

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

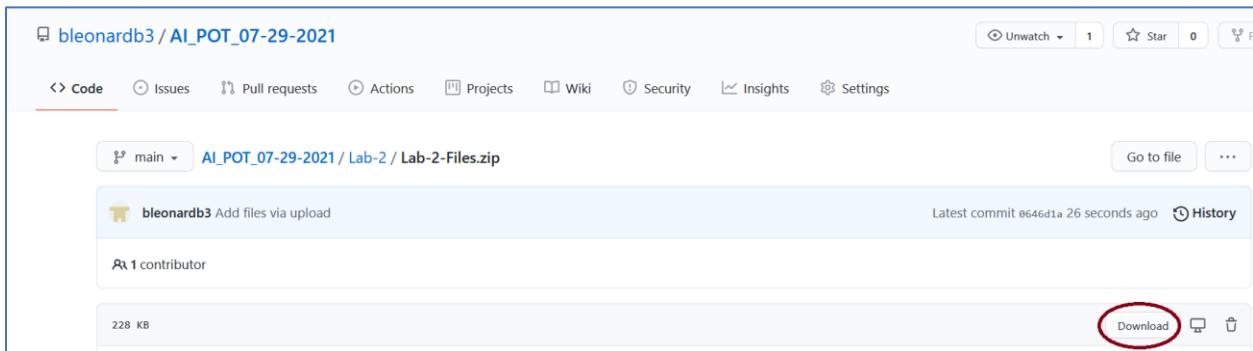
Table of Contents

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

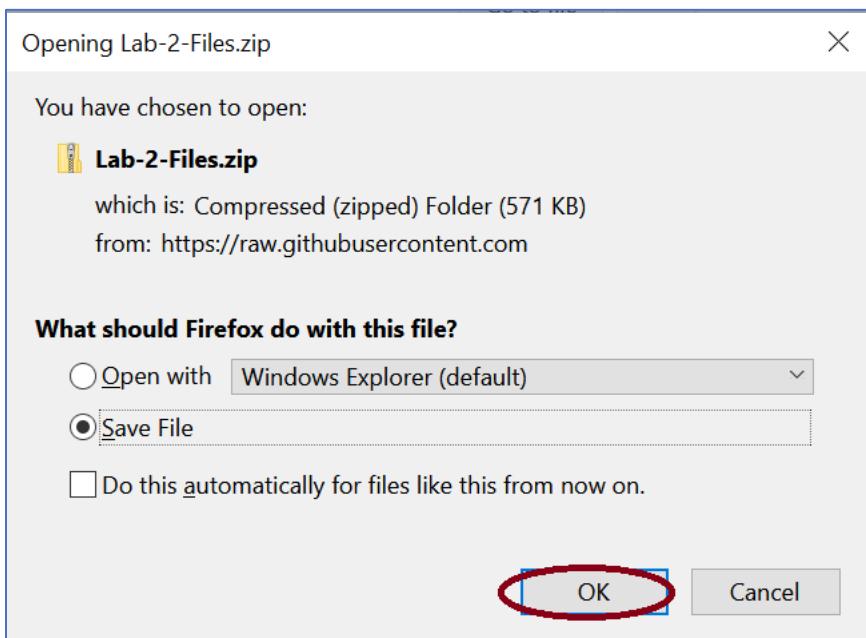
Prerequisites

Download the data files to the Desktop

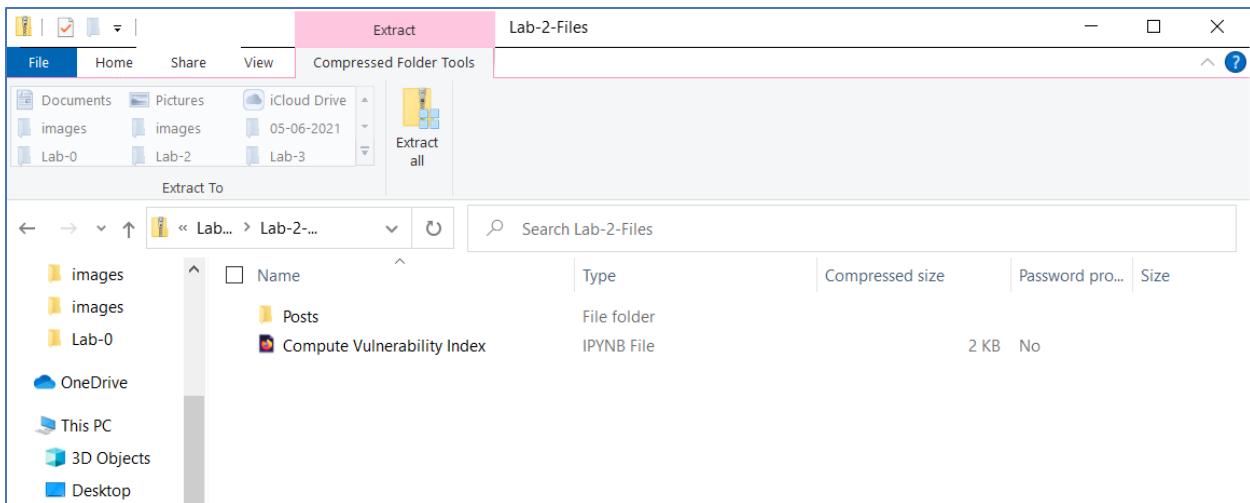
1. Click on https://github.com/bleonardb3/AI_POT_07-29-2021/blob/main/Lab-2/Lab-2-Files.zip
2. Click on the **Download** button.



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

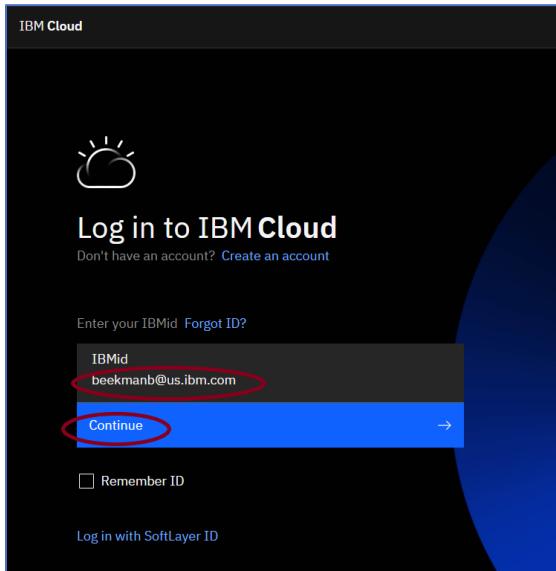
The goal of this lab is to familiarize the user with the Watson Discovery service. Watson Discovery is an enterprise AI search technology that leverages machine learning, including natural language processing (NLP), to retrieve specific answers to your questions and analyze trends and relationships buried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

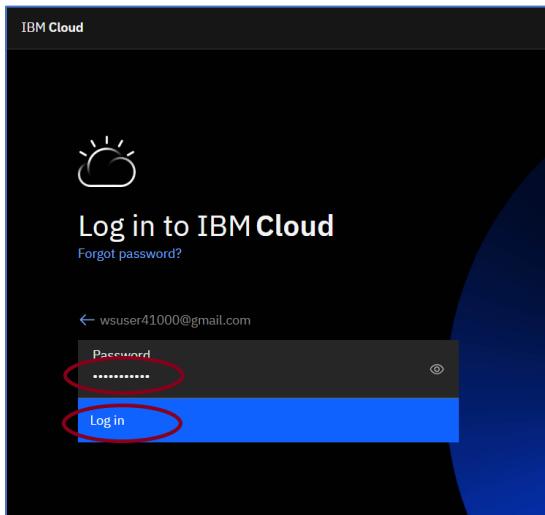
1. Create a Discovery collection
2. Upload the documents
3. Add the entity model from Knowledge Studio
4. Perform custom entity extraction
5. Retrieve the analyzed files using the Discovery API
6. Calculate the COVID-19 vulnerability index
7. Create a collection for a COVID-19 publication
8. Perform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

Exercise 1: Create a Discovery collection

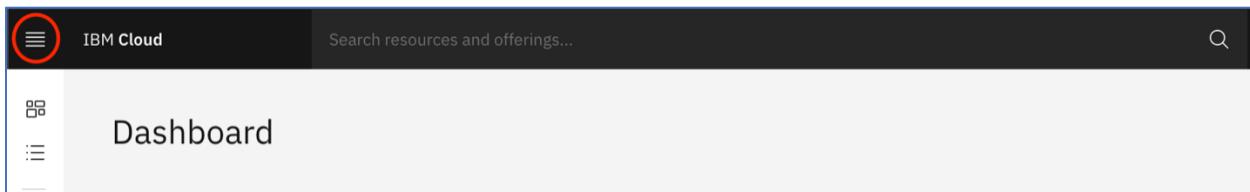
1. Log into your IBM Cloud account by typing **cloud.ibm.com** into the URL address bar of your Firefox or Chrome browser.
2. If you have been logged out, enter your **IBMid** and click **Continue**.



