

# **IBM Training**

## **Student Exercises**

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

Legal Copyright: © Copyright IBM Corp. 2020  
*Course materials may not be reproduced in whole or in part without the prior written permission of IBM*

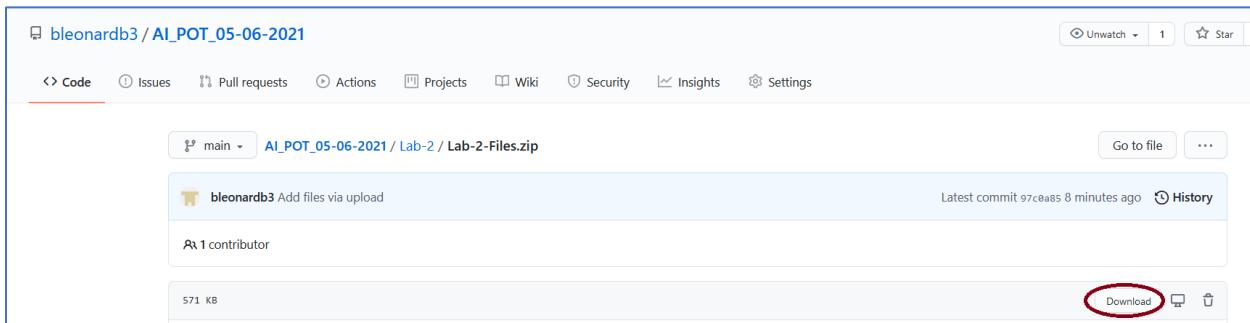
## Table of Contents

|  |           |
|--|-----------|
| <b><i>Prerequisites</i></b> .....  | <b>3</b>  |
| Download the data files to the Desktop.....                                    | 3         |
| <b><i>Introduction</i></b> .....   | <b>4</b>  |
| <b><i>Objectives</i></b> .....   | <b>4</b>  |
| <b><i>Exercise 1: Create a Discovery collection</i></b> .....                  | <b>4</b>  |
| <b><i>Exercise 2: Upload the documents</i></b> .....                           | <b>8</b>  |
| 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         |
| <b><i>Exercise 3: Add the entity model from Knowledge Studio</i></b> .....     | <b>12</b> |
| <b><i>Exercise 4: Perform Custom Entity Extraction</i></b> .....               | <b>17</b> |
| <b><i>Exercise 5: Calculate the COVID-19 vulnerability index</i></b> .....     | <b>19</b> |
| <b><i>Exercise 7: Create a collection for a COVID-19 publication</i></b> ..... | <b>25</b> |
| <b><i>Exercise 8: Perform Smart Document Understanding</i></b> .....           | <b>28</b> |
| <b><i>Exercise 9: Create and run Natural Language Queries</i></b> .....        | <b>35</b> |
| <b><i>Exercise 10: Improve accuracy with Relevancy Training</i></b> .....      | <b>41</b> |

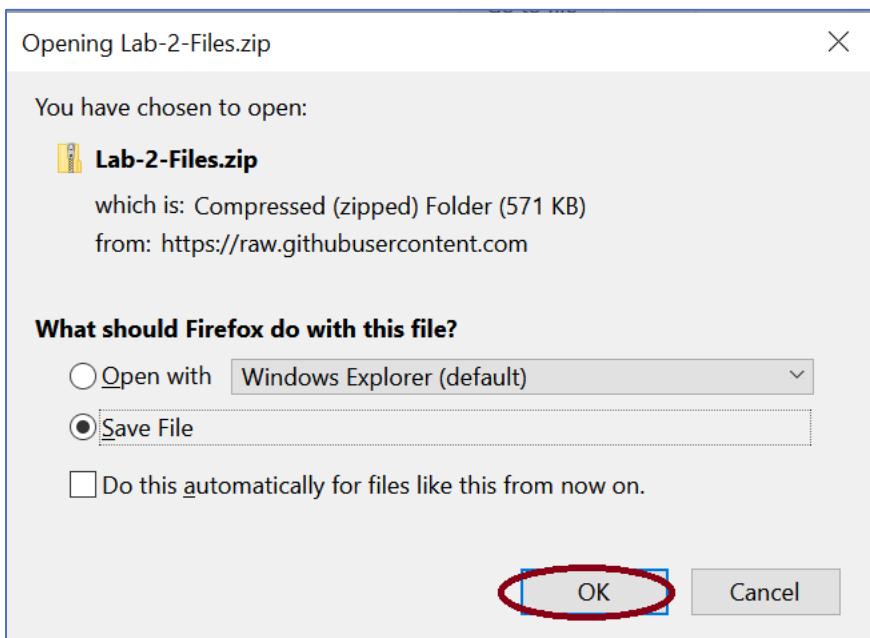
## Prerequisites

Download the data files to the Desktop

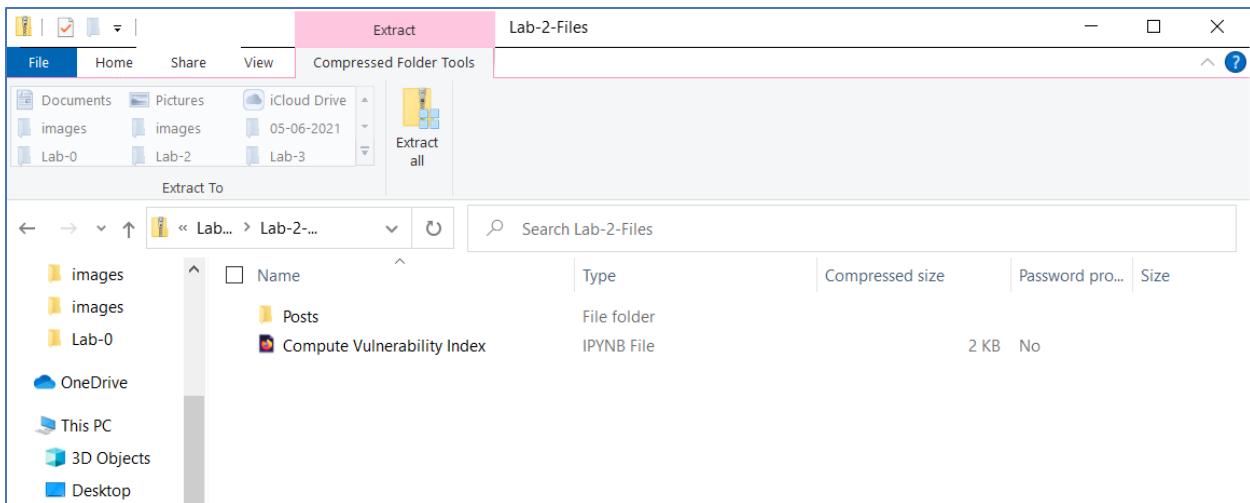
1. Click on [https://github.com/bleonardb3/AI\\_POT\\_05-06-2021/blob/main/Lab-2/Lab-2-Files.zip](https://github.com/bleonardb3/AI_POT_05-06-2021/blob/main/Lab-2/Lab-2-Files.zip)
2. Click on the **Download** button.



3. Click **OK**.



4. Extract the file contents. You should have the following directories 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

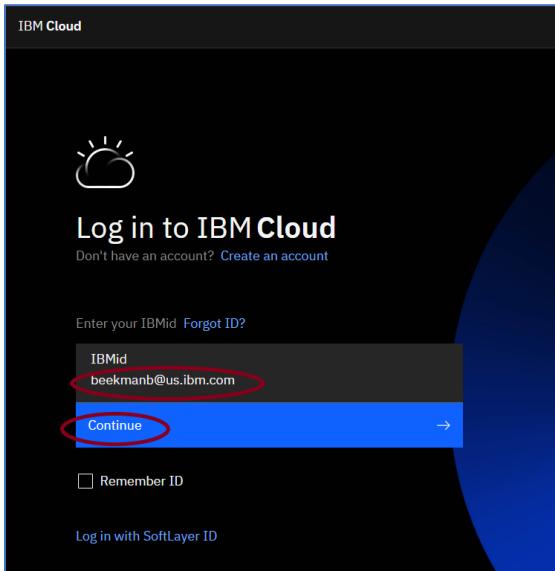
The goal of this lab is to familiarize the user with the Watson Discovery service. Watson Discovery is an enterprise AI search technology that leverages machine learning, including natural language processing (NLP), to retrieve specific answers to your questions and analyze trends and relationships buried 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:

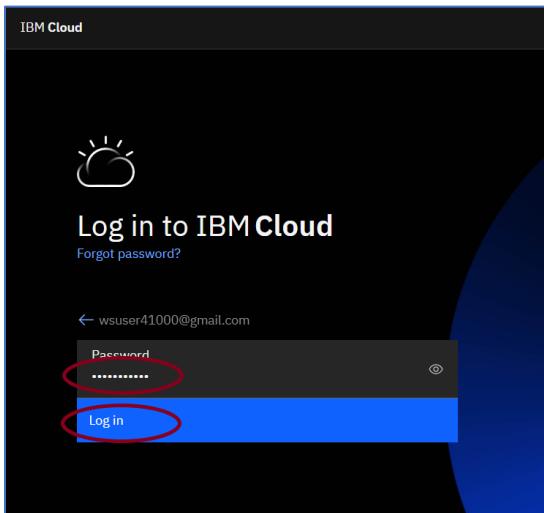
1. Create a Discovery collection
2. Upload the documents
3. Add the entity model from Knowledge Studio
4. Perform custom entity extraction
5. Retrieve the analyzed files using the Discovery API
6. Calculate the COVID-19 vulnerability index
7. Create a collection for a COVID-19 publication
8. Perform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

## Exercise 1: Create a Discovery collection

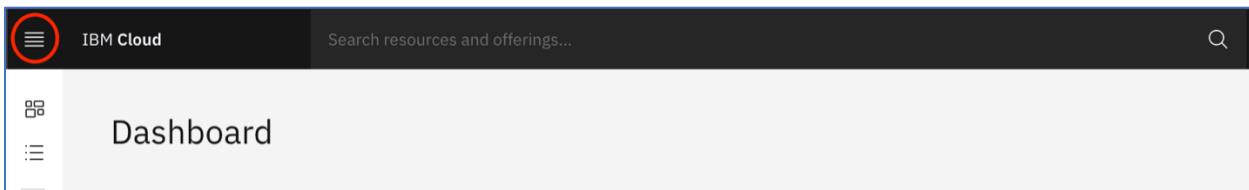
1. Log into your IBM Cloud account by typing **cloud.ibm.com** into the URL address bar of your Firefox or Chrome browser.
2. If you have been logged out, enter your **IBMid** and click **Continue**.



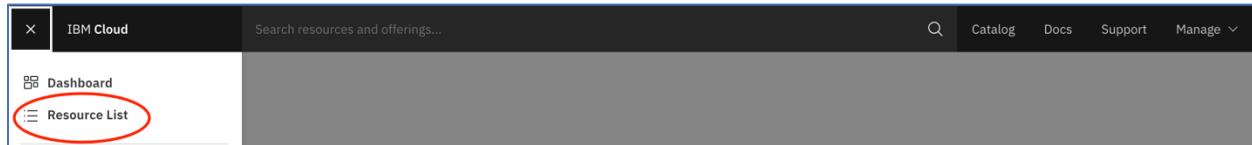
3. Enter your **Password** and click **Log in**.



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.

A screenshot of the "Resource list" interface. The title "Resource list" is at the top left. On the right side, there's a "Create resource" button. The main area shows a table with columns: Name, Group, Location, Offering, Status, and Tags. A filter bar is above the table. The "Name" column contains entries like "Discovery-kf". The entry "Discovery-kf" is circled with a red marker.

7. On the Manage screen, click on **Show Credentials** and copy the **API key** and **URL** for your Discovery instance into a Notepad or text editor for later use. We will be using these credentials when calling the Discovery API later in the lab.

A screenshot of the "Watson Discovery" manage screen. At the top left, it says "Start by launching the tool". On the right, there are "Plan" and "Advanced" sections, with "Advanced" currently selected. Below that is a "Upgrade" button. In the center, there are three buttons: "Launch Watson Discovery" (blue), "Getting started tutorial" (white), and "API reference" (white). The "API reference" button is highlighted with a blue border. Below these buttons is a section titled "Credentials". It shows an "API key:" field containing "Jfxr729jMa\_YBYKGpsxf4awFhfUP-Dy8EwpjGDZ5mqcy" and an "URL:" field containing "https://gateway.watsonplatform.net/discovery/api". Both fields are circled with red markers.

