

IBM Training

Student Exercises

**Lab-1: Develop Socioeconomic
Annotators for COVID-19**

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
Introduction	3
Objectives.....	3
Exercise 1: Create a Watson Knowledge Studio Instance	4
Exercise 2: Create a Watson Discovery Instance.....	6
Exercise 3: Create a Type System	7
Exercise 4: Create a Dictionary	12
Exercise 5: Upload a corpus of documents.....	15
Exercise 6: Perform Manual Annotation.....	17
Exercise 7: Train and create a machine learning (ML) annotator	24
Exercise 8: Save and Deploy the ML Annotator to Discovery	29

Prerequisites

1. Make sure you have an IBM Cloud account before starting these labs.

Please make sure that you have already registered and currently have a working IBM Cloud account before starting these labs. If you do not have an IBM Cloud account, please inform the instructors right away.

2. Download all lab files to Desktop

- Go to the GitHub repository for this workshop:
https://github.com/bleonardb3/AI_POT_06-11-2020
- Click the down arrow on the **Clone or download button** and select **Download ZIP**. This will download all of the folders for this workshop as a zip file to your computer. You will use these folders with the instruction guides in order to complete each lab.
- Extract the downloaded zip file and save it on your Desktop.

Introduction

This lab will cover the development of socioeconomic annotators for COVID-19 in order to create a COVID-19 vulnerability index. IBM Watson Knowledge Studio will be used to develop the socioeconomic annotators.

Objectives

The goal of this lab is to familiarize the user with the Watson Knowledge Studio service. Watson Knowledge Studio lets you build a machine learning annotator by applying a type system, dictionary pre-annotator and human annotation on a training corpus of unstructured documents. Upon training and evaluation, the machine learning annotator can be saved and deployed to Watson Discovery for automated entity extraction. Watson Discovery is an enterprise AI search technology that leverages machine learning, including natural language processing, 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, Watson Discovery can be trained on the language of your domain. Both Watson Knowledge Studio and Watson Discovery can be deployed on any cloud or on-premises environment.

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

1. Provision an instance of Watson Knowledge Studio
2. Provision an instance of Watson Discovery
3. Create a type system
4. Create a dictionary
5. Upload a corpus of documents
6. Perform manual annotation

7. Train and create a machine learning (ML) annotator
8. Save and deploy the ML annotator to Watson Discovery

Exercise 1: Create a Watson Knowledge Studio Instance

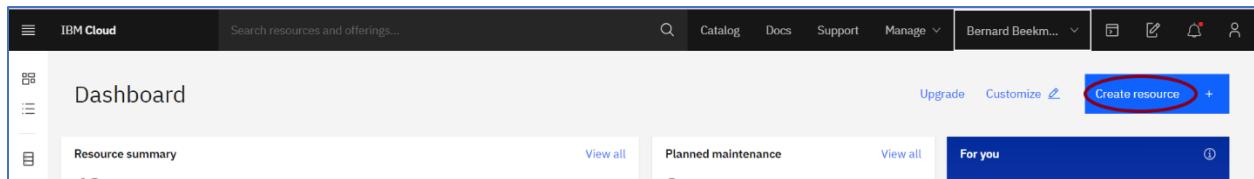
1. Log into your IBM Cloud account by typing **cloud.ibm.com** into the URL address bar of your Firefox or Chrome browser.
2. Enter your **IBMid** and click **Continue**.

The screenshot shows the 'Log in to IBM Cloud' page. At the top, there's a logo and a link to 'Create an account'. Below that, a text input field is labeled 'IBMid' with the value 'beekmanrb@us.ibm.com'. A red circle highlights this input field. Below it is a blue button labeled 'Continue' with a red circle around it. To the right of the 'Continue' button is a right-pointing arrow. At the bottom left, there's a checkbox for 'Remember ID' and a link 'Log in with SoftLayer ID'.