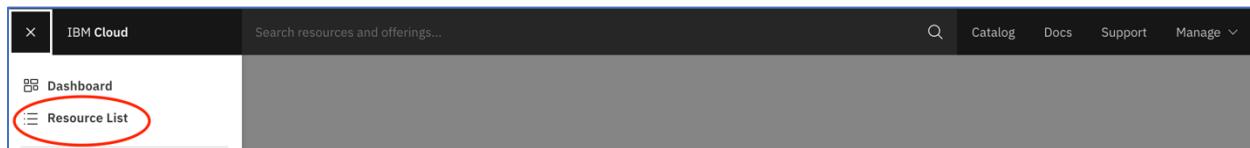
3. Enter your **Password** and click **Log in**.



4. Click on the icon.

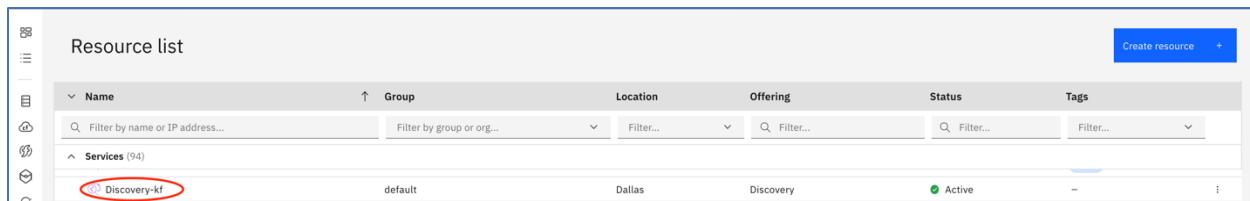


5. Select **Resource List** from the drop-down menu.



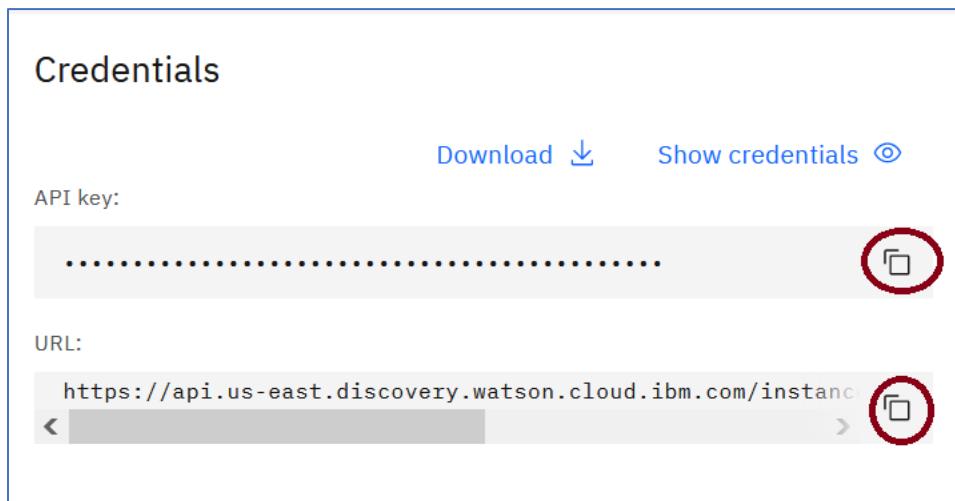
The screenshot shows the IBM Cloud interface. At the top, there's a navigation bar with 'IBM Cloud' and a search bar. Below the navigation bar is a sidebar with two main options: 'Dashboard' and 'Resource List'. The 'Resource List' option is circled in red. The main content area is titled 'Resource list' and displays a table of services. The table has columns for Name, Group, Location, Offering, Status, and Tags. A blue 'Create resource' button is located at the top right of the table area.

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



This screenshot shows the 'Resource list' interface with the 'Services' section expanded. It lists 94 services, including one named 'Discovery-kf' which is circled in red. The table columns are identical to the previous screenshot: Name, Group, Location, Offering, Status, and Tags. The status for 'Discovery-kf' is shown as 'Active'.

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.



This screenshot shows the 'Credentials' section of the Manage screen. It displays two pieces of sensitive information: an 'API key' and a 'URL'. Both items have a copy icon (with a square icon) adjacent to them, which are both circled in red. Below each item is a 'Download' link and a 'Show credentials' link.

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.

The screenshot shows the 'Manage' section of the Watson Discovery interface. On the left, there's a sidebar with 'Manage' selected, followed by 'Getting started', 'Service credentials', 'Plan', and 'Connections'. The main area has a heading 'Start by launching the tool' and three buttons: 'Launch Watson Discovery' (circled in red), 'Getting started tutorial', and 'API reference'. To the right, there's a 'Plan' section showing 'Lite' and a 'Upgrade' button. At the bottom, there's a 'Credentials' section with fields for 'API key' and 'URL', along with download and show credentials links. A 'FEEDBACK' button is on the far right.

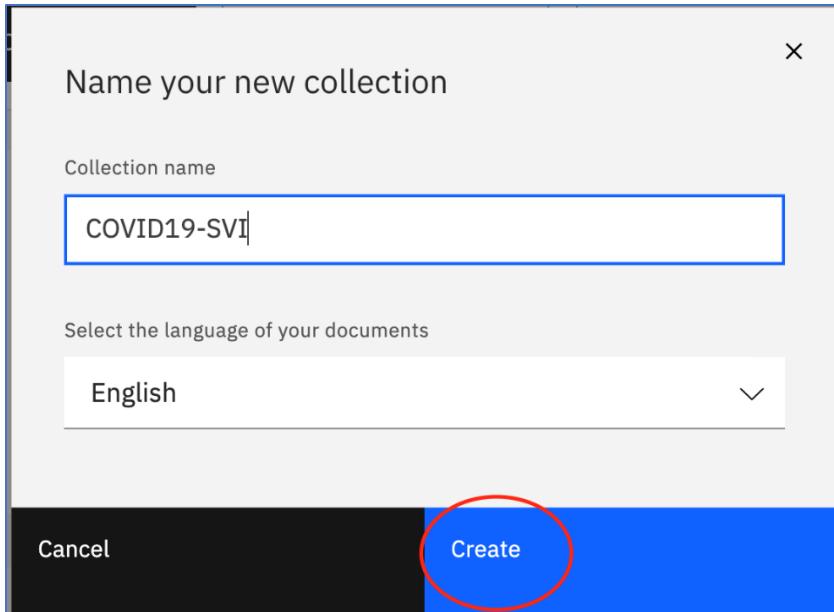
9. On the Manage data screen, click **Upload your own data** to create a new collection.

The screenshot shows the 'Manage data' screen. It features a sidebar with icons for 'Manage data', 'Create a new data collection', 'Create COVID-19 Kit', 'Upload your own data' (circled in red), and 'Connect a data source'. The main content area displays a message about managing private data collections and provides links for creating a COVID-19 kit or uploading own data.

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

The screenshot shows a dialog box titled 'Set up Discovery for your private data.' It states that the 'Lite plan' is suitable for trial & experimentation. It includes a section for 'Building a production application?' with information about the Advanced plan and a link to upgrade. At the bottom, there are 'Cancel' and 'Set up with current plan' buttons.

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.

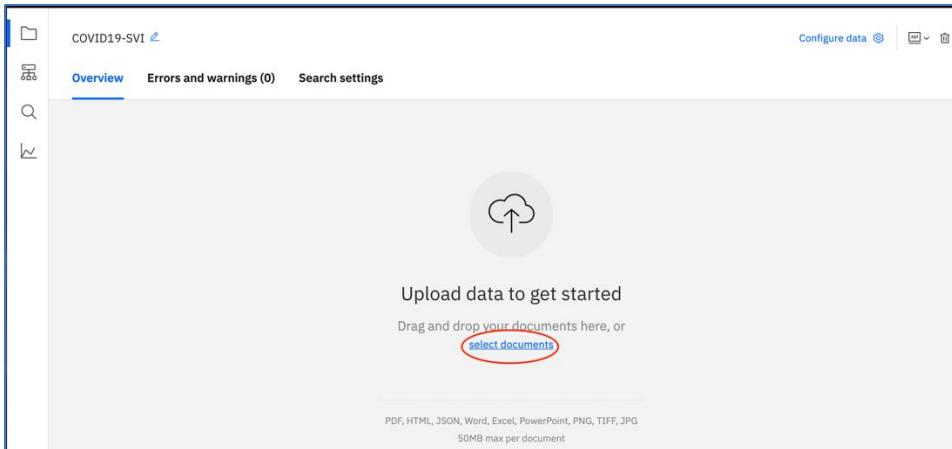
Sentiment analysis

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.