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

Resource list /  
Discovery-kf Active Add tags ↗

Manage  
Getting started  
Service credentials  
Plan  
Connections

Start by launching the tool

Launch Watson Discovery Getting started tutorial API reference

Credentials

API key: [Download](#) Show credentials

URL: <https://api.us-south.discovery.watson.cloud.ibm.com/instances/ffe3214a-5b4b-4d1d-8529-7ac4>

Plan  
Lite Upgrade

Details Actions... FEEDBACK

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

IBM Watson Discovery

Manage data Collections of your private data and pre-enriched data to configure and query against. [Learn more.](#)

Create a new data collection Create COVID-19 Kit [Upload your own data](#) Connect a data source

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

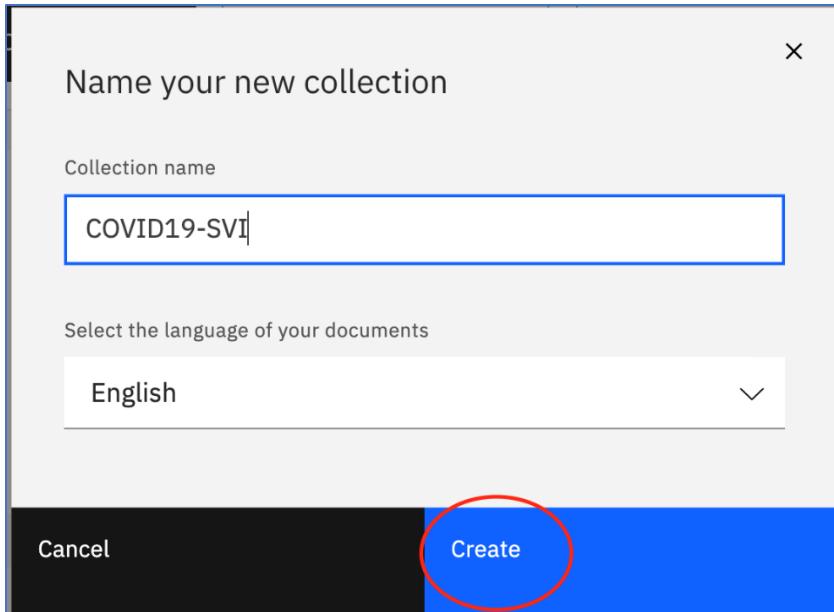
# Set up Discovery for your private data.

Your current **Lite plan** is suitable for trial & experimentation.

**Building a production application?**  
The Advanced plan provides more storage, better performance with dedicated clusters and support for a production grade workload.  
[Upgrade plan in IBM Cloud](#)

Cancel 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 content mining and passage retrieval and query the analyzed document 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.

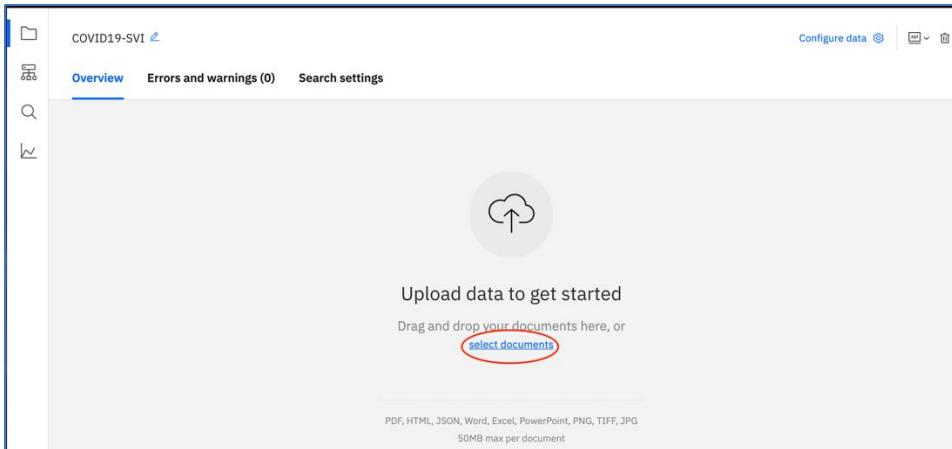
## Sentiment analysis

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.

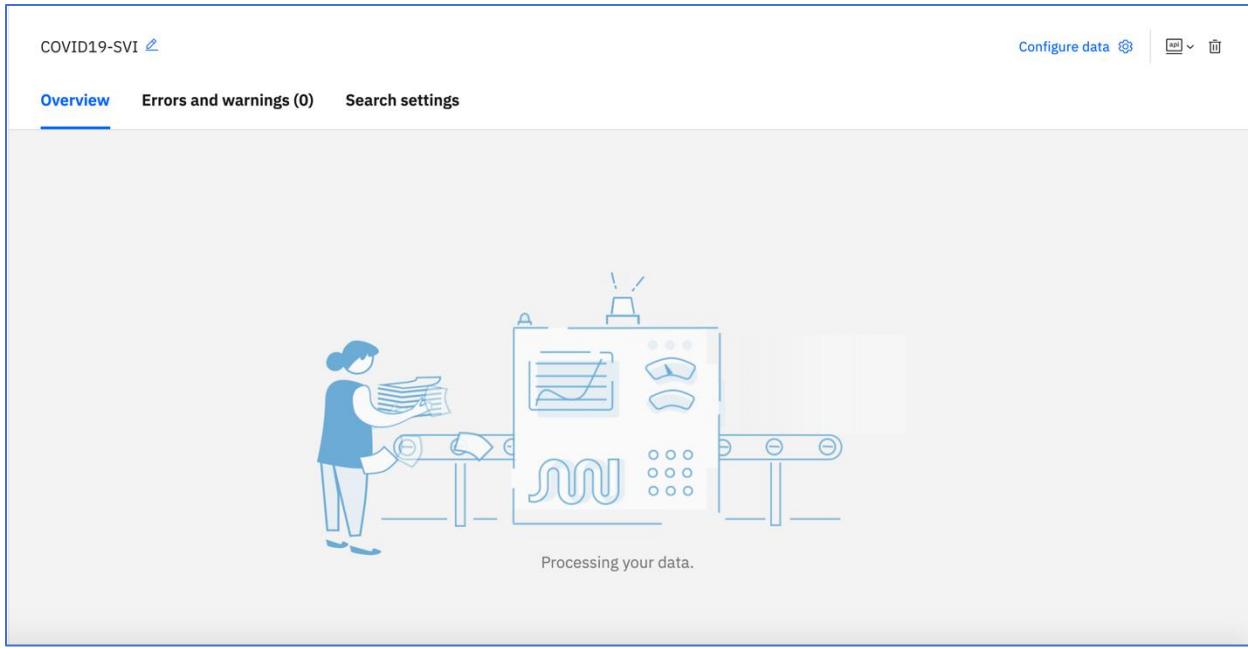
1. Click **select documents**.



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 of the documents.



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

COVID19-SVI

**Overview** Errors and warnings (0) Search settings

364 documents

0 documents failed [View details](#)

Created on 5/18/2020 12:32:41 pm EDT  
Last updated 5/18/2020 12:32:41 pm EDT

Upload documents

Identified 5 fields from your data

- text
- author\_fullname
- extracted\_metadata
- id
- title

Added 4 enrichments to your data

Entity Extraction

Seattle (36) | DC (21) | \$600 (10) | 2 weeks (10) | \$0 (9)

Sentiment Analysis

|          |     |         |    |          |     |
|----------|-----|---------|----|----------|-----|
| positive | 28% | neutral | 1% | negative | 72% |
|----------|-----|---------|----|----------|-----|

Now you're ready to query!

Entities of type **Quantity** which have negative sentiment

Run

Documents that contain English-language films, but not Landlord

Run

Top entities with their average, min, max sentiment score

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.

The screenshot shows the 'Versions' page in Watson Knowledge Studio. At the top, there's a 'Machine Learning Model' section with two buttons: 'Go to Pre-annotation page' and 'Export current model'. Below this is a table titled 'Version History and Deployment'.

| Version   | Base            | Creation Date | Entity Scores      | Relation Scores | Description     | Action  |
|---|-----------------|---------------|--------------------|-----------------|-----------------|---|
| 1.1   | Current Version |               | 0.65 (0.69 / 0.62) | N/A             |                 | <a href="#">Create Version</a>  |
| 1.0   | 05/18/2020      |               | 0.65 (0.69 / 0.62) | N/A             | 368docs-85-10-5 | <a href="#">Promote</a> <a href="#">Delete</a> <a href="#">Deploy</a> |
| <b>Deployed Models (1)</b><br>Model ID: 63d1efc3-6d00-4273-a034-7034a996c8f0 Service ID: 03b54347-0aad-4da9-b59a-e1f2df1070cc <a href="#">Undeploy</a> <a href="#">Status</a> |                 |               |                    |                 |                 |   |

## Exercise 3: Add the entity model from Knowledge Studio

1. Click **Configure Data**.

The screenshot shows the 'Overview' page in Watson Knowledge Studio. On the right side, there's a 'Configure data' button with a red circle around it. The page also displays the workspace name 'COVID19-SVI', document count (364), and various status metrics.

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**.

COVID19-SVI / Configure data

Identify fields   Manage fields **Enrich fields**

Enrich your data with additional Watson insights

Set up rules for which fields you want to apply enrichments to. [Learn more.](#)

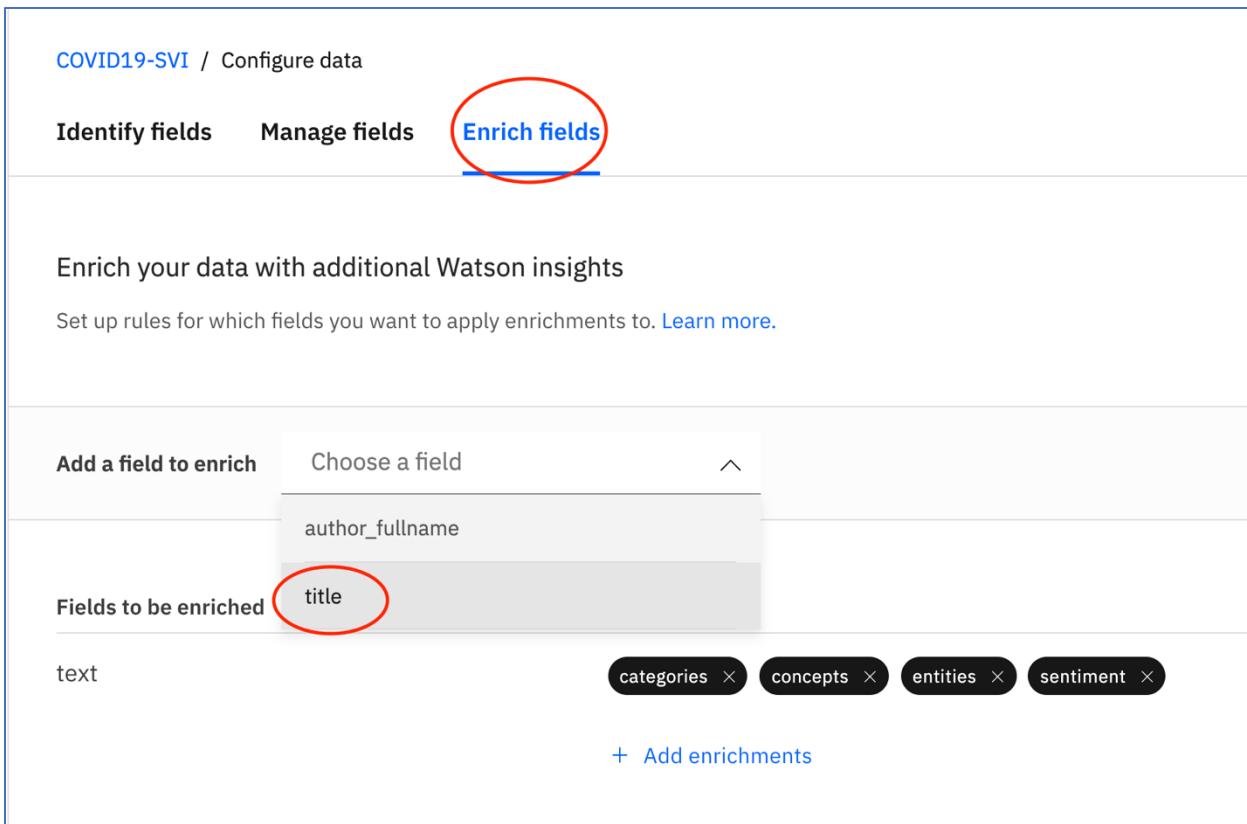
Add a field to enrich Choose a field ^

author\_fullname

Fields to be enriched title

text categories × concepts × entities × sentiment ×

+ Add enrichments



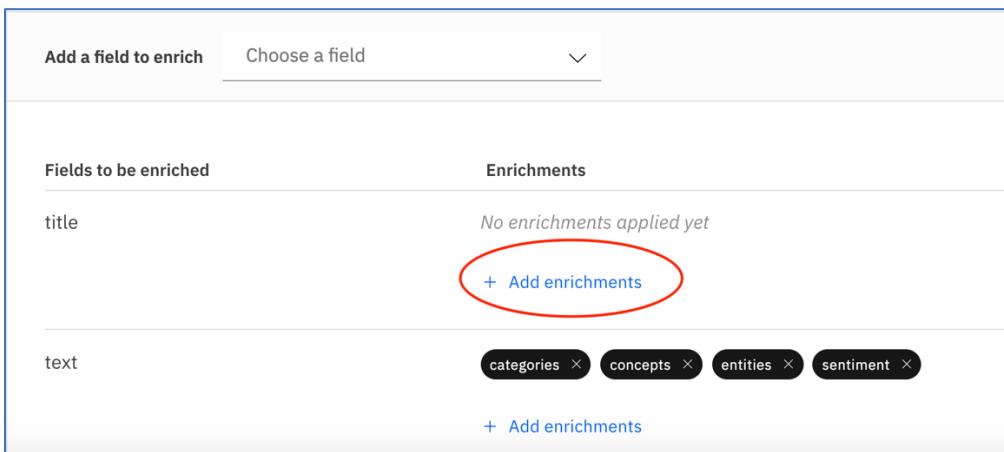
3. To the title field of each post, we will apply entity extraction. Click **+Add enrichments** on the title row.