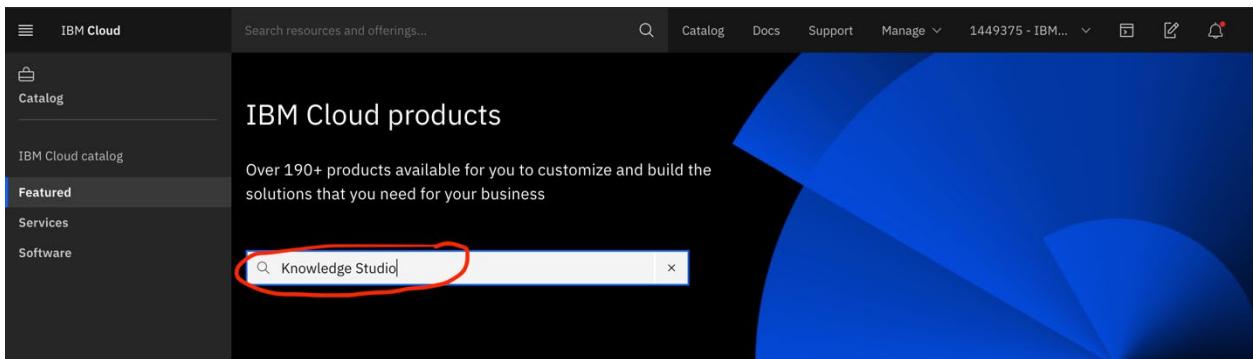
3. Enter your **Password** and click **Log in**.

The screenshot shows the same 'Log in to IBM Cloud' page after the user has entered their IBMID. Now, the password input field is highlighted with a red circle. It contains the value 'wsuser41000@gmail.com'. Below it is a blue 'Log in' button with a red circle around it. To the right of the 'Log in' button is a right-pointing arrow. The rest of the page remains the same, with the 'Forgot password?' link and the 'Remember ID' checkbox.

4. Click Create Resource.



5. Enter Knowledge Studio and click the <Enter> key.



6. Click on Knowledge Studio.

A screenshot of the search results for "Knowledge Studio". The results are titled "Search results for 'Knowledge Studio'" with "1 result". The single result is a card for "Knowledge Studio" by IBM, Services, AI. The card features a small icon, the service name in blue, and the subtitle "Teach Watson the language of your domain.". Below the card, it says "Lite • Free • IAM-enabled". The entire card area is circled in red.

7. Click on the Lite plan and click Create.

The screenshot shows the IBM Cloud Catalog interface for creating a Knowledge Studio instance. On the left, there's a sidebar with 'Create' and 'About' buttons. Below it, a dropdown menu shows 'Dallas' selected. The main area has a heading 'Select a region' and a note about pricing. A table lists three plans: 'Lite', 'Standard', and 'Premium'. The 'Lite' plan is circled in red. The 'Create' button at the bottom right of the table is also circled in red.

Exercise 2: Create a Watson Discovery Instance

1. Enter **Discovery** into the *Search resources and offerings* bar and click on **Discovery** under *Catalog Results*.

The screenshot shows the IBM Cloud Catalog search results for 'Discovery'. The search bar at the top contains 'Discovery'. In the 'Resource Results' section, there are entries for 'Discovery-ew', 'Discovery-x2', and 'Watson-Discovery'. In the 'Catalog Results' section, there are entries for 'API Connect' and 'Discovery'. The 'Discovery' entry under Catalog Results is circled in red.