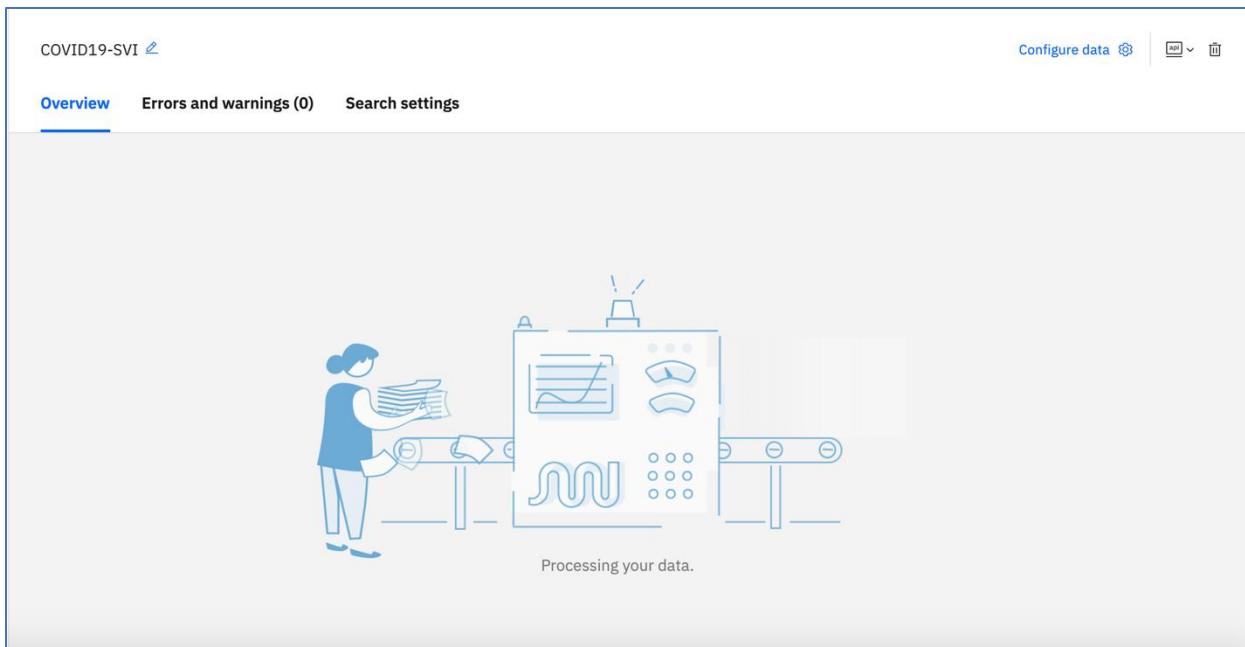
1. Click **select documents**.



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:

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	<i>Current Version</i>		0.65 (0.69 / 0.62)	N/A		Create Version
1.0	05/18/2020		0.65 (0.69 / 0.62)	N/A	368docs-85-10-5	Promote Delete Deploy
Deployed Models (1) Model ID: 63d1efc3-6d00-4273-a034-7034a996c8f0 Service ID: 03b54347-0aad-4da9-b59a-e1f2df1070cc						Undeploy Status

Exercise 3: Add the entity model from Knowledge Studio

1. Click **Configure Data**.

The screenshot shows the 'COVID19-SVI' workspace overview. It includes sections for 'Overview', 'Errors and warnings (0)', and 'Search settings'. On the left, there are icons for folder, document, search, and refresh. In the center, it shows '364 documents' and '0 documents failed'. On the right, there are buttons for 'Upload documents' and 'Configure data' (which is circled in red).

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.](#)

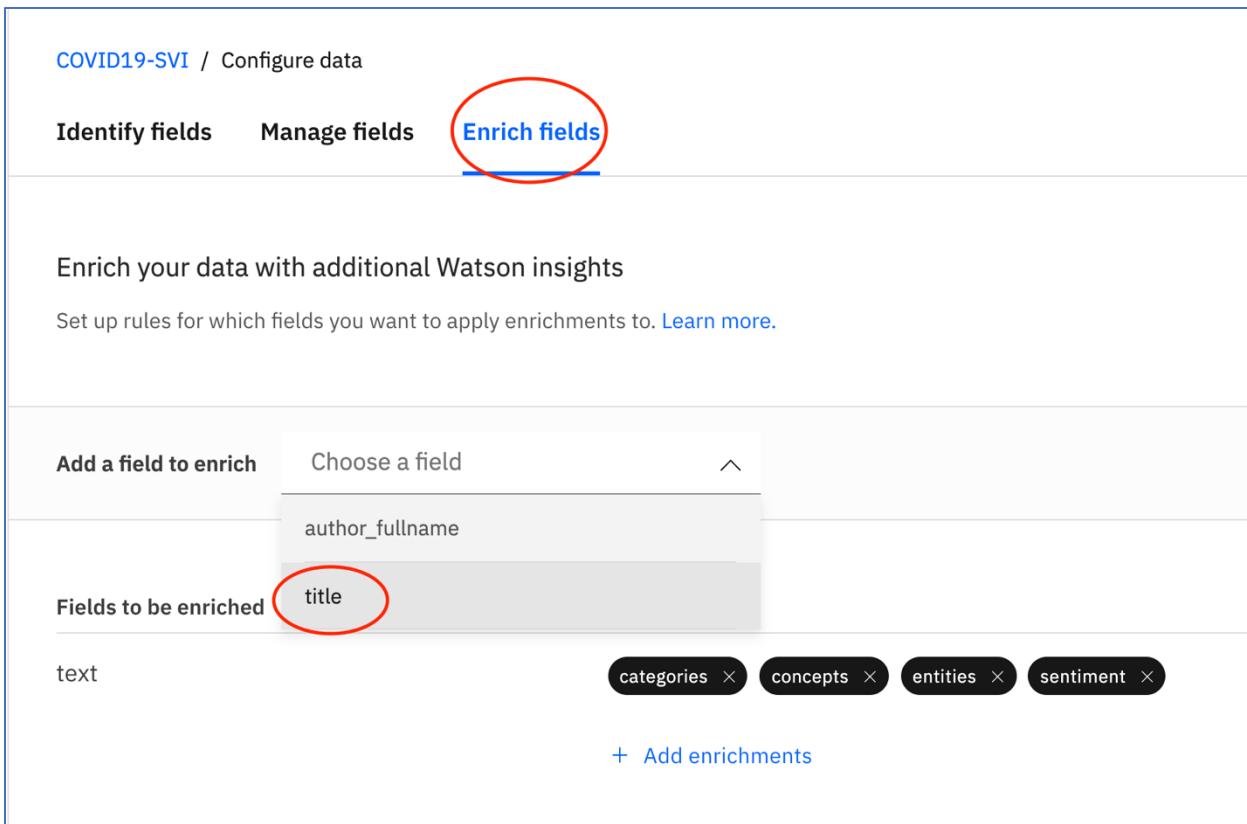
Add a field to enrich Choose a field ^

author_fullname

Fields to be enriched title

text categories × concepts × entities × sentiment ×

+ Add enrichments



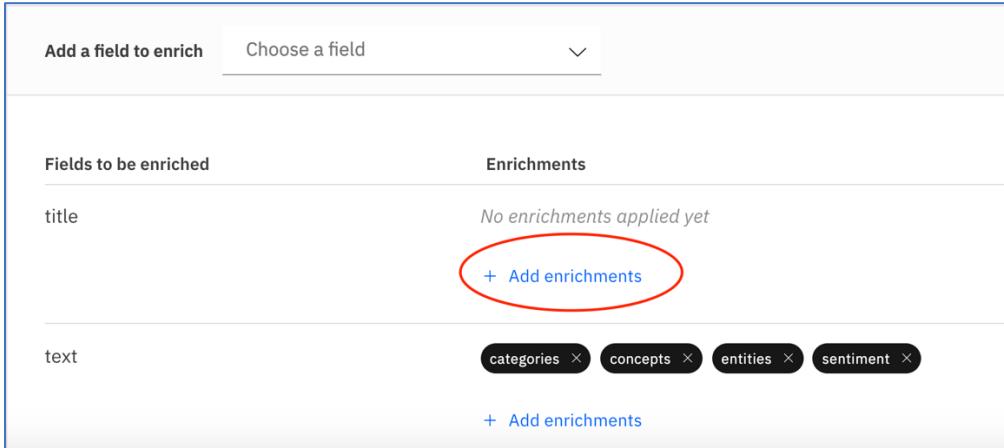
3. To the title field of each post, we will apply entity extraction. Click **+Add enrichments** on the title row.