Add a field to enrich Choose a field ^

Fields to be enriched

| Fields to be enriched | Enrichments                                     |
|-----------------------|---|
| title                 | No enrichments applied yet<br>+ Add enrichments |
| text                  | categories × concepts × entities × sentiment ×  |

+ Add enrichments



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

Add Enrichments

x

title:

**Keyword Extraction**  
Determines important keywords in this field, ranks them, and optionally detects the sentiment.

[Learn more](#) [Add](#)

**Sentiment Analysis**  
Identifies the overall positive or negative sentiment within this field.

[Learn more](#) [Add](#)

**Concept Tagging**  
Identifies general concepts that aren't necessarily directly referenced in this field.

[Learn more](#) [Add](#)

**Category Classification**  
Classifies this field into a hierarchy of categories that's five levels deep.

[Learn more](#) [Add](#)

**Semantic Role Extraction**  
Parses sentences into subject, action, and object form and returns additional semantic information.

[Learn more](#) [Add](#)

**Emotion Analysis**  
Analyzes the emotions (anger, disgust, fear, joy, and sadness) in this field.

[Learn more](#) [Add](#)

**Entity Extraction**  
Extracts people, companies, organizations, cities, geographic features, and more from this field. You can also create and add custom entity models with Watson Knowledge Studio.

[Learn more](#) [Add](#)

**Relation Extraction**  
Recognizes when two entities are related and identifies the type of relation. You can also create and add custom relation models with Watson Knowledge Studio.

[Learn more](#) [Add](#)

**Custom Model ID** ⓘ  
 [Apply](#)

**Custom Model ID** ⓘ  
 [Apply](#)



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.

## Add Enrichments

title: entities ×

**Keyword Extraction**  
Determines important keywords in this field, ranks them, and optionally detects the sentiment.

[Learn more](#) Add

**Sentiment Analysis**  
Identifies the overall positive or negative sentiment within this field.

[Learn more](#) Add

**Concept Tagging**  
Identifies general concepts that aren't necessarily directly referenced in this field.

[Learn more](#) Add

**Category Classification**  
Classifies this field into a hierarchy of categories that's five levels deep.

[Learn more](#) Add

**Semantic Role Extraction**  
Parses sentences into subject, action, and object form and returns additional semantic information.

[Learn more](#) Add

**Emotion Analysis**  
Analyzes the emotions (anger, disgust, fear, joy, and sadness) in this field.

[Learn more](#) Add

**Entity Extraction**  
Extracts people, companies, organizations, cities, geographic features, and more from this field. You can also create and add custom entity models with Watson Knowledge Studio

[Learn more](#) ✓ Added!

**Relation Extraction**  
Recognizes when two entities are related and identifies the type of relation. You can also create and add custom relation models with Watson Knowledge Studio.

[Learn more](#) Add

**Custom Model ID** ⓘ

63d1efc3-6d00-4273-a034-Apply

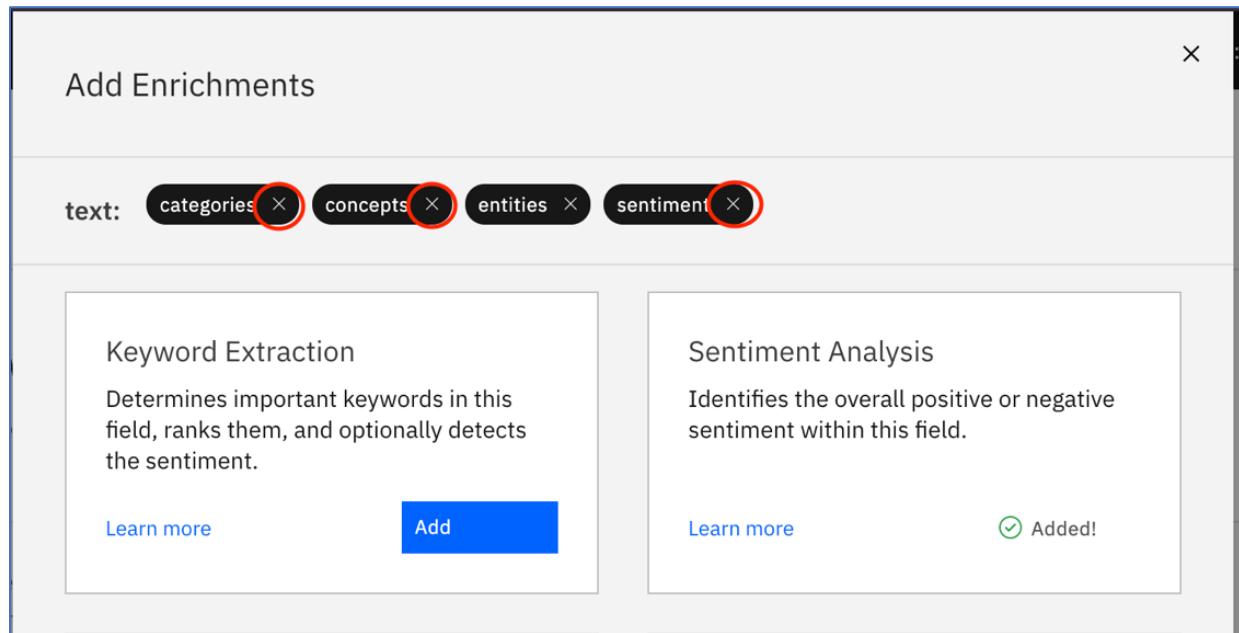
Custom Model ID ⓘ

enter model ID  Apply

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

| Fields to be enriched | Enrichments   |
|-----------------------|---|
| title                 | <a href="#">entities</a> <a href="#">×</a>  |
| text                  | <a href="#">categories</a> <a href="#">×</a> <a href="#">concepts</a> <a href="#">×</a> <a href="#">entities</a> <a href="#">×</a> <a href="#">sentiment</a> <a href="#">×</a><br><a href="#">+ Add enrichments</a> |

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.

Add Enrichments

text: entities

|  |  |
|--|--|
| Keyword Extraction<br>Determines important keywords in this field, ranks them, and optionally detects the sentiment.<br><br>Learn more      Add  | Sentiment Analysis<br>Identifies the overall positive or negative sentiment within this field.<br><br>Learn more      Add  |
| Concept Tagging<br>Identifies general concepts that aren't necessarily directly referenced in this field.<br><br>Learn more      Add   | Category Classification<br>Classifies this field into a hierarchy of categories that's five levels deep.<br><br>Learn more      Add  |
| Semantic Role Extraction<br>Parses sentences into subject, action, and object form and returns additional semantic information.<br><br>Learn more      Add   | Emotion Analysis<br>Analyzes the emotions (anger, disgust, fear, joy, and sadness) in this field.<br><br>Learn more      Add   |
| Entity Extraction<br>Extracts people, companies, organizations, cities, geographic features, and more from this field. You can also create and add custom entity models with Watson Knowledge Studio.<br><br>Learn more <input checked="" type="checkbox"/> Added!<br><br>Custom Model ID <input type="text" value="63d1efc3-6d00-4273-a034-"/> <input type="button" value="Apply"/> | Relation Extraction<br>Recognizes when two entities are related and identifies the type of relation. You can also create and add custom relation models with Watson Knowledge Studio.<br><br>Learn more      Add |

## Exercise 4: Perform Custom Entity Extraction

- 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**.

COVID19-SVI / Configure data

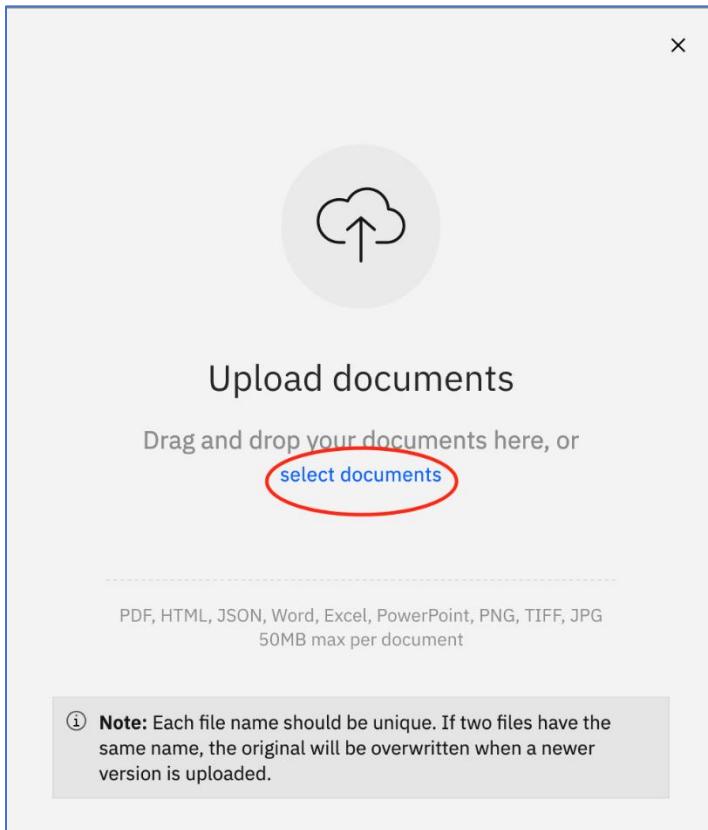
Identify fields   Manage fields   **Enrich fields**

Enrich your data with additional Watson insights  
Set up rules for which fields you want to apply enrichments to. [Learn more](#).

Add a field to enrich   Choose a field

| Fields to be enriched | Enrichments                       |
|-----------------------|-----------------------------------|
| title                 | entities                          |
|                       | + <a href="#">Add enrichments</a> |
| text                  | entities                          |
|                       | + <a href="#">Add enrichments</a> |

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**.

| Name                      | Date Modified           | Size      | Kind          |
|---------------------------|-------------------------|-----------|---------------|
| relevant-seattle-292.json | Apr 28, 2020 at 2:54 AM | 725 bytes | JSON Document |
| relevant-seattle-267.json | Apr 28, 2020 at 2:54 AM | 633 bytes | JSON Document |
| relevant-seattle-285.json | Apr 28, 2020 at 2:54 AM | 568 bytes | JSON Document |
| relevant-seattle-221.json | Apr 28, 2020 at 2:54 AM | 476 bytes | JSON Document |
| relevant-seattle-139.json | Apr 28, 2020 at 2:53 AM | 675 bytes | JSON Document |
| relevant-seattle-172.json | Apr 28, 2020 at 2:53 AM | 1 KB      | JSON Document |
| relevant-seattle-114.json | Apr 28, 2020 at 2:53 AM | 720 bytes | JSON Document |
| relevant-seattle-199.json | Apr 28, 2020 at 2:53 AM | 478 bytes | JSON Document |
| relevant-seattle-126.json | Apr 28, 2020 at 2:53 AM | 472 bytes | JSON Document |
| relevant-seattle-134.json | Apr 28, 2020 at 2:53 AM | 528 bytes | JSON Document |
| relevant-seattle-175.json | Apr 28, 2020 at 2:53 AM | 440 bytes | JSON Document |
| relevant-seattle-151.json | Apr 28, 2020 at 2:53 AM | 783 bytes | JSON Document |
| relevant-seattle-155.json | Apr 28, 2020 at 2:53 AM | 635 bytes | JSON Document |
| relevant-seattle-137.json | Apr 28, 2020 at 2:53 AM | 295 bytes | JSON Document |
| relevant-seattle-36.json  | Apr 28, 2020 at 2:52 AM | 510 bytes | JSON Document |
| relevant-seattle-14.json  | Apr 28, 2020 at 2:52 AM | 542 bytes | JSON Document |
| relevant-seattle-13.json  | Apr 28, 2020 at 2:52 AM | 3 KB      | JSON Document |
| relevant-seattle-12.json  | Apr 28, 2020 at 2:52 AM | 775 bytes | JSON Document |
| relevant-seattle-65.json  | Apr 28, 2020 at 2:52 AM | 2 KB      | JSON Document |
| relevant-seattle-94.json  | Apr 28, 2020 at 2:52 AM | 896 bytes | JSON Document |
| relevant-seattle-42.json  | Apr 28, 2020 at 2:52 AM | 6 KB      | JSON Document |
| relevant-seattle-64.json  | Apr 28, 2020 at 2:52 AM | 955 bytes | JSON Document |
| relevant-seattle-5.json   | Apr 28, 2020 at 2:52 AM | 640 bytes | JSON Document |
| relevant-seattle-86.json  | Apr 28, 2020 at 2:52 AM | 790 bytes | JSON Document |
| relevant-seattle-8.json   | Apr 28, 2020 at 2:52 AM | 714 bytes | JSON Document |
| relevant-seattle-32.json  | Apr 28, 2020 at 2:52 AM | 856 bytes | JSON Document |
| relevant-seattle-24.json  | Apr 28, 2020 at 2:52 AM | 653 bytes | JSON Document |
| relevant-seattle-97.json  | Apr 28, 2020 at 2:52 AM | 615 bytes | JSON Document |
| relevant-seattle-44.json  | Apr 28, 2020 at 2:52 AM | 1 KB      | JSON Document |
| relevant-seattle-25.json  | Apr 28, 2020 at 2:52 AM | 350 bytes | JSON Document |
| relevant-seattle-79.json  | Apr 28, 2020 at 2:52 AM | 924 bytes | JSON Document |
| relevant-seattle-51.json  | Apr 28, 2020 at 2:52 AM | 303 bytes | JSON Document |

