in the right corner of the screen.

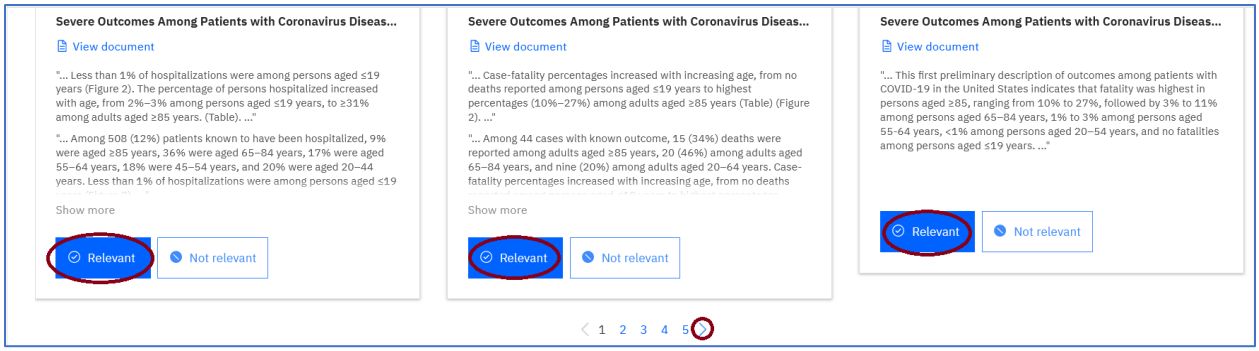
5. You should now be able to see all 4 queries listed on the screen:



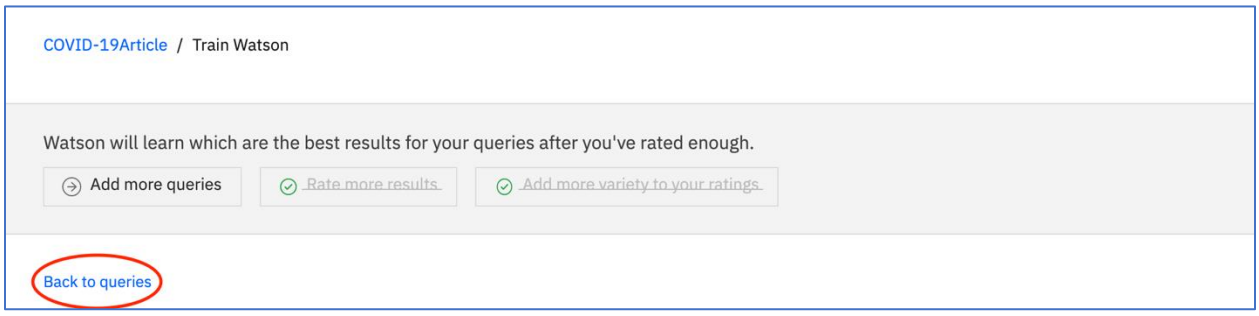
6. Let's select the query "What was the percentage of fatalities among people that are less than 19 years old?" Click **Rate Results** for that query.

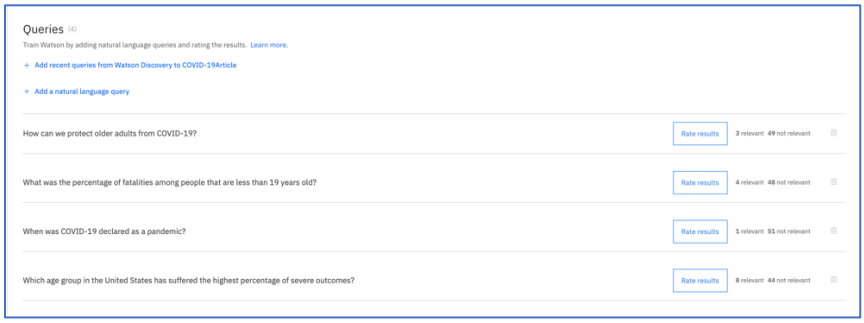


7. Review all the passages and tag all of the passages that include information about fatalities for people under 19 as **Relevant** and anything else as **Not relevant**. Make sure that you review all the passages for this query by clicking through the pages of results.



8. When you are done reviewing all the passages, click on **Back to queries** to return to the list of our natural language queries.





Now that we have completed Relevancy Training, we should be able to get more accurate results when we run trained queries in the future. This is especially useful for conversational applications containing a virtual agent (such as Watson Assistant in Lab 3), which require accurate real-time responses to user inquiries that are often more detailed and long-tail.

You have completed Lab 2!