Add a field to enrich Choose a field ^

Fields to be enriched

Fields to be enriched	Enrichments
title	No enrichments applied yet + 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

title:

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

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

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

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

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

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

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.

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.

Custom Model ID ⓘ
enter 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

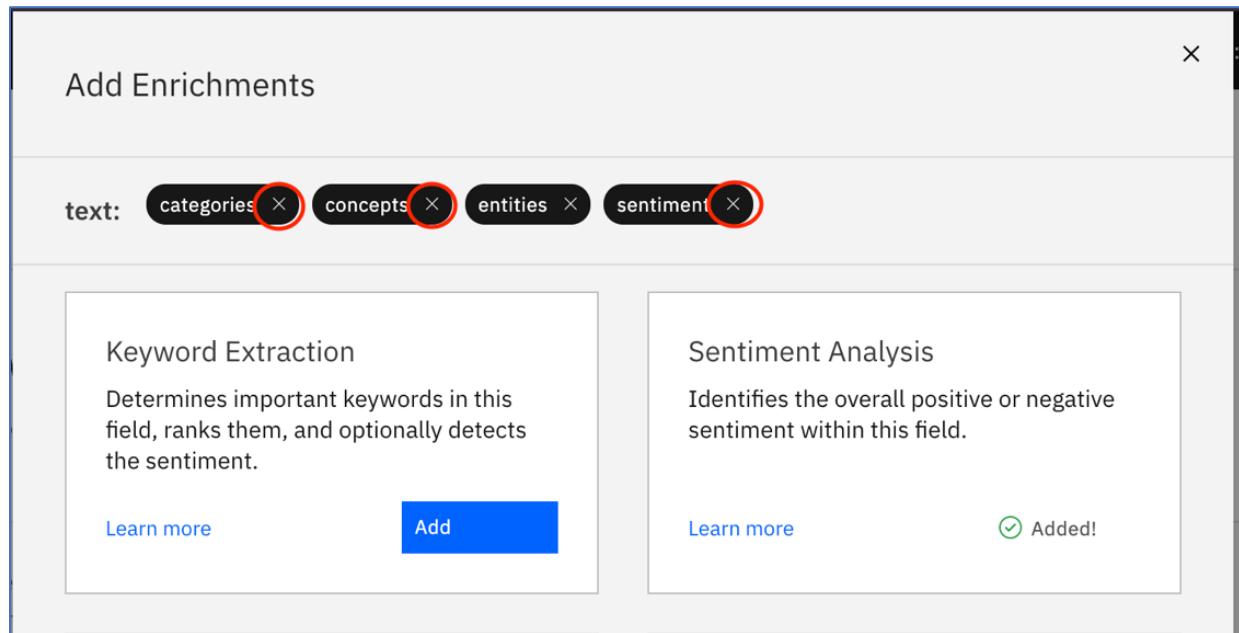
Custom Model ID ⓘ
 Apply

Custom Model ID ⓘ
 Apply

6. 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	entities X
text	categories X concepts X entities X sentiment X + Add enrichments

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 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 <input checked="" type="checkbox"/> Added! Custom Model ID <input type="text" value="63d1efc3-6d00-4273-a034-"/> <input type="button" value="Apply"/>	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

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

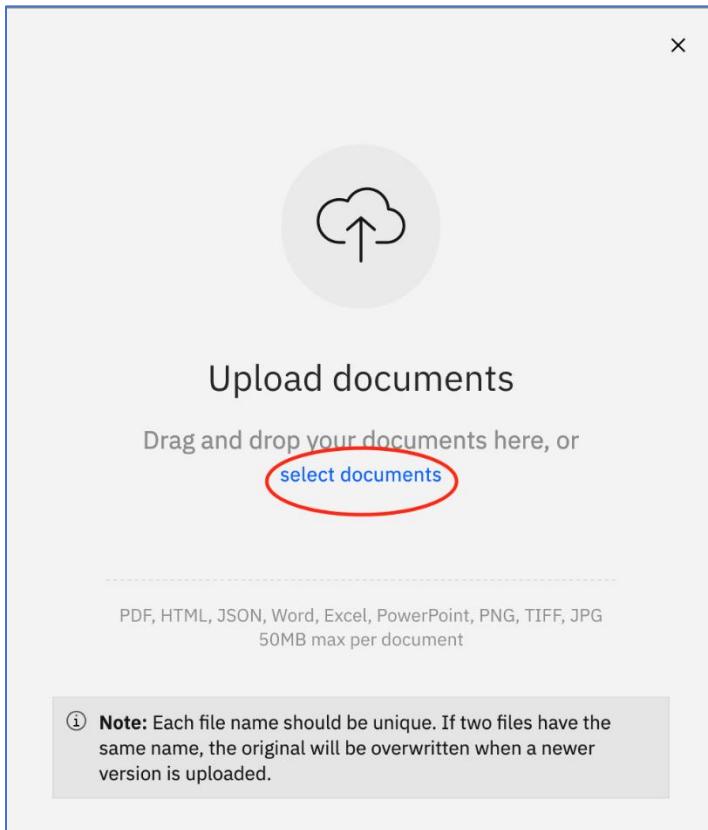
Identify fields Manage fields **Enrich fields**

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
	+ Add enrichments
text	entities
	+ Add enrichments

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' tab is selected, displaying 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 this, the 'Identified 5 fields from your data' section lists:

- text
- author_fullname
- extracted_metadata
- id
- title

The 'Added 1 enrichment to your data' section shows:

- Entity Extraction**
- unemployment (61) | rent (27) | apartment (26) | lease (17) | apartments (10)

A note says: **8 enrichments available.** [Add enrichments](#)

The right sidebar contains three query examples:

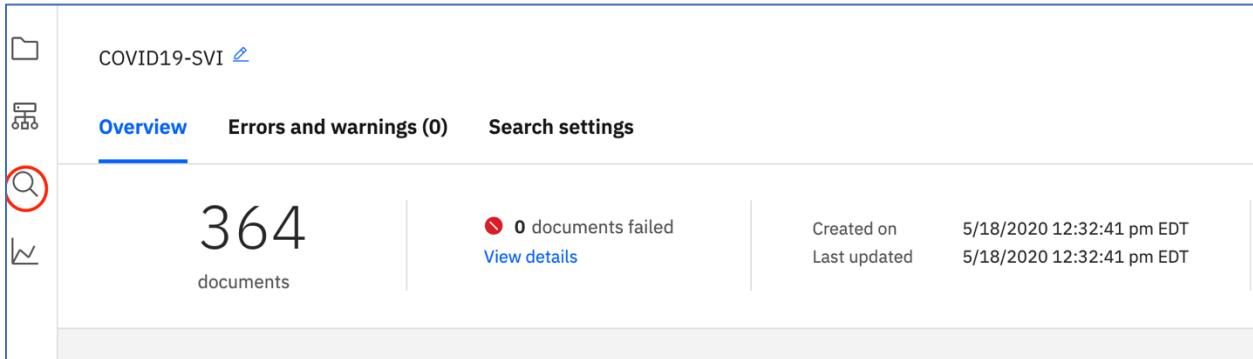
- Top people related to /society/work/unemployment** [Run](#)
- Entities of type Unemployed which have negative sentiment** [Run](#)
- Top entities with their average, min, max sentiment score** [Run](#)

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

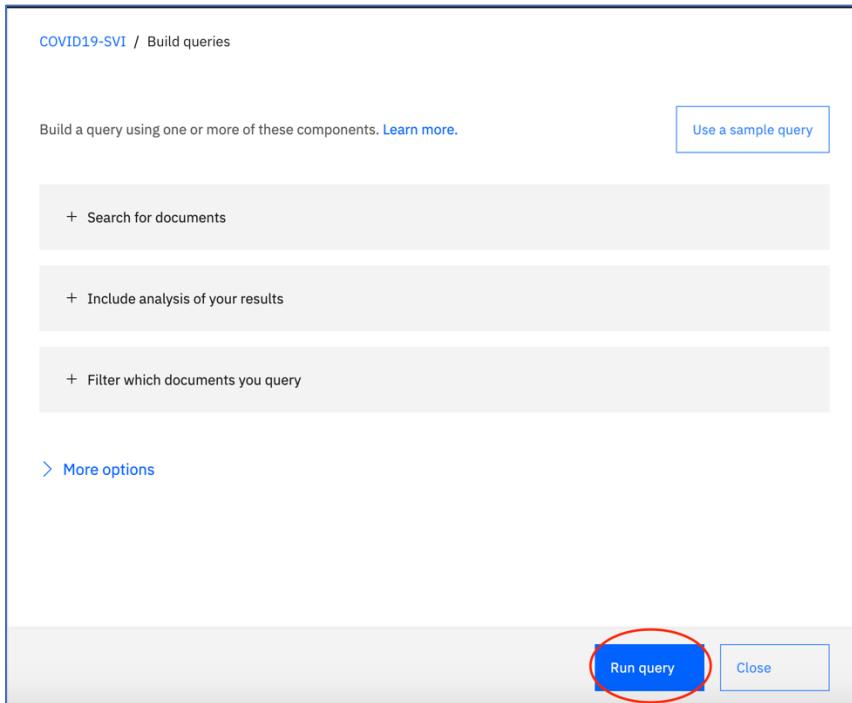
Let's start by viewing the entity extraction output inside the Discovery UI.