Cancel

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:

The screenshot shows the IBM Watson Discovery interface for the 'COVID19-SVI' dataset. The top navigation bar includes 'Configure data' and 'Upload documents'. The main overview section displays the following information:

- 364 documents**
- 0 documents failed** (View details)
- Created on**: 5/18/2020 12:32:41 pm EDT
- Last updated**: 5/18/2020 12:32:41 pm EDT
- Upload documents**

Below the overview, there are three main sections:

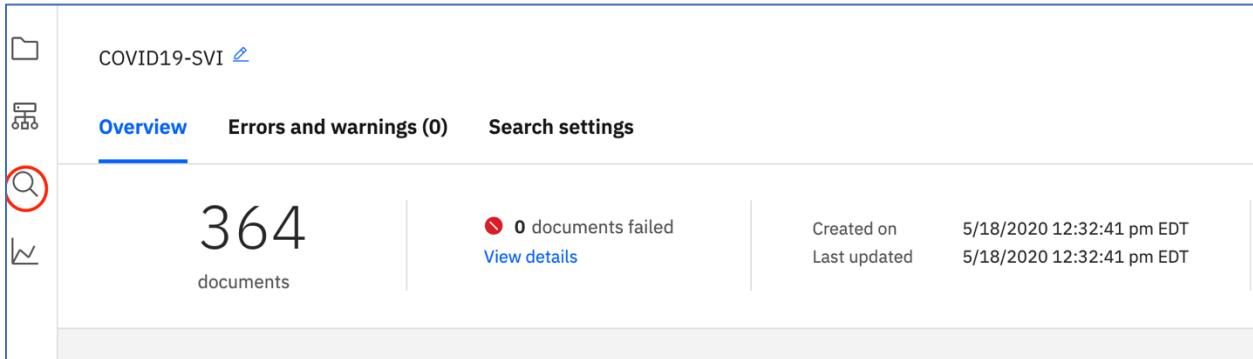
- Identified 5 fields from your data:** text, author\_fullname, extracted\_metadata, id, title.
- Added 1 enrichment to your data:** Entity Extraction, which includes terms like unemployment (61), rent (27), apartment (26), lease (17), and apartments (10). It also notes 8 enrichments available and a link to add more.
- Now you're ready to query!** This section contains three examples:
  - Top people related to /society/work/unemployment (Run button)
  - Entities of type Unemployed which have negative sentiment (Run button)
  - Top entities with their average, min, max sentiment score (Run button)

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

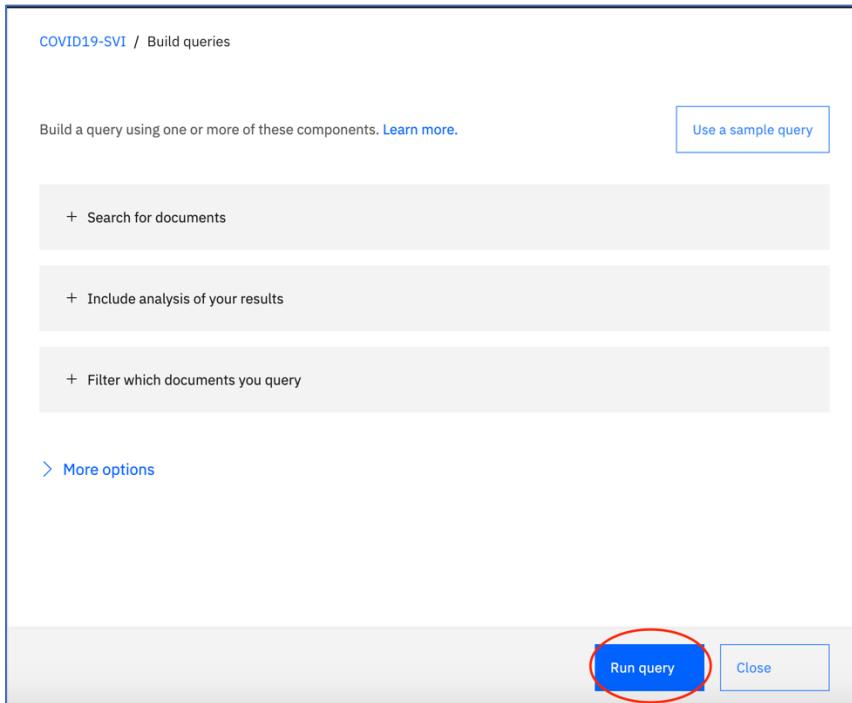
Let's start by viewing the entity extraction output inside the Discovery UI.

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



The screenshot shows the Watson Studio interface. On the left, there's a sidebar with icons for files, datasets, and queries. The 'Query' icon is circled in red. The main area is titled 'COVID19-SVI' with a blue link. Below it, there are three tabs: 'Overview' (which is selected and highlighted in blue), 'Errors and warnings (0)', and 'Search settings'. The 'Overview' section displays a large number '364' and the text 'documents'. To the right of this, it says '0 documents failed' with a 'View details' link. At the bottom right, it shows 'Created on 5/18/2020 12:32:41 pm EDT' and 'Last updated 5/18/2020 12:32:41 pm EDT'.

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**.



The screenshot shows the 'Build queries' page. At the top, it says 'COVID19-SVI / Build queries'. Below that, there's a message 'Build a query using one or more of these components. [Learn more.](#)' and a 'Use a sample query' button. There are three main options listed: '+ Search for documents', '+ Include analysis of your results', and '+ Filter which documents you query'. At the bottom, there's a 'More options' link and two buttons: 'Run query' (which is circled in red) and 'Close'.

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.

The screenshot shows the Watson Discovery service interface. At the top, there are tabs for 'Summary' and 'JSON', with 'JSON' being the active tab and circled in red. Below the tabs is a 'Train Watson to improve results' button. A 'Query URL' field contains the value `https://gateway.watsonplatform.net/discovery/api/v1/environment`. The main content area displays a JSON response with a single result document. The document has fields like 'matching\_results', 'session\_token', 'passages', and 'results'. One of the 'results' objects is expanded, showing an 'id' (f7181e61aec2d2de8e67f34ff137ca07), 'result\_metadata', 'author\_fullname' (t2\_mguoa), and an 'enriched\_text' field containing a long paragraph about unemployment insurance. There are also 'entities' and other nested fields. At the bottom left are 'Run query' and 'Close' buttons.

```

{
  "matching_results": 364,
  "session_token": "1_0SECbKHa0o613VG4_xrczEdrmX",
  "passages": [],
  "results": [
    {
      "id": "f7181e61aec2d2de8e67f34ff137ca07",
      "result_metadata": {...},
      "author_fullname": "t2_mguoa",
      "enriched_title": {...},
      "text": "I have reached a point of not knowing what else to do, other than reach out to news agencies to possibly shed light on the situation of Unemployment Insurance here in NYC. \n\nI, along with thousands of others, filed for UIB over a month ago and have yet to receive any word about whether my claim is accepted or not. Getting in touch with the DoL is impossible, their lines don't even have a queue option. After spending over 5 minutes in a touch-tone operating system, you are advised that they are 'experiencing a high volume of callers' and to 'try back later'. After calling everyday for weeks now, I am no closer to getting an answer than I was when I started this. \n\nGroups on Facebook like Restaurant Worker Solidarity NYC will show hundreds of posts of people in the same position. None of us have any idea about what to do, and there's no money coming in. I haven't received any income for a month and a half, have not paid rent, am depending on my sweetheart to help pay for groceries, and still have tuition pay...",
      "enriched_text": {...},
      "entities": [...]
    }
  ]
}

```

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.

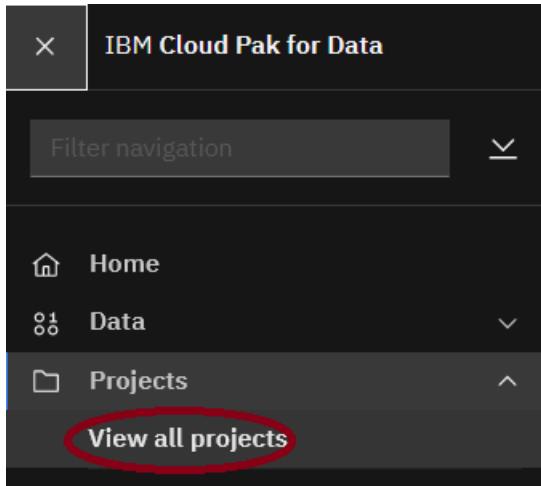
- Click on the **IBM Cloud Pak for Data** tab to navigate to Watson Studio. If you had closed this tab, you can enter <http://dataplatform.cloud.ibm.com> in the browser to navigate to Watson Studio.



- Click on the hamburger  icon



- Click on **View all projects**.



## 7. Click on Watson Studio Labs.

A screenshot of the "Projects" page in IBM Cloud Pak for Data. The title bar says "Projects". Below it is a search bar with placeholder text "Which project are you looking for?" and a dropdown menu set to "All my projects". The main area shows a table of projects. The first row, "Watson Studio Labs", is highlighted with a red oval. The columns are "Name", "Role", "Storage", "Collaborators", and "Creator". The "Name" column shows "Watson Studio Labs", "Role" shows "Admin", "Storage" shows "COS", "Collaborators" shows a user icon, and "Creator" shows "Jack Doe".

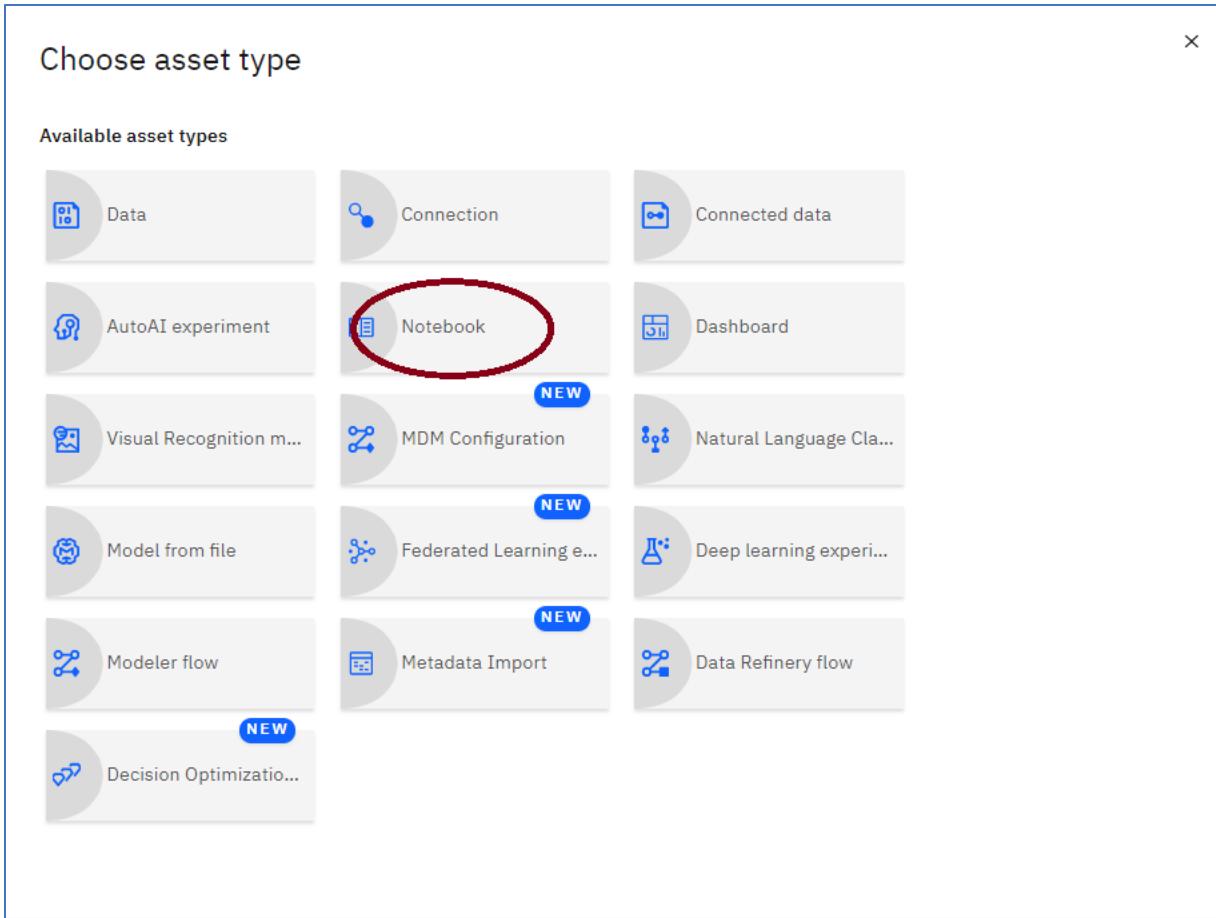
## 8. Click on the Assets tab.