2. Select the **Lite** plan and click **Create**.

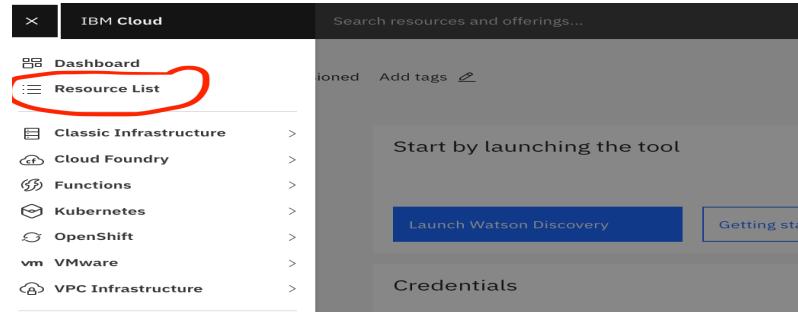
The screenshot shows the Watson Discovery service page. On the left, there's a table comparing three plans: Lite, Advanced, and Premium. The 'Lite' plan is circled in red. The 'Advanced' plan details pricing based on document tiers, ranging from up to 50,000 docs to up to 100M docs. The 'Premium' plan includes everything in Advanced plus usage and training data in an isolated environment, high availability, and HIPAA compliance. On the right, a summary panel shows the service is 'Free', located in the Dallas region, with the plan set to 'Lite'. A 'Create' button is highlighted with a red circle.

Although we will be using this Watson Discovery instance in Lab 2, we need to provision this instance in order to link it to the deployed machine learning annotator, which we will create by the end of this lab. The machine learning annotator will be used by Watson Discovery to perform entity extraction in Lab 2.

Exercise 3: Create a Type System

1. Select the Navigation Menu icon on the top left corner of the screen (the hamburger icon) and click **Resource List** on the drop down menu.

The screenshot shows the IBM Cloud Resource List interface. It displays a single instance named 'Discovery-x2' which is 'Provisioned'. The instance has 'Add tags' and 'Actions...' buttons. On the left, a sidebar shows 'Manage' sections for 'Getting started', 'Service credentials', 'Plan', and 'Connections'. The 'Plan' section shows 'Lite' and an 'Upgrade' button. The main area has a 'Start by launching the tool' section with 'Launch Watson Discovery', 'Getting started tutorial', and 'API reference' buttons.



- Under Services, click on your Knowledge Studio instance (for a new IBM Cloud account, you should only see Knowledge Studio and Discovery listed here).

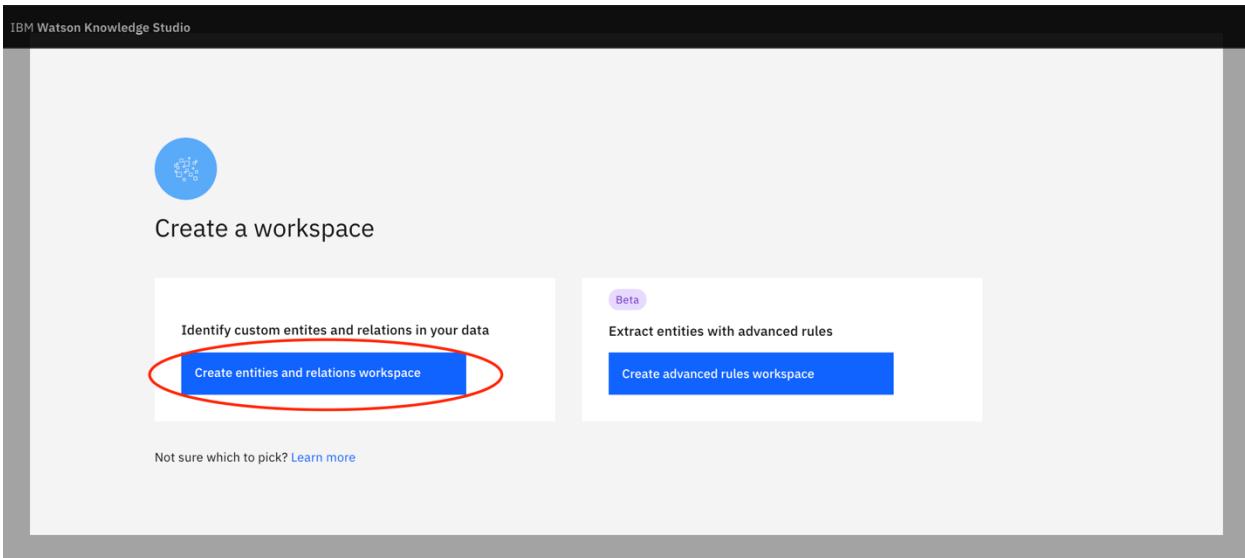
This screenshot shows the 'Resource list' page in the IBM Cloud interface. The left sidebar lists various service categories like Devices, VPC infrastructure, Clusters, Cloud Foundry apps, and Cloud Foundry services. Under 'Cloud Foundry services', there are 13 items listed, including 'Knowledge Catalog-va' and 'Knowledge Studio-2c'. The 'Knowledge Studio-2c' entry is circled in red.