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



The screenshot shows the 'COVID19-SVI' collection overview. On the left, there are four icons: a folder, a document, a magnifying glass (circled in red), and a list. The main area displays '364 documents'. Below this, it says '0 documents failed' with a 'View details' link. On the 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 under 'COVID19-SVI'. It has a sidebar with options: '+ Search for documents', '+ Include analysis of your results', '+ Filter which documents you query', and a 'More options' link. At the bottom, there are 'Run query' and 'Close' buttons. The 'Run query' button is circled in red.

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 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.

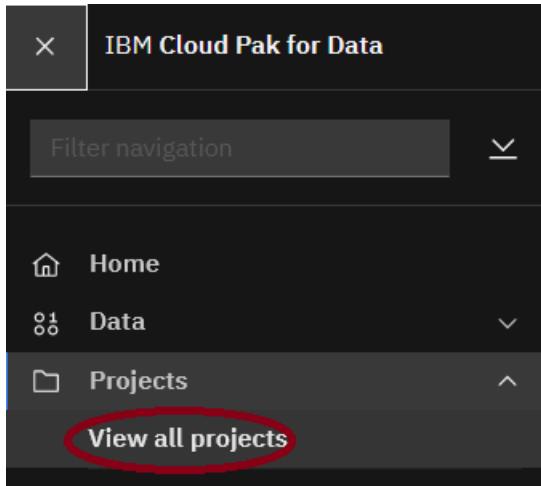
- Click on the **IBM Cloud Pak for Data** tab to navigate to Watson Studio. If you had closed this tab, you can enter <http://dataplatform.cloud.ibm.com> in the browser to navigate to Watson Studio.



- 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 top navigation bar shows "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 displays a table of projects. The first row shows columns for "Name", "Role", "Storage", "Collaborators", and "Creator". The "Name" column lists "Watson Studio Labs", which is circled in red. The "Role" column 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 top navigation bar shows "Projects / Watson Studio Labs". Below it is a horizontal menu with tabs: "Overview", "Assets" (which is highlighted with a blue box and circled in red), "Environments", "Jobs", "Access Control", and "Settings".

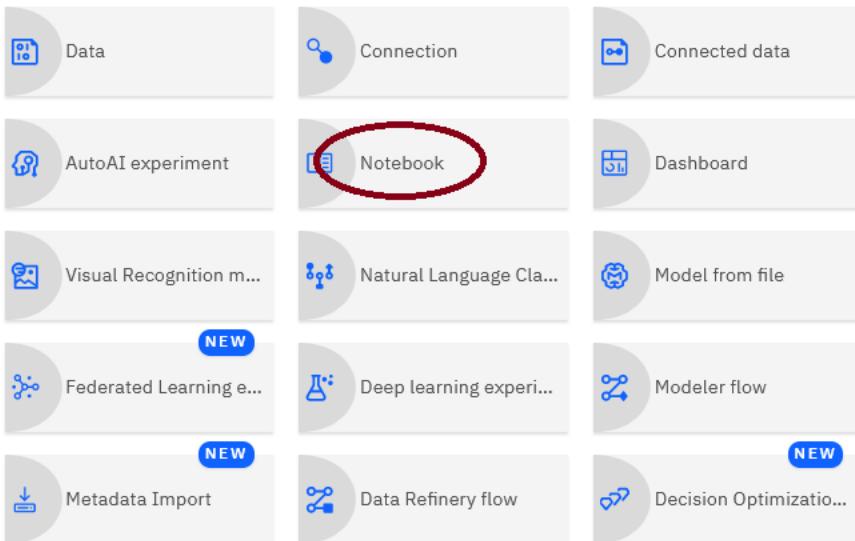
9. Click on Add to project.

A screenshot of the "Assets" tab for the "Watson Studio Labs" project. The top navigation bar shows "IBM Cloud Pak for Data" and "All". There is a search bar and an "Upgrade" button. The main area shows a horizontal menu with tabs: "Overview", "Assets" (which is highlighted with a blue box and circled in red), "Environments", "Jobs", "Access Control", and "Settings". In the top right corner, there is a "Launch IDE" button and an "Add to project" button with a plus sign, which is also circled in red.

10. Click Notebook

Choose asset type

Available asset types



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

New notebook

Blank **From file** From URL

Name
Type notebook name here

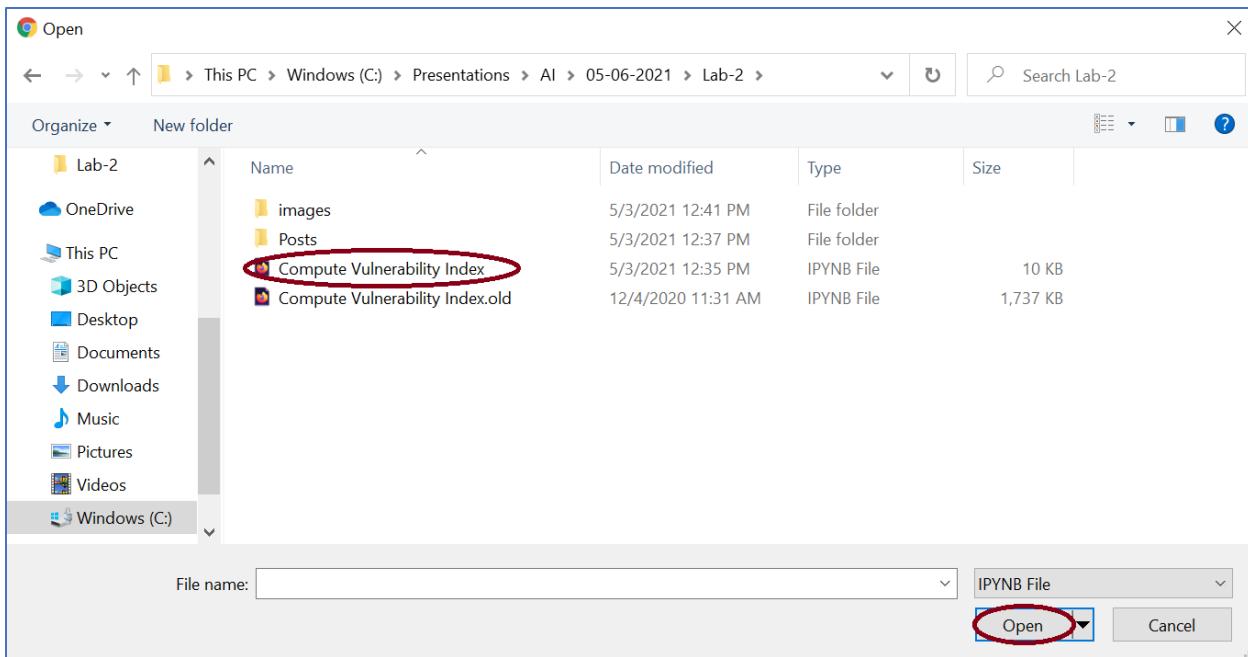
Description (optional)
Type your description here

Select runtime
Default Python 3.7 XS (2 vCPU 8 GB RAM)