A screenshot of the "Watson Studio Labs" project view. The title bar says "Projects / Watson Studio Labs". Below it is a navigation bar with tabs: "Overview", "Assets" (which is highlighted with a blue box and a red oval), "Environments", "Jobs", "Access Control", and "Settings".

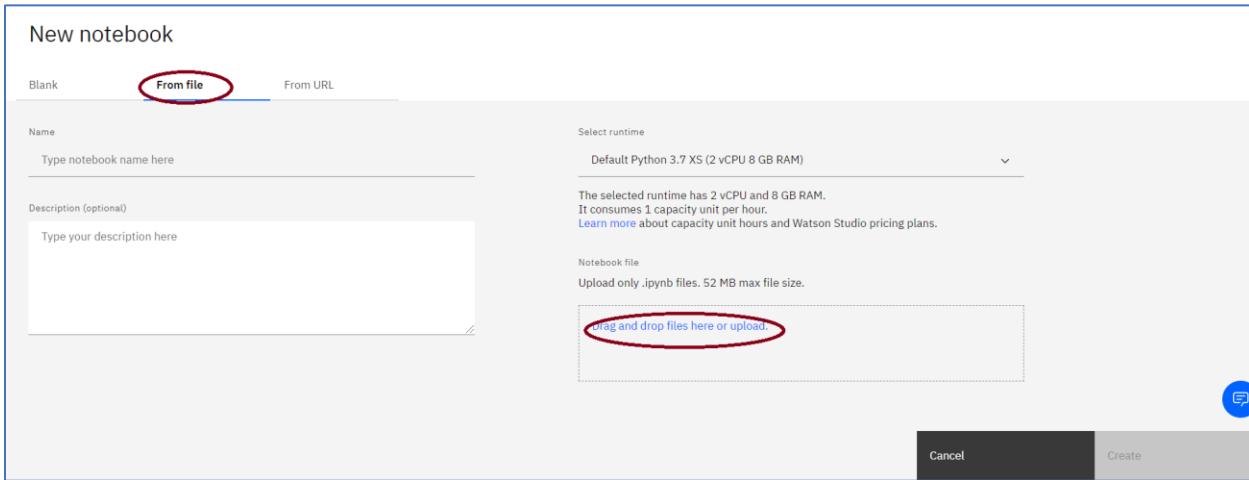
## 9. Click on Add to project.

A screenshot of the "Assets" tab for the "Watson Studio Labs" project. The title bar includes "IBM Cloud Pak for Data" and "Upgrade". Below it is a search bar and a "Launch IDE" button. The main area shows a table with columns: "Overview", "Assets" (which is highlighted with a blue box and a red oval), "Environments", "Jobs", "Access Control", and "Settings". At the top right of the main area is a "Launch IDE" button with a plus sign and the text "Add to project" with a red oval around it.

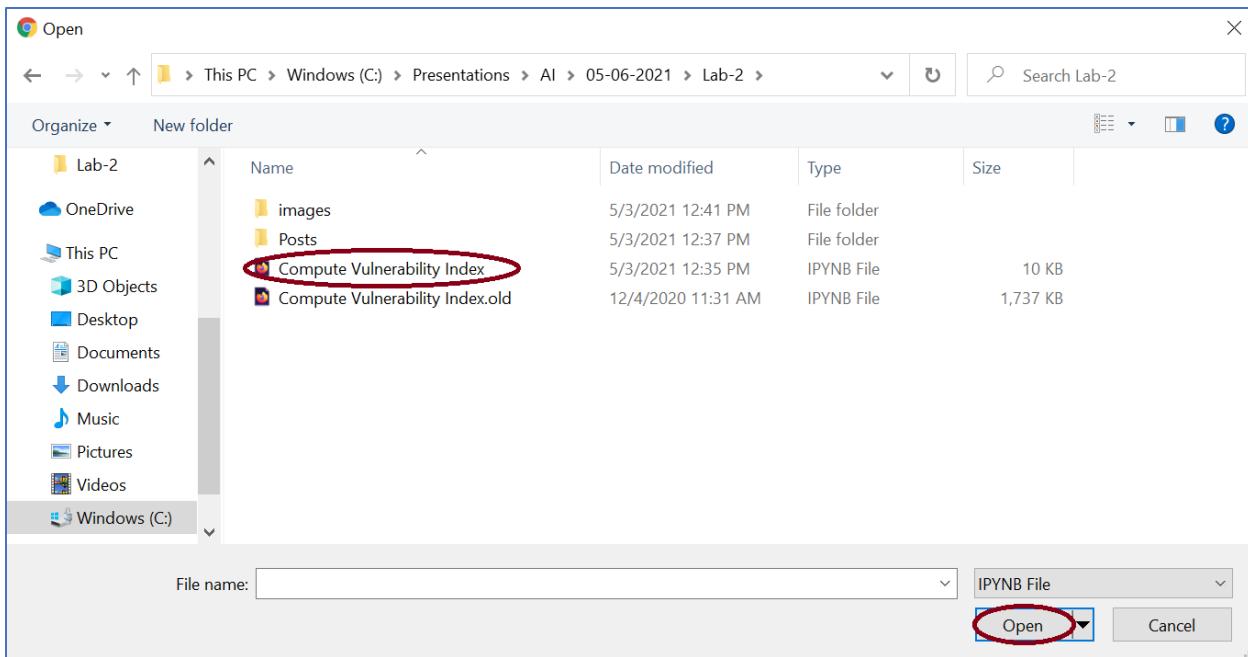
## 10. Click Notebook



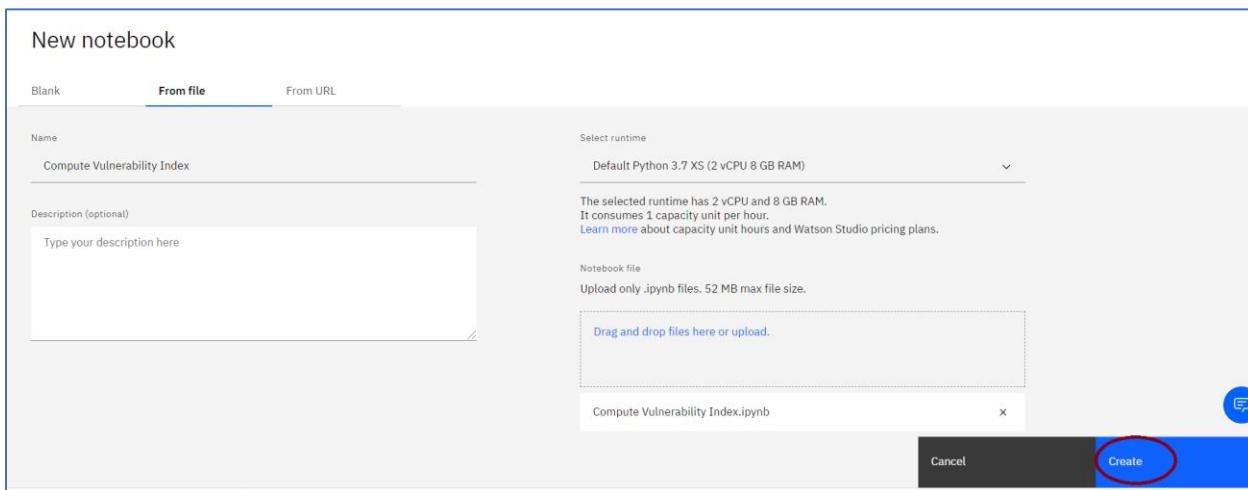
11. Click on the **File** tab, and then click on **Drag or drop files here or upload**.



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, then highlight the INSERT WATSON DISCOVERY API KEY as shown below**

```
#### Insert the Watson Discovery API key
In [ ]: authenticator = IAMAuthenticator('INSERT WATSON DISCOVERY API KEY')

discovery = DiscoveryV1(
    version='2019-04-30',
    authenticator=authenticator)
discovery.set_service_url('https://gateway.watsonplatform.net/discovery/api')
```

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



```
#### Insert the Watson Discovery API key
In [ ]: authenticator = IAMAuthenticator('Cnq8n_Xv4Hkvg0H7nxgzadRVLzMgycIrlaCtZRmyAkWf')
discovery = DiscoveryV1(
    version='2019-04-30',
    authenticator=authenticator)
discovery.set_service_url('https://gateway.watsonplatform.net/discovery/api')

#### Display Watson Discovery Environment ID
```

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.

16. 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>

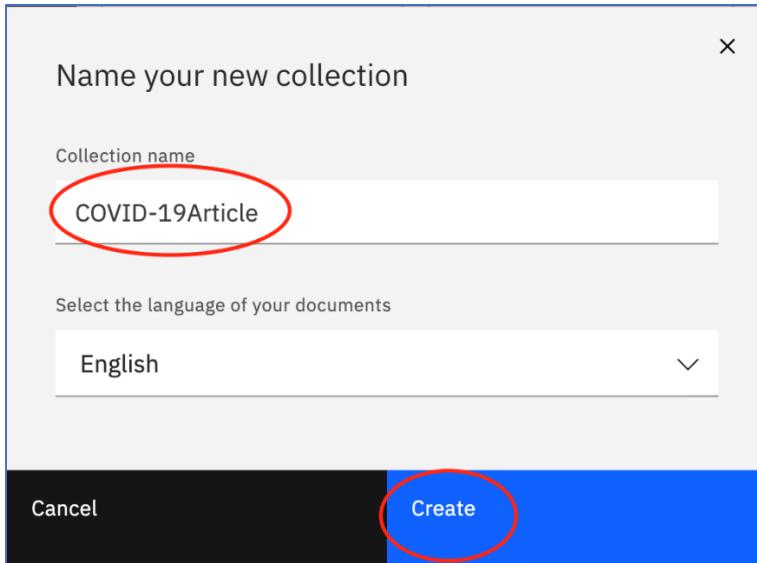
2. Download the file to your workstation. Open up the downloaded pdf file to verify that you have the complete article, which spans 4 pages. We will be uploading this article to a brand new collection in Watson Discovery.

The screenshot shows the first page of a CDC report. At the top, a red-bordered box contains the text "Please note: This report has been corrected." Below this, the title "Morbidity and Mortality Weekly Report" is displayed. The main title of the report is "Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020". The subtitle "CDC COVID-19 Response Team" is below the title. A small note at the bottom left indicates the report was posted on March 18, 2020. The right side of the screenshot is a solid gray vertical bar.

3. Open up your Discovery instance (you may have to log in again) and click **Upload your own data**.

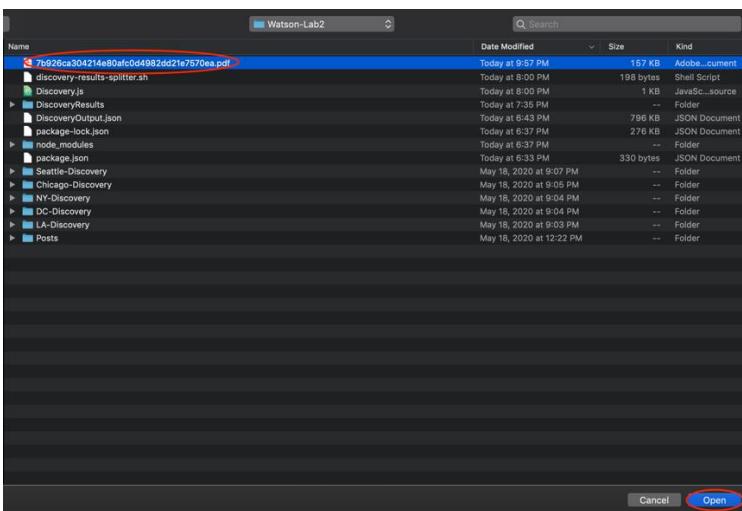
The screenshot shows the "Manage data" section of the Watson Discovery interface. It features a sidebar with icons for folder, document, and search. The main area displays the text "Manage data" and "Collections of your private data and pre-enriched data to configure and query against. [Learn more](#)". Below this are three buttons: "Create a new data collection", "Create COVID-19 Kit", and "Upload your own data" (which is circled in red). There is also a "Connect a data source" button.