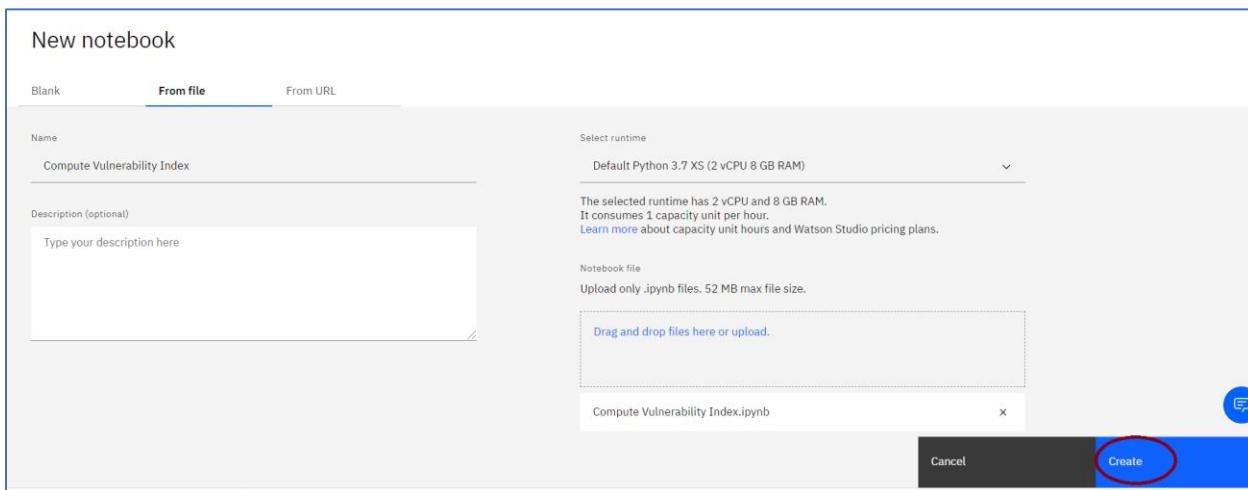
Notebook file
Upload only .ipynb files. 52 MB max file size.
drag and drop files here or upload

Cancel Create

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.

```
In [ ]: authenticator = IAMAuthenticator('X-Yabvwoipsj2rK-rO3kmpE12tB3m56DvaH48ErBqzOA')

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

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

```
In [ ]: authenticator = IAMAuthenticator('X-Yabvwoipsj2rK-rO3kmpE12tB3m56DvaH48ErBqzOA')

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

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

```
In [ ]: authenticator = IAMAuthenticator('X-Yabvwoipsj2rK-rO3kmpE12tB3m56DvaH48ErBqzOA')

discovery = DiscoveryV1(
    version='2019-04-30',
    authenticator=authenticator)
discovery.set_service_url('https://api.us-east.discovery.watson.cloud.ibm.com/instances/fec4b57b-660e-4bdc-b110-610b19fc2067')
```

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.

Please note: This report has been corrected.

Morbidity and Mortality Weekly Report

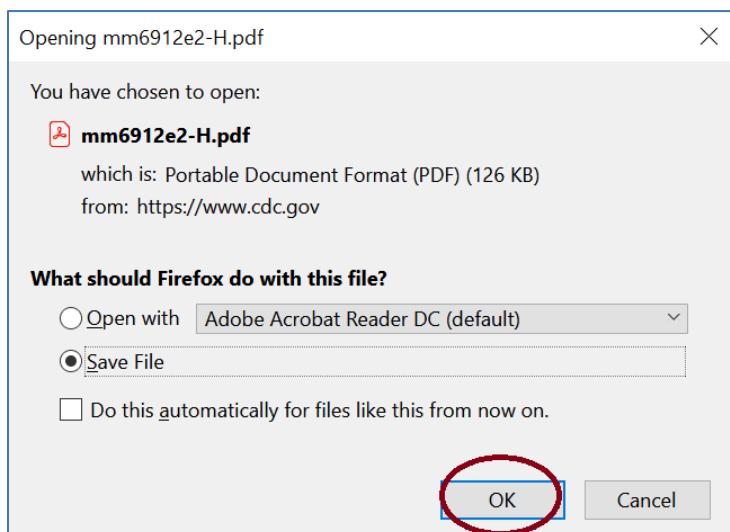
Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020

CDC COVID-19 Response Team

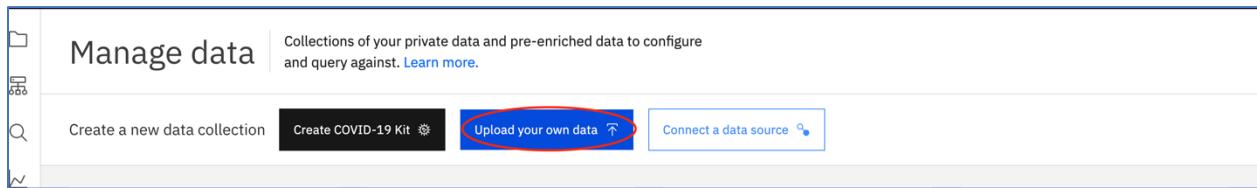
On March 18, 2020, this report was posted as an MMWR Early Release on the MMWR website (<https://www.cdc.gov/mmwr>).

Globally, approximately 170,000 confirmed cases of coronavirus disease 2019 (COVID-19) caused by the 2019 novel coronavirus (SARS-CoV-2) have been reported, including an interest, including hospitalization status (1,514), ICU admission (2,253), death (2,001), and age (386). Because of these missing data, the percentages of hospitalizations, ICU admissions, and deaths (case-fatality percentages) were estimated as a range. The lower bound of these percentages was estimated

3. Click OK.

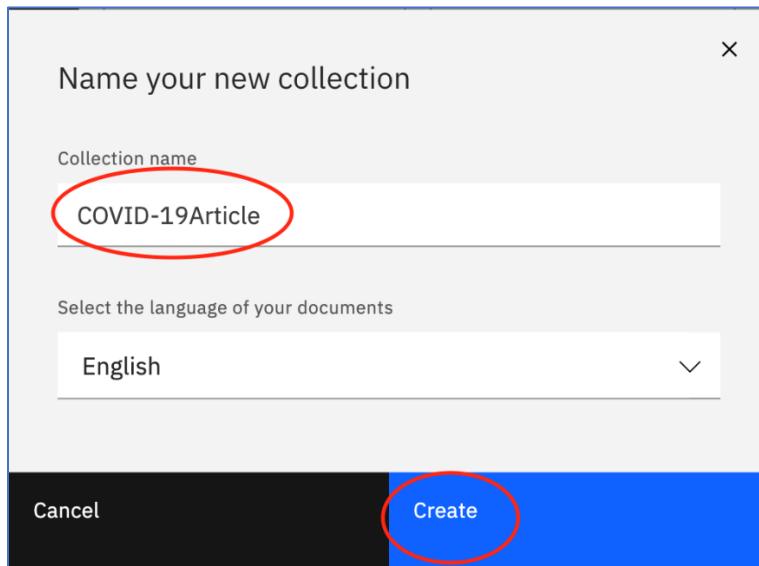


4. Open up your Discovery instance (you may have to log in again) and click **Upload your own data**. If you are not at the right panel, click on the Document  icon.



The screenshot shows the 'Manage data' interface. At the top, there's a header with a folder icon, the text 'Manage data', and a description: 'Collections of your private data and pre-enriched data to configure and query against. [Learn more](#)'. Below the header are three buttons: 'Create a new data collection', 'Create COVID-19 Kit', and 'Upload your own data' (which is circled in red). There's also a 'Connect a data source' button. The interface has a clean, modern design with a light blue header and white background.

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



The screenshot shows a modal dialog box titled 'Name your new collection'. It contains a 'Collection name' input field where 'COVID-19Article' is typed and circled in red. Below it is a 'Select the language of your documents' dropdown menu set to 'English'. At the bottom of the dialog are two buttons: 'Cancel' on the left and a large blue 'Create' button on the right, which is also circled in red.

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

COVID-19Article

Overview Errors and warnings (0) Search settings

Configure data

Upload data to get started

Drag and drop your documents here, or **select documents**

PDF, HTML, JSON, Word, Excel, PowerPoint, PNG, TIFF, JPG
50MB max per document

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

Search Downloads

Name	Date modified	Type	Size
S3 Screenshots	7/20/2021 3:23 PM	File	3
Optimizationv07-29-2021	7/22/2021 1:35 PM	Adobe Acrobat D...	3,7
Optimization Lab Instructions 1.2	7/21/2021 1:35 PM	Adobe Acrobat D...	3,7
Optimization Lab Instructions 1.2	7/20/2021 6:01 PM	Microsoft Word D...	11,8
mm6912e2-H	7/24/2021 5:27 PM	Adobe Acrobat D...	1
COVIDStats	7/22/2021 7:47 PM	CSV File	5,1
COVIDStats	7/22/2021 7:42 PM	File	5,8
COVIDStatistics	7/22/2021 7:12 PM	CSV File	6,0
CognosAnalyticsv07-29-2021	7/23/2021 9:53 AM	Adobe Acrobat D...	2,4
Cognos Slides	7/19/2021 2:34 PM	Microsoft PowerP...	30,8
Cognos Lab 2.0	7/20/2021 6:01 PM	Microsoft Word D...	5,0

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

COVID-19Article

Overview Errors and warnings (0) Search settings

Configure data

Processing your data.

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

COVID19 Article

Overview Errors and warnings (0) Search settings

1 document

0 documents failed View details

Created on Last updated 5/4/2021 10:54:08 am EDT 5/4/2021 10:54:08 am EDT

Upload documents

Identified 1 field from your data

Added 4 enrichments to your data

Now you're ready to query!

Entity Extraction: 19 years | 44 years | 64 years | 84 years | 85 years

Sentiment Analysis: 0% positive, 0% neutral, 100% negative

Concept Tagging: Ageing | Death | Health care | Medicine | Old age

Category Classification: health and fitness → disease

Top entities with their average, min, max sentiment score: Run

Entities of type JobTitle which have negative sentiment: Run

Documents about Washington as a Location with a very negative

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.

COVID-19Article

Overview Errors and warnings (0) Search settings

1 document

0 documents failed View details

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

Upload documents

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.

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

- 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.

The labels below.

- answer
- author
- footer
- header
- question
- subtitle
- table_of_contents
- text
- title
- image
- table

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.
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.

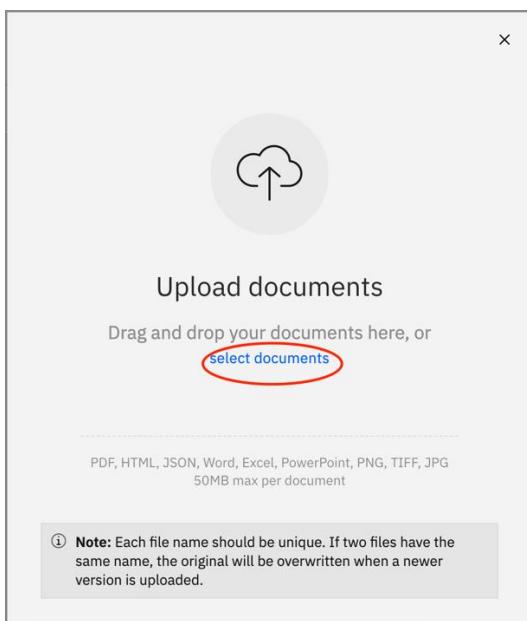
- answer
- author
- footer
- header
- question
- subtitle
- table_of_contents
- text
- title
- image
- table

Submit page

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

The screenshot shows a document review interface. At the top right, there is a button labeled "Apply changes to collection" with a checkmark icon, which is circled in red. Below this, there are navigation controls (back, forward, search, etc.) and a "Field labels" section. The "Field labels" section includes a "+ Create new" button and a dropdown menu with the option "answer".

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**.

The screenshot shows a Windows File Explorer window displaying the "Downloads" folder. The left sidebar shows standard folder icons. The main area lists files with columns for Name, Date modified, Type, and Size. A red circle highlights the file "mm6912e2-H.pdf" in the list. Other visible files include "Optimizationv07-29-2021", "Optimization Lab Instructions 1.2", "COVIDStats", "COVIDStatistics", "CognosAnalyticsv07-29-2021", "Cognos Slides", and "Cognos Lab 2.0".

Name	Date modified	Type	Size
S3 Screenshots	7/20/2021 3:23 PM	File	3
Optimizationv07-29-2021	7/22/2021 1:35 PM	Adobe Acrobat D...	3,7
Optimization Lab Instructions 1.2	7/21/2021 1:35 PM	Adobe Acrobat D...	3,7
Optimization Lab Instructions 1.2	7/20/2021 6:01 PM	Microsoft Word D...	11,8
mm6912e2-H.pdf	7/24/2021 5:27 PM	Adobe Acrobat D...	1
COVIDStats	7/22/2021 7:47 PM	CSV File	5,1
COVIDStatistics	7/22/2021 7:42 PM	File	5,8
CognosAnalyticsv07-29-2021	7/23/2021 9:53 AM	Adobe Acrobat D...	2,4
Cognos Slides	7/19/2021 2:34 PM	Microsoft PowerP...	30,8
Cognos Lab 2.0	7/20/2021 6:01 PM	Microsoft Word D...	5,0

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

The screenshot shows the Watson Studio interface for a collection named 'COVID-19Article'. On the left, there's a sidebar with icons for folder, file, search, and refresh. The main area has a title bar with 'COVID-19Article' and a 'Configure data' button with a red circle around it. Below the title bar, there are tabs for 'Overview', 'Errors and warnings (0)', and 'Search settings'. The 'Overview' tab is selected. It displays the following information: '1 document', '0 documents failed' (with a 'View details' link), 'Created on 5/24/2020 10:15:40 pm EDT', 'Last updated 5/24/2020 10:15:40 pm EDT', and a blue 'Upload documents' button. At the bottom, it says 'Identified 1 field from your data', 'Added 4 enrichments to your data', and 'Now you're ready to query!'. A red box highlights the 'Configure data' button.

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**.

This screenshot is identical to the one above, showing the 'COVID-19Article' collection overview. The 'Configure data' button in the top right corner is specifically highlighted with a red oval.

13. Click on **Manage fields**.

The screenshot shows the 'Manage fields' section of the collection configuration. At the top, there are three tabs: 'Identify fields', 'Manage fields' (which is highlighted with a red circle), and 'Enrich fields'. Below these tabs, there are two main sections: 'Identify fields to index' and 'Improve query results by splitting your documents'. Under 'Identify fields to index', there is a list of fields (answer, author, caption, footer, graphTitle, header, image) each with an 'On' toggle switch. Under 'Improve query results by splitting your documents', there is a button labeled '+ 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**.

The screenshot shows two sections side-by-side. On the left, under 'Identify fields to index', there is a table with three rows: 'answer' (On), 'author' (On), and 'caption' (On). On the right, under 'Improve query results by splitting your documents', there is a note about splitting documents into segments based on fields, followed by a button labeled '+ Split document' which is circled in red.

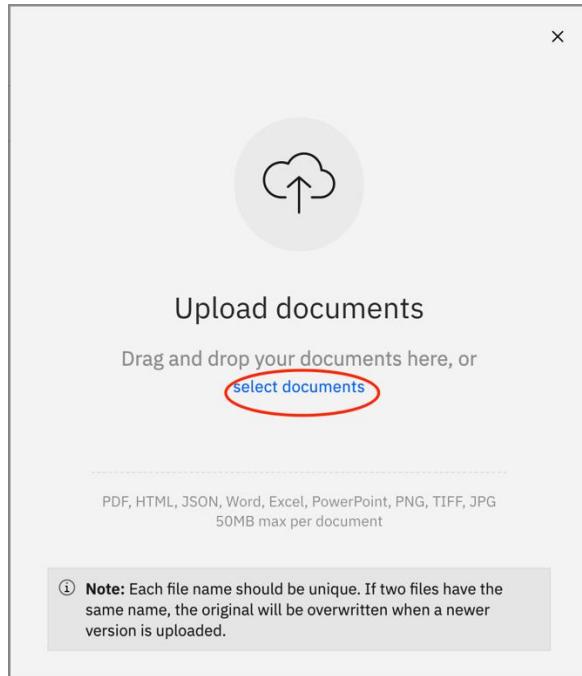
15. Click **Select field** and choose **text**.

The screenshot shows a dropdown menu titled 'Split document on each occurrence of'. The 'Select field' dropdown is open, showing several options: 'subtitle', 'table', 'table_of_contents', and 'text'. The 'text' option is highlighted and circled in red.

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

The screenshot shows a blue button labeled 'Apply changes to collection' with a circular arrow icon, which is circled in red. Below the button, there is a note about improving query results by splitting documents and a link to learn more.

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



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

	Name	Date modified	Type	Size
	S3 Screenshots	7/20/2021 3:23 PM	File	3
	Optimizationv07-29-2021	7/22/2021 1:35 PM	Adobe Acrobat D...	3,7
	Optimization Lab Instructions 1.2	7/21/2021 1:35 PM	Adobe Acrobat D...	3,7
	Optimization Lab Instructions 1.2	7/20/2021 6:01 PM	Microsoft Word D...	11,8
	mm6912e2-H	7/24/2021 5:27 PM	Adobe Acrobat D...	1
	COVIDStats	7/22/2021 7:47 PM	CSV File	5,1
	COVIDStats	7/22/2021 7:42 PM	File	5,8
<input checked="" type="checkbox"/>	COVIDStatistics	7/22/2021 7:12 PM	CSV File	6,0
	CognosAnalyticsv07-29-2021	7/23/2021 9:53 AM	Adobe Acrobat D...	2,4
	Cognos Slides	7/19/2021 2:34 PM	Microsoft PowerP...	30,8
	Cognos Lab 2.0	7/20/2021 6:01 PM	Microsoft Word D...	5,0

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).