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

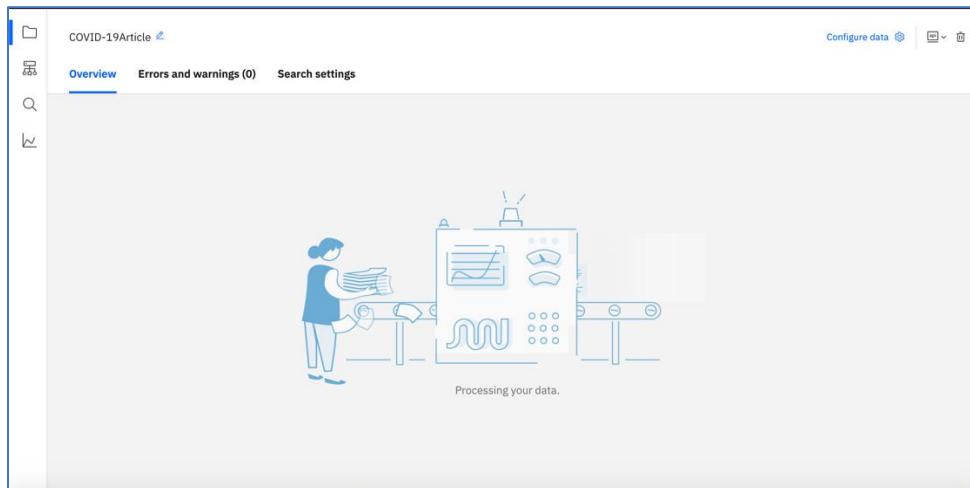


5. Inside the new collection, click **select documents**.

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



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



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

| Identified 1 field from your data |  | Added 4 enrichments to your data  |  |   | Now you're ready to query!     |   |
|-----------------------------------|--|---|--|---|--------------------------------|---|
| <b>text</b>                       | <a href="#">Need to identify more fields? Add fields</a> | <b>Entity Extraction</b>  | <b>Sentiment Analysis</b>                  | <b>Concept Tagging</b>  | <b>Category Classification</b> | <a href="#">Top entities with their average, min, max sentiment score</a>     |
|                                   |  | 19 years (1)   44 years (1)   64 years (1)<br>84 years (1)   85 years (1) | 0% positive<br>0% neutral<br>100% negative | Ageing (1)   Death (1)   Health care (1)   Medicine (1)   Old age (1) | health and fitness → disease   | <a href="#">Run</a>   |
|                                   |  |   |  |   |                                | <a href="#">Entities of type JobTitle which have negative sentiment</a>       |
|                                   |  |   |  |   |                                | <a href="#">Run</a>   |
|                                   |  |   |  |   |                                | <a href="#">Documents about Washington as a Location with a very negative</a> |

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....” 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.

Please note: This report has been updated.

## Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) - United States, February 12–March 16, 2020

(Continued from previous page)

On March 16, 2020, the agency received its 4th COVID-19 daily update from the CDC. Globally approximately 17,000 confirmed cases of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have been reported, including approximately 2,000 deaths in approximately 170 countries (1). Of the 17,000 cases, 1,000 were reported in the United States. The COVID-19 outbreak is pandemic (2). Data from China indicate that the risk for severe outcomes among patients with COVID-19 associated illness and death does not increase with age (3). In contrast, data from the United States show that the risk for death associated with COVID-19 is greater in older persons (4). In the report, COVID-19 cases in the United States that occurred during February 12–March 16, 2020, were categorized by age, sex, race/ethnicity, admission to intensive care (ICU), and death status. Data from the United States indicate that 1,000 COVID-19 cases in the United States have been reported to CDC, with multiple case reports. Cases like those being tracked by CDC are often described as “probable.” Of these cases, 10% are hospitalized, 30% are ICU admissions, and 80% of deaths are associated with hospitalization, while approximately 10% with the highest proportion of deaths among persons aged ≥65 years. In contrast, no ICU admissions or deaths were reported among persons aged ≤19 years. Thus, if probabilities suddenly appeared data for laboratory-confirmed cases of COVID-19 among persons in the report include both COVID-19 cases confirmed by state or local public health departments and cases reported by a provider at the state or local public health laboratory and confirmation at CDC. No data on serious underlying health conditions are available. Data are about cases preliminary and are subject to review for confirmation of

the labels below.

+ Create new Upgrade

- answer
- author
- footer
- header
- question
- subtitle
- table\_of\_contents
- text
- title
- image
- table

Viewing: Your training

Submit 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. 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**.

Identify document elements using the labels below.

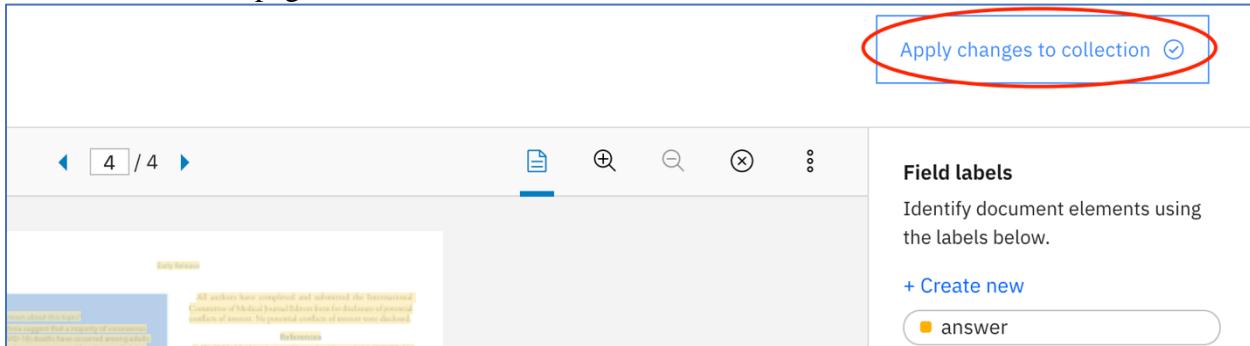
+ Create new Upgrade

- answer
- author
- footer
- header
- question
- subtitle
- table\_of\_contents
- text
- title
- image
- table

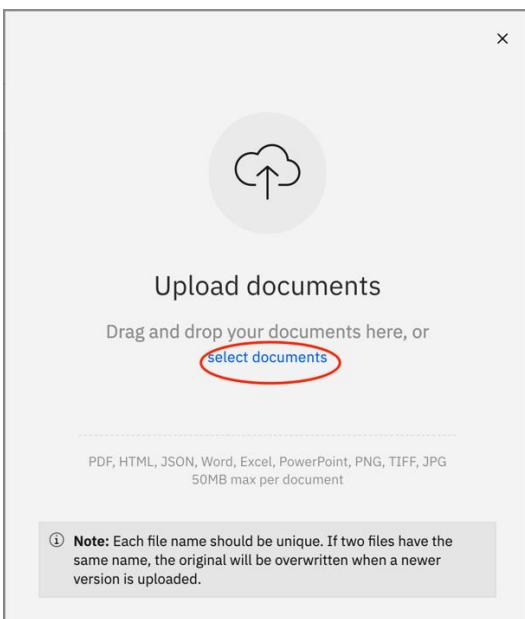
Viewing: Your training

Submit page

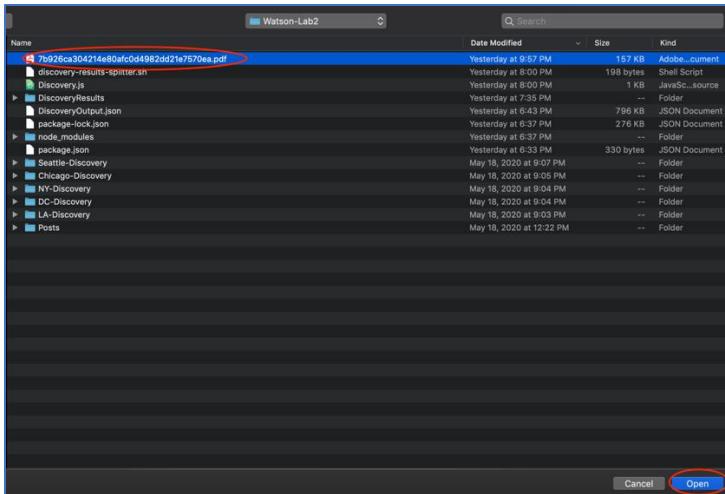
8. Click the **Apply changes to collection** button on the top right corner of the screen to save the labelled pages.



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.

| Created on                | Last updated              |
|---------------------------|---------------------------|
| 5/24/2020 10:15:40 pm EDT | 5/24/2020 10:15:40 pm EDT |

Identified 1 field from your data      Added 4 enrichments to your data      Now you're ready to query!

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**.

| Created on                | Last updated              |
|---------------------------|---------------------------|
| 5/24/2020 10:15:40 pm EDT | 5/24/2020 10:15:40 pm EDT |

13. Click on **Manage fields**.

COVID-19Article / Configure data

Identify fields **Manage fields** Enrich fields

Identify fields to index

All fields are indexed by default. Switch off any fields you do not want to be indexed. [Learn more.](#)

answer  On  
author  On  
caption  On  
footer  On  
graphTitle  On  
header  On  
image  On

Improve query results by splitting your documents

You can split your documents into segments based on fields. Once split, each segment is a separate document that will be enriched, indexed, and returned as a query separately. [Learn more.](#)

+ Split document

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**.

Identify fields to index

All fields are indexed by default. Switch off any fields you do not want to be indexed. [Learn more.](#)

answer  On  
author  On  
caption  On

Improve query results by splitting your documents

You can split your documents into segments based on fields. Once split, each segment is a separate document that will be enriched, indexed, and returned as a query separately. [Learn more.](#)

+ Split document

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

Improve query results by splitting your documents

You can split your documents into segments based on fields. Once split, each segment is a separate document that will be enriched, indexed, and returned as a query separately. [Learn more.](#)

Split document on each occurrence of

Select field

subtitle  
table  
table\_of\_contents  
**text**  
title

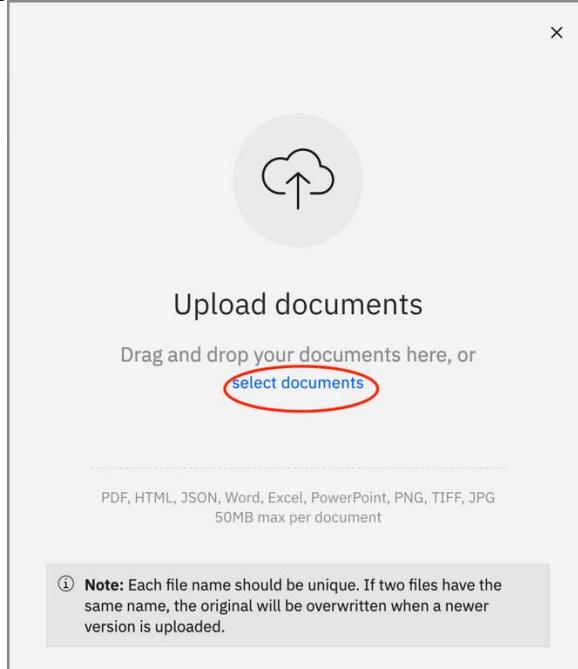
16. Click **Apply changes to collection** for the journal article to be split on each occurrence of the text field.

Apply changes to collection

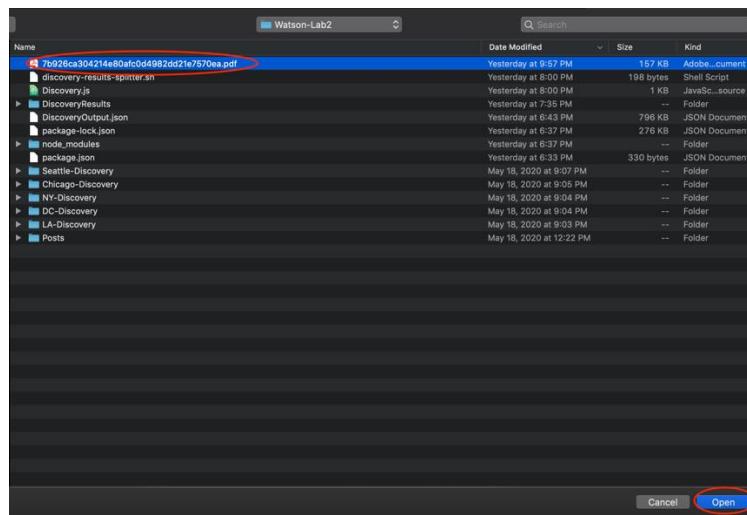
Improve query results by splitting your documents

You can split your documents into segments based on fields. Once split, each segment is a separate document that will be enriched, indexed, and returned as a query separately. [Learn more.](#)

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 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).

The screenshot shows the Watson Discovery Overview page. At the top, it says "Overview" and "Errors and warnings (0)" with a "Search settings" button. Below that, it shows "51 documents". To the right, there's a status bar with "0 documents failed" and "View details", "Created on 5/4/2021 10:54:08 am EDT", "Last updated 5/4/2021 10:54:08 am EDT", and a "Upload documents" button.

On the left, under "Identified 2 fields from your data", it lists "text" and "title". Below that, a button says "Need to identify more fields? Add fields".

In the center, under "Added 4 enrichments to your data", there are four boxes: "Entity Extraction" (CDC (13) | United States (9) | 19 years (7) | US Department of Health and H... (6) |), "Sentiment Analysis" (4% positive, 68% neutral, 28% negative), "Concept Tagging" (Epidemiology (11) | Medicine (11) | United States (8) | Death (7) | Ageing (6) |), and "Category Classification" (health and fitness → disease → epidemic). Below these, a button says "5 enrichments available. Add enrichments".

On the right, under "Now you're ready to query!", there are three sections: "Documents that contain Epidemiology, but not Medicine" (Run), "Top entities with their average, min, max sentiment score" (Run), and "Entities of type Organization which have negative sentiment" (Run).

## 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.

The screenshot shows the Watson Discovery Overview page. On the left, there are several icons: a folder, a document, a magnifying glass (circled in red), and a list. The magnifying glass icon represents the Query page. Below the icons, it says "COVID-19Article" with a link. Underneath the icons, it says "Overview" (which is blue and underlined), "Errors and warnings (0)", and "Search settings". To the right, it shows "52 documents" and a status bar with "0 documents failed" and "View details".

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**.

COVID-19Article / Build queries

Build a query using one or more of these components. [Learn more.](#)

**Use a sample query**

+ Search for documents Entities of type **JobTitle** which have negative sentiment

+ Include analysis of your results Top entities with their average, min, max sentiment score

**Most common entity types and their top entities** (highlighted with a red oval)

+ Filter which documents you query Top people related to /health and fitness/disease

+ More options Documents about Washington as a Location with a very negative sentiment

Documents that contain Ageing, but not Death

Entities of type **JobTitle** which have positive sentiment

**Run query** **Close**

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

Train Watson to improve results

**Summary** **JSON**

Query URL: <https://gateway.watsonplatform.net/discovery/api/v1/environments/9>

**Aggregations**

- term(enriched\_text.entities.type) **Quantity** (82)
  - term(enriched\_text.entities.text) **19 years** (7)
  - term(enriched\_text.entities.text) **84 years** (6)
  - term(enriched\_text.entities.text) **64 years** (5)
  - term(enriched\_text.entities.text) **85 years** (5)
  - term(enriched\_text.entities.text) **31%** (4)
  - term(enriched\_text.entities.text) **44 years** (4)
  - term(enriched\_text.entities.text) **53%** (3)
  - term(enriched\_text.entities.text) **54 years** (3)
  - term(enriched\_text.entities.text) **60 years** (3)
  - term(enriched\_text.entities.text) **65 years** (3)
- term(enriched\_text.entities.type) **Location** (32)
  - term(enriched\_text.entities.text) **United States** (8)
  - term(enriched\_text.entities.text) **Atlanta** (6)
  - term(enriched\_text.entities.text) **China** (5)
  - term(enriched\_text.entities.text) **Washington** (3)
  - term(enriched\_text.entities.text) **GA** (2)
  - term(enriched\_text.entities.text) **Geneva** (2)
  - term(enriched\_text.entities.text) **Switzerland** (2)
  - term(enriched\_text.entities.text) **U.S.** (2)
  - term(enriched\_text.entities.text) **Japan** (1)
  - term(enriched\_text.entities.text) **Wuhan** (1)
- term(enriched\_text.entities.type) **Organization** (32)
  - term(enriched\_text.entities.text) **CDC** (15)
  - term(enriched\_text.entities.text) **US Department of Health and Human Services** (8)

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

Include analysis of your results

Write an aggregation query using the Discovery Query Language

```
nested(enriched_text.entities).term(enriched_text.entities.type,count:5).term(enriched_text.entities.text)
```

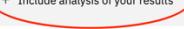
Build in visual mode  

?

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

COVID-19Article / Build queries

Build a query using one or more of these components. [Learn more.](#) [Use a sample query](#)

- + Search for documents
- + **Include analysis of your results** 
- + Filter which documents you query

> More options

Run query Close

## 6. From the **Field** drop-down, select **enriched\_text.entities.text**

Include analysis of your results

Output Field Count

Top values   10

+ Add child aggregations

+ Add top-level aggregation

enriched\_text.entities.disambiguation.name

enriched\_text.entities.disambiguation.subtype

enriched\_text.entities.sentiment.label

**enriched\_text.entities.text** 

+ Filter which documents you query

enriched\_text.entities.type

Run query Close

## 7. Click + **Add child aggregation**

Include analysis of your results

[Edit in query language](#)

| Output     | Field                       | Count |
|------------|-----------------------------|-------|
| Top values | enriched_text.entities.text | 10    |

+ Add condition

+ Add child aggregation **(circled)**

+ Add top-level aggregation

term(enriched\_text.entities.text,count:10)

- Under **Output**, select **Average** and under **Field**, select **enriched\_text.entities.sentiment.score**. Click **Run query** to view the results.

Include analysis of your results

[Edit in query language](#)

| Output     | Field                       | Count |
|------------|-----------------------------|-------|
| Top values | enriched_text.entities.text | 10    |

+ Add condition

| Output  | Field                                  |
|---------|--|
| Average | enriched_text.entities.sentiment.score |

+ Add child aggregation

+ Add top-level aggregation

**Run query** **CLOSE**

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

Train Watson to improve results

**Summary** **JSON**

Query URL <https://gateway.watsonplatform.net/discovery/api/v1/environments/9>

**Aggregations**

```

term(enriched_text.entities.text) CDC (15)
• average (enriched_text.entities.sentiment.score): 0.08059964556962025
term(enriched_text.entities.text) United States (8)
• average (enriched_text.entities.sentiment.score): -0.04446059322033898
term(enriched_text.entities.text) 19 years (7)
• average (enriched_text.entities.sentiment.score): -0.0229655641025641
term(enriched_text.entities.text) 84 years (6)
• average (enriched_text.entities.sentiment.score): -0.007347934426229509
term(enriched_text.entities.text) Atlanta (6)
• average (enriched_text.entities.sentiment.score): 0.1365073750000001
term(enriched_text.entities.text) US Department of Health and Human Services (6)
• average (enriched_text.entities.sentiment.score): 0.14098004761904762
term(enriched_text.entities.text) 64 years (5)
• average (enriched_text.entities.sentiment.score): -0.008300444444444445
term(enriched_text.entities.text) 85 years (5)
• average (enriched_text.entities.sentiment.score): -0.02315672413793103
term(enriched_text.entities.text) China (5)
• average (enriched_text.entities.sentiment.score): -0.06504192105263158
term(enriched_text.entities.text) 31% (4)
• average (enriched_text.entities.sentiment.score): -0.027981041666666664

```

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.

Include analysis of your results

Write an aggregation query using the Discovery Query Language

```
term(enriched_text.entities.text,count:10).average(enriched_text.entities.sentiment.score)
```

11. Click on + Search for documents.

Build a query using one or more of these components. [Learn more.](#)

[Use a sample query](#)

+ Search for documents

+ Include analysis of your results

+ Filter which documents you query

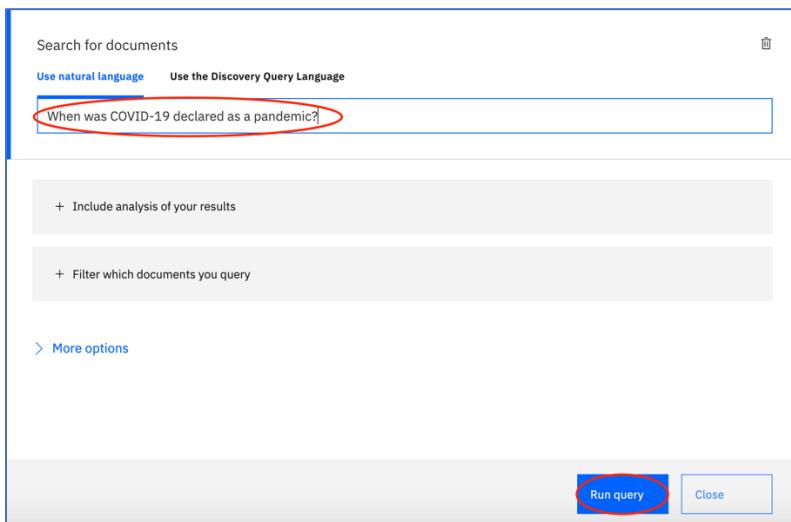
Here we can create natural language queries to ask specific questions about our document. After a quick glance at our article, we can ask the following questions:

- 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 in order to answer this question.

A screenshot of a search results page titled 'Passages'. The page displays five retrieved documents as passages:

- "On March 11, 2020, the World Health Organization declared the COVID-19 outbreak a pandemic (2). Data from China have indicated that older adults, particularly those with serious underlying health conditions, are at higher risk for severe COVID-19-associated illness and death than are younger persons (3)."
- "Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020"
- "Data from China have indicated that older adults, particularly those with serious underlying health conditions, are at higher risk for severe COVID-19-associated illness and death than are younger persons (3). Although the majority of reported COVID-19 cases in China were mild (81%), approximately"
- "Although the majority of reported COVID-19 cases in China were mild (81%), approximately 80% of deaths occurred among adults aged ≥60 years; only one (0.1%) death occurred in a person aged ≤19 years (3). In this report, COVID-19 cases in the United States that occurred during February"
- "CDC. Coronavirus disease 2019 (COVID-19): if you are at higher risk."

14. Repeat steps 13 and 14 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?

**Passages**

"FIGURE 2. COVID-19 hospitalizations,\* intensive care unit (ICU) admissions,\$^{\dagger}\$ and deaths,\$^{\ddagger}\$ by age group – United States, February 12–"

"TABLE. Hospitalization, intensive care unit (ICU) admission, and case-fatality percentages for reported COVID-19 cases, by age group – United States, February 12–March 16, 2020 Age group (yrs)"

". \$^{\dagger}\$ Cases identified before February 28 were aggregated and reported during March 1–3. aged ≥65 years with the highest percentage of severe outcomes among persons aged ≥85 years. These findings are similar to data from China, which indicated >80% of deaths occurred among persons aged ≥60 years (3 )."

"Overall, 31% of cases, 45% of hospitalizations, 53% of ICU admissions, and 80% of deaths associated with COVID-19 were among adults aged ≥65 years with the highest percentage of severe outcomes among persons aged ≥85 years. In contrast, no ICU admissions or deaths were reported among persons aged ≤19 years."

"What is added by this report? This first preliminary description of outcomes among patients with COVID-19 in the United States indicates that fatality was highest in persons aged ≥85, ranging from 10% to 27%, followed by 3% to 11% among persons aged 65–84 years, 1% to 3% among persons aged 55–64 years, <1% among persons aged 20–54 years, and no fatalities among persons aged ≤19 years."

How can we protect older adults from COVID19?

**Passages**

"COVID-19 can result in severe disease, including hospitalization, admission to an intensive care unit, and death, especially among older adults. Everyone can take actions, such as social distancing, to help slow the spread of COVID-19 and protect older adults from severe illness."

"to further reduce the risk of being exposed (7). Persons of all ages and communities can take actions to help slow the spread of COVID-19 and protect older adults. ↑ <https://www.cdc.gov/coronavirus/2019-ncov/downloads/communitymitigation-strategy.pdf>. \$^{\dagger}\$ [https://www.whitehouse.gov/wp-content/uploads/2020/03/03.16.20\\_coronavirus-guidance\\_8.5x11\\_315PM.pdf](https://www.whitehouse.gov/wp-content/uploads/2020/03/03.16.20_coronavirus-guidance_8.5x11_315PM.pdf)."

"Social distancing is recommended for all ages to slow the spread of the virus, protect the health care system, and help protect vulnerable older adults. Further, older adults should maintain adequate supplies of nonperishable foods and at least a 30-day supply of necessary medications, take precautions"

"What are the implications for public health practice? COVID-19 can result in severe disease, including hospitalization, admission to an intensive care unit, and death, especially among older adults."

"\* The risk for serious disease and death in COVID-19 cases among persons in the United States increases with age. Social distancing is recommended for all ages to slow the spread of the virus, protect the health care system, and help protect vulnerable older adults."

What is % among people less than 19?

**Passages**

"This first preliminary description of outcomes among patients with COVID-19 in the United States indicates that fatality was highest in persons aged ≥85, ranging from 10% to 27%, followed by 3% to 11% among persons aged 65–84 years, 1% to 3% among persons aged 55–64 years, <1% among persons aged 20–54 years, and no fatalities among persons aged ≤19 years."

"Less than 1% of hospitalizations were among persons aged ≤19 years (Figure 2). The percentage of persons hospitalized increased with age, from 2%–3% among persons aged ≤19 years, to ≥31% among adults aged ≥85 years. (Table)."'

"Case-fatality percentages increased with increasing age, from no deaths reported among persons aged ≤19 years to highest percentages (10%–27%) among adults aged ≥85 years (Table) (Figure 2)."

"aged 20–44 years (Figure 2). No ICU admissions were reported among persons aged ≤19 years. Percentages of ICU admissions were lowest among adults aged 20–44 years (2%–4%) and highest among adults aged 75–84 years (11%–31%) (Table)."

"Overall, 31% of cases, 45% of hospitalizations, 53% of ICU admissions, and 80% of deaths associated with COVID-19 were among adults aged ≥65 years with the highest percentage of severe outcomes among persons aged ≥85 years. In contrast, no ICU admissions or deaths were reported among persons aged ≤19 years."

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 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.

The screenshot shows the Watson Discovery interface with the 'Summary' tab selected. At the top right, there is a blue button labeled 'Train Watson to improve results'. Below it, the 'Query URL' is displayed as <https://gateway.watsonplatform.net/discovery/api/v1/environments/9>. The main area is titled 'Passages'.