The screenshot shows the Watson Discovery Overview page. At the top, there are tabs for 'Overview' (which is selected), 'Errors and warnings (0)', and 'Search settings'. Below this, it displays '51 documents' and '0 documents failed'. It also shows the creation date as '5/4/2021 10:54:08 am EDT' and the last update date as '5/4/2021 10:54:08 am EDT'. There is a 'Upload documents' button with a file icon.

On the left, it says 'Identified 2 fields from your data' with 'text' and 'title' listed. A link 'Need to identify more fields? Add fields' is provided. On the right, it says 'Added 4 enrichments to your data' with sections for Entity Extraction, Sentiment Analysis, Concept Tagging, and Category Classification. Entity Extraction shows CDC (13) | United States (9) | 19 years (7) | US Department of Health and H... (6). Sentiment Analysis shows 4% positive, 68% neutral, and 28% negative. Concept Tagging shows Epidemiology (11) | Medicine (11) | United States (8) | Death (7) | Ageing (6). Category Classification shows health and fitness → disease → epidemic. A note says 'Now you're ready to query!' with three examples: '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. The magnifying glass icon in the sidebar is circled in red. The main area shows 'COVID-19Article' (with a link), 'Overview' (selected), 'Errors and warnings (0)', 'Search settings', '52 documents', and '0 documents failed'.

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**.

The screenshot shows the Watson Discovery interface under the 'Build queries' section for a COVID-19 article. It displays a list of pre-built query components:

- + Search for documents
- + Include analysis of your results
- + Filter which documents you query
- > More options

On the right, a list of sample queries is shown, with the first one, "Most common entity types and their top entities", circled in red. Other queries include:

- Entities of type JobTitle which have negative sentiment
- Top entities with their average, min, max sentiment score
- Top people related to /health and fitness/disease
- 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

At the bottom are "Run query" and "Close" buttons.

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

The screenshot shows the Watson Discovery results page for the query URL: <https://gateway.watsonplatform.net/discovery/api/v1/environments/9>. The "Summary" tab is active. The results are presented in sections:

- 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**

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

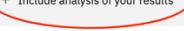
```
nested(enriched_text.entities).term(enriched_text.entities.type,count:5).term(enriched_text.entities.text)
```

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  Select field  10 

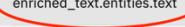
+ Add child aggregations 

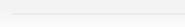
+ Add top-level aggregation 

enriched_text.entities.disambiguation.name

enriched_text.entities.disambiguation.subtype

enriched_text.entities.sentiment.label

enriched_text.entities.text 

+ Filter which documents you query 

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 top-level aggregation

Run query (circled) **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

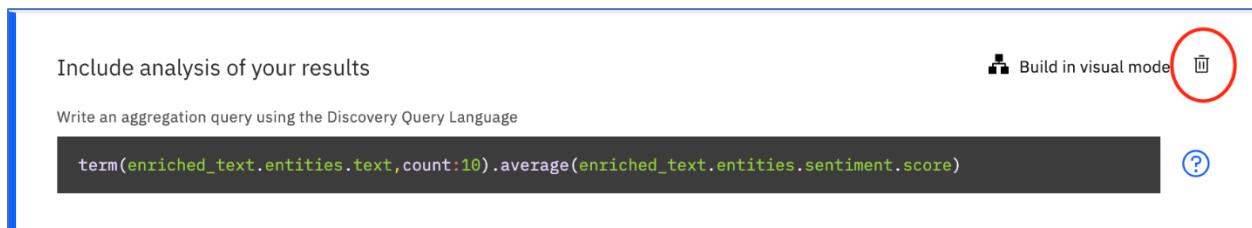
```

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): -0.02798104166666664

```

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.

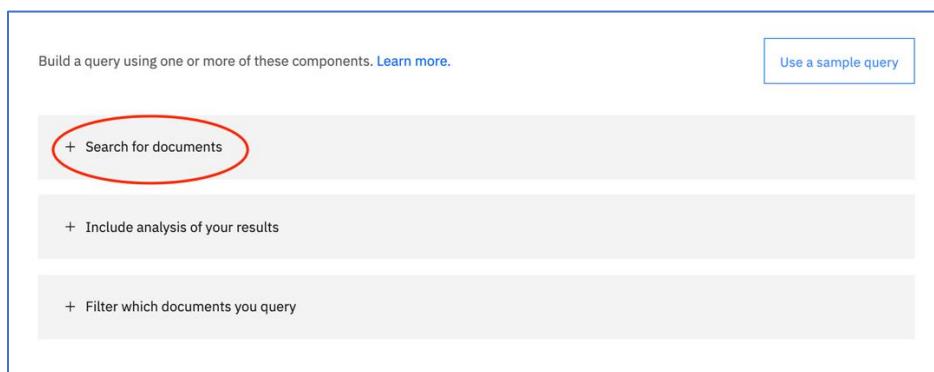


Include analysis of your results

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

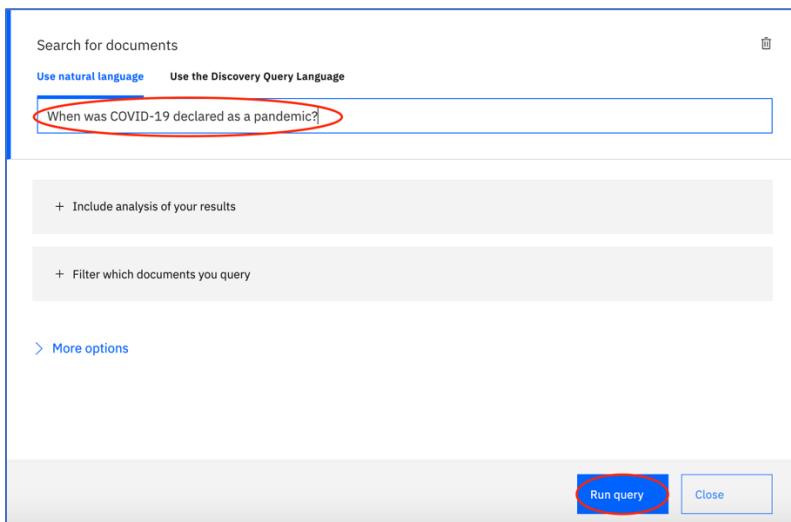
Here we can create natural language queries to ask specific questions about our document. After a quick glance at our article, we can ask the following questions:

- 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 to answer this question.

A screenshot of a search results page titled 'Passages'. It displays five retrieved passages related to COVID-19. Passage 1: "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)." Passage 2: "Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020" Passage 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" Passage 4: "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" Passage 5: "CDC. Coronavirus disease 2019 (COVID-19): if you are at higher risk."

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?

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."

"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 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.

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 a sidebar with icons for folder, file, search, and list. The main area displays instructions: 'Watson will learn which are the best results for your queries after you've rated enough.' Below are three buttons: '+ Add more queries', '+ Rate more results', and '+ Add more variety to your ratings'. Under the 'Queries' section, there is a note: 'Train Watson by adding natural language queries and rating the results. [Learn more.](#)' followed by a red-circled '+ Add recent queries from Watson Discovery to COVID-19Article' button. At the bottom, there is a link '+ Add a natural language 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. One query, 'Which age group in the United States has suffered the highest percent...', is selected with a checked checkbox. A red circle highlights both the selected query 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:

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

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

+ Add a natural language query

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?

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

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.

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?

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

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.

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%–3% among persons aged ≤19 years, to ≥31% among adults aged ≥85 years. (Table)..."

"... Among 508 (12%) patients known to have been hospitalized, 9% were aged ≥85 years, 36% were aged 65–84 years, 17% were aged 55–64 years, 18% were 45–54 years, and 20% were aged 20–44 years. Less than 1% of hospitalizations were among persons aged ≤19 years. (Table)..."

Show more

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). ..."

"... Among 44 cases with known outcome, 15 (34%) deaths were reported among adults aged ≥85 years, 20 (46%) among adults aged 65–84 years, and nine (20%) among adults aged 20–64 years. 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). ..."

Show more

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 65–84 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. ..."

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.

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 55 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, 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!