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**

The screenshot shows the 'Train Watson' interface for the 'COVID-19Article' environment. It includes icons for folder, file, search, and list. Below the search bar, it says 'Watson will learn which are the best results for your queries after you've rated enough.' with three buttons: '+ Add more queries', '+ Rate more results', and '+ Add more variety to your ratings'. Under the 'Queries' section, it says '(0)' and 'Train Watson by adding natural language queries and rating the results. [Learn more.](#)'. A red circle highlights the '+ Add recent queries from Watson Discovery to COVID-19Article' button. Below it are '+ Add a natural language query' and '+ Add a recent query'.

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**.

The screenshot shows a pop-up window with a list of queries. The first query, 'Which age group in the United States has suffered the highest percent...', has a checked checkbox. A red circle highlights both the checkbox and the 'Add to training list' button at the bottom right. Navigation arrows and a 'Cancel' button are also visible.

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.

Select recent queries from Watson Discovery for COVID-19Article to train



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

The screenshot shows the Watson Discovery interface for training a COVID-19 Article. It includes a sidebar with icons for file, search, and refresh, and a header 'COVID-19Article / Train Watson'. Below the header, it says 'Watson will learn which are the best results for your queries after you've rated enough.' with three buttons: 'Add more queries', 'Rate more results', and 'Add more variety to your ratings'. There are two sections: '+ Add recent queries from Watson Discovery to COVID-19Article' and '+ Add a natural language query'. Under the first section, there are four queries with 'Rate results' buttons and 'Not rated yet' status: 'How can we protect older adults from COVID-19?', 'What was the percentage of fatalities among people that are less than 19 years old?', 'When was COVID-19 declared as a pandemic?', and 'Which age group in the United States has suffered the highest percentage of severe outcomes?'. Each query row also has a trash icon.

6. Let's work with the first query (How can we protect older adults from COVID-19?) by clicking on the **Rate Results** button on the same row.

The screenshot shows the same Watson Discovery interface as before, but with a red circle highlighting the 'Rate results' button for the first query: 'How can we protect older adults from COVID-19?'. The rest of the interface remains the same, including the sidebar, header, and other query sections.

7. We can go through all of the documents returned for this query and mark each as Relevant or Not relevant. For this question, we are looking for passages that can answer the question of how to protect older adults from COVID-19. Any passage that answers this question should be marked as **Relevant**; any other passage should be marked as **Not relevant**. Make sure that you review all the passages for this query by clicking through the pages of results.

Watson will learn which are the best results for your queries after you've rated enough.

Add more queries    Rate more results    Add more variety to your ratings

How can we protect older adults from COVID-19?

Rate some documents as relevant or not relevant results for this query. [Learn more.](#)

[View document](#)

... COVID-19 can result in severe disease, including hospitalization, admission to an intensive care unit, and death, especially among older adults. Everyone can take actions, such as social distancing, to help slow the spread of COVID-19. ...

... to further reduce the risk of being exposed? (7). Persons of all ages and communities can take actions to help slow the spread of COVID-19 and protect older adults. <sup>†</sup> Acknowledgments State and local health

Show more

Relevant    Not relevant

[View document](#)

... Social distancing is recommended for all ages to slow the spread of the virus, protect the health care system, and help protect vulnerable older adults. Further, older adults should maintain adequate supplies of nonprescription goods and at least a 30-day supply of necessary medications. ...

... 69 Summary What is already known about this topic? Early data from CDC's COVID-19 Response Team indicate that approximately 19 deaths have occurred among adults aged ≥60 years old among persons with serious underlying health conditions. ...

Show more

Relevant    Not relevant

[View document](#)

... Administration for Community Living, 2017 profile of older Americans. ...

... 69 Summary What is already known about this topic? Early data from CDC's COVID-19 Response Team indicate that approximately 19 deaths have occurred among adults aged ≥60 years old among persons with serious underlying health conditions. ...

Show more

Relevant    Not relevant

1 2 3 4 5

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

Watson will learn which are the best results for your queries after you've rated enough.

Add more queries    Rate more results    Add more variety to your ratings

[Back to queries](#)

How can we protect older adults from COVID-19?

Rate some documents as relevant or not relevant results for this query. [Learn more.](#)

- You should have about 3 relevant results and 49 not relevant results for the first query. Click on the **Rate Results** button next to What was the percentage of fatalities among people that are less than 19 years old?

Queries (4)

Train Watson by adding natural language queries and rating the results. [Learn more.](#)

+ Add recent queries from Watson Discovery to COVID-19Article  
+ Add a natural language query

How can we protect older adults from COVID-19?

Rate results   3 relevant 49 not relevant

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

Rate results   Not rated yet

- Review all of the passages for this second query and tag all of the passages that mention the fatality rate for people <= 19 years old as **Relevant** and anything else as **Not relevant**.

COVID-19Article / Train Watson

Watson will learn which are the best results for your queries after you've rated enough.

Add more queries  Rate more results  Add more variety to your ratings.

[Back to queries](#)

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

Rate some documents as relevant or not relevant results for this query. [Learn more.](#)

Severe Outcomes Among Patients with Coronavirus Disease ...  
 View document

"... Less than 1% of hospitalizations were among persons aged ≥19 years (Figure 2). The percentage of persons hospitalized increased with age, from 2%–7% among persons aged ≥19 years, to ≥35% among adults aged ≥85 years."...

Show less  
 Relevant  Not relevant

Severe Outcomes Among Patients with Coronavirus Disease ...  
 View document

"... This first preliminary description of outcomes among patients with COVID-19 in the United States indicates that fatality was highest in persons aged ≥85, ranging from 10% to 27%, followed by 3% to 11% among persons aged 65–74 years, and 1% to 3% among persons aged 55–64 years, <1% among persons aged 20–54 years, and no fatalities among persons aged ≤19 years..."

Relevant  Not relevant

Severe Outcomes Among Patients with Coronavirus Disease ...  
 View document

"... Case-fatality percentages increased with increasing age, from no deaths reported among persons aged ≤19 years to highest percentages (10%–27%) among adults aged ≥85 years (Table) (Figure 2)."...

Relevant  Not relevant

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

COVID-19Article / Train Watson

Watson will learn which are the best results for your queries after you've rated enough.

Add more queries  Rate more results  Add more variety to your ratings.

[Back to queries](#)

12. Click on the **Rate Results** button next to When was COVID-19 declared as a pandemic?

Queries (4)

Train Watson by adding natural language queries and rating the results. [Learn more.](#)

+ Add recent queries from Watson Discovery to COVID-19Article  
+ Add a natural language query

---

How can we protect older adults from COVID-19?  
 Rate results 3 relevant 49 not relevant

---

What was the percentage of fatalities among people that are less than 19 years old?  
 Rate results 4 relevant 48 not relevant

---

When was COVID-19 declared as a pandemic?  
 Rate results Not rated yet

13. Review all of the passages for this third query and tag all of the passages that mention exactly when COVID-19 was declared as a pandemic as **Relevant** and anything else as **Not relevant**.

Watson will learn which are the best results for your queries after you've rated enough.

Add more queries    Rate more results    Add more variety to your ratings

[Back to queries](#)

When was COVID-19 declared as a pandemic?

Rate some documents as relevant or not relevant results for this query. [Learn more.](#)

**Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) in the United States**

View document

... On March 11, 2020, the World Health Organization declared the COVID-19 outbreak a pandemic. (2.) Data from China have indicated that older adults, particularly those with underlying health conditions, are at higher risk for severe COVID-19-associated illness and death than are younger persons (3.)...

... Data from China have indicated that older adults, particularly those with serious underlying health conditions, are at higher risk for severe COVID-19-associated illness and death than are younger persons (3.). Although the majority of reported COVID-19 cases in China were mild (81%), approximately 80% of deaths occurred among adults aged ≥60 years; only one (0.1%) death occurred in a person aged <19 years (3.). In this report, COVID-19 cases in the United States that occurred during February...

Show less

Relevant    Not relevant

**Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) in the United States**

View document

... Approximately 49 million U.S. persons are aged ≥65 years (4), and many of these adults, who are at risk for severe COVID-19-associated illness, might depend on services and support to maintain their health and independence. ...

... Approximately 49 million U.S. persons are aged ≥65 years (9), and many of these adults, who are at risk for severe COVID-19-associated illness, might depend on services and support to maintain their health and independence. To prepare for potential COVID-19 illness among persons at high risk, family members and caregivers of older adults should know what medications they are taking and ensure...

... To prepare for potential COVID-19 illness among persons at high risk, family members and caregivers of older adults should know what medications they are taking and taking any medications that are required and ensure medical supplies are available. Long-term care facilities should be particularly vigilant to prevent the introduction and spread of COVID-19 (10). In addition, clinicians who...

Show less

Relevant    Not relevant

**Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) in the United States**

View document

... On March 11, 2020, the World Health Organization declared the COVID-19 outbreak a pandemic. (2.) Data from China have indicated that older adults, particularly those with underlying health conditions, are at higher risk for severe COVID-19-associated illness and death than are younger persons (3.)...

... Data from China have indicated that older adults, particularly those with serious underlying health conditions, are at higher risk for severe COVID-19-associated illness and death than are younger persons (3.). Although the majority of reported COVID-19 cases in China were mild (81%), approximately 80% of deaths occurred among adults aged ≥60 years; only one (0.1%) death occurred in a person aged <19 years (3.). In this report, COVID-19 cases in the United States that occurred during February...

Show less

Relevant    Not relevant

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

[COVID-19Article](#) / Train Watson

Watson will learn which are the best results for your queries after you've rated enough.

Add more queries    Rate more results    Add more variety to your ratings

[Back to queries](#)

15. Click on the **Rate Results** button next to Which age group in the United States has suffered the highest percentage of severe outcomes?

**Queries (4)**

Train Watson by adding natural language queries and rating the results. [Learn more.](#)

+ Add recent queries from Watson Discovery to COVID-19Article  
+ Add a natural language query

|  |  |
|--|--|
| How can we protect older adults from COVID-19?   | <input type="button"/> Rate results   3 relevant 49 not relevant |
| What was the percentage of fatalities among people that are less than 19 years old?          | <input type="button"/> Rate results   4 relevant 48 not relevant |
| When was COVID-19 declared as a pandemic?  | <input type="button"/> Rate results   1 relevant 51 not relevant |
| Which age group in the United States has suffered the highest percentage of severe outcomes? | <input type="button"/> Rate results   Not rated yet              |

16. Review all of the passages for this final query and tag all passages that mention the age group suffering the most from severe outcomes as **Relevant** and anything as **Not relevant**.

Watson will learn which are the best results for your queries after you've rated enough.

[Add more queries](#) [Rate more results](#) [Add more variety to your ratings](#)

[Back to queries](#)

Which age group in the United States has suffered the highest percentage of severe outcomes?

Rate some documents as relevant or not relevant results for this query. [Learn more](#).

**Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) in the United States**

[View document](#)

"... In contrast, persons aged ≤19 years appear to have milder COVID-19 illness, with almost no hospitalizations or deaths reported to date in the United States in this age group. The overall case fatality rate is 0.9%..."

"... Discussion Since February 12, 4,226 COVID-19 cases were reported in the United States; 31% of cases, 45% of hospitalizations, 53% of ICU admissions, and 80% of deaths occurred among adults aged ≥65 years with the highest percentage of severe outcomes among persons aged ≥85 years. These findings are similar to data from China, which indicated >80% of deaths occurred among persons aged ≥60 years..."

Show less

[Relevant](#) [Not relevant](#)

**Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) in the United States**

[View document](#)

"... fatality percentages for reported COVID-19 cases, by age group — United States, February 12–March 16, 2020 Age group (yrs) (no. of cases) %\* Hospitalization ICU admission Case-fatality 0–19 (123) 1.6..."

[Relevant](#) [Not relevant](#)

**Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) in the United States**

[View document](#)

"... Overall, 31% of cases, 45% of hospitalizations, 53% of ICU admissions, and 80% of deaths occurred among adults aged ≥65 years, among whom aged ≥85 years with the highest percentage of severe outcomes among persons aged ≥85 years. In contrast, no ICU admissions or deaths were reported among persons aged ≤19 years. ..."

"... In this report, COVID-19 cases in the United States that occurred during February 12–March 16, 2020 and severity of disease (hospitalization, admission to intensive care unit [ICU], and death) were analyzed by age group. As of March 16, a total of 4,226 COVID-19 cases in the United..."

Show more

[Relevant](#) [Not relevant](#)

< 1 2 3 4 5 >

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

COVID-19Article / Train Watson

Watson will learn which are the best results for your queries after you've rated enough.

[Add more queries](#) [Rate more results](#) [Add more variety to your ratings](#)

[Back to queries](#)

**Queries (4)**

Train Watson by adding natural language queries and rating the results. [Learn more](#).

+ Add recent queries from Watson Discovery to COVID-19Article

+ Add a natural language query

How can we protect older adults from COVID-19? [Rate results](#) 3 relevant 49 not relevant

What was the percentage of fatalities among people that are less than 19 years old? [Rate results](#) 4 relevant 48 not relevant

When was COVID-19 declared as a pandemic? [Rate results](#) 1 relevant 51 not relevant

Which age group in the United States has suffered the highest percentage of severe outcomes? [Rate results](#) 8 relevant 44 not relevant

Now that we have completed Relevancy Training for our natural language queries, we should be able to get more accurate results when we run one of these 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!**