Name	Group	Location	Offering	Status	Tags
AI-OpenScale	default	Dallas	Watson OpenScale	Active	-
App ID-app	default	Dallas	App ID	Active	-
Compare-n-Comply	default	Dallas	Compare and Comply	Active	-
Continuous-Delivery	default	Dallas	Continuous Delivery	Active	-
Db2-app	default	Dallas	Db2	Active	-
IBM Cognos Dashboard Embedded-ix	default	Dallas	IBM Cognos Dashboard Embedded	Active	-
Knowledge Catalog-va	default	Dallas	Knowledge Catalog	Active	-
Knowledge Studio-2c	default	Dallas	Knowledge Studio	Active	-

- Click **Launch Knowledge Studio** to start your instance of Watson Knowledge Studio.

This screenshot shows the details page for the 'Knowledge Studio-2c' instance. It includes sections for 'Manage', 'Getting started', and 'Plan'. The 'Plan' section shows a 'Lite' plan with an 'Upgrade' button. In the center, there is a 'Start by launching the tool' section with a prominent blue 'Launch Watson Knowledge Studio' button, which is circled in red.

- Select **Create entities and relations workspace**.



5. Type **COVID19-Vulnerability** for the Workspace name and click **Create**.

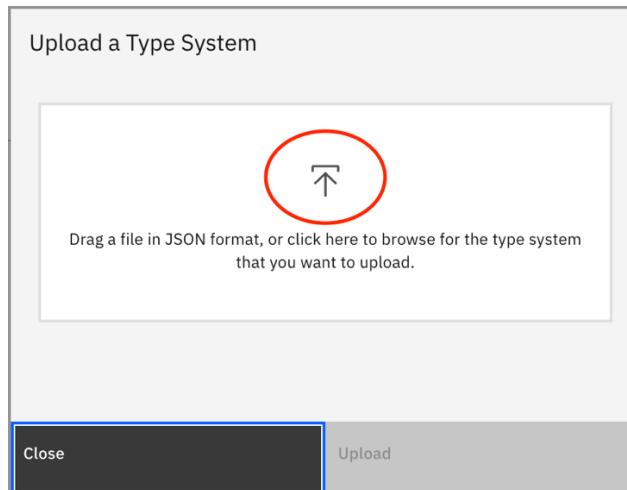
A screenshot of the "Create Workspace" dialog box. It has fields for "Workspace name" (containing "COVID19-Vulnerability") and "Language of documents" (set to "English"). There are also "Advanced Options" and "Cancel" buttons. The "Create" button at the bottom right is highlighted with a red oval.

Inside of this workspace, we will create a type system consisting of the custom entities of the COVID-19 vulnerability index, create a dictionary, perform manual annotation and upload a training corpus for the development of the entity recognition machine learning model.

6. Although we can manually enter the entity types for our type system, we will instead upload the type system file downloaded from the GitHub repository.
On the Entity Types screen, click **Upload**.

The screenshot shows the Entity Types section of the IBM Watson Knowledge Studio interface. On the left, there's a sidebar with navigation links like 'Back to Workspaces', 'Assets', 'Documents', 'Entity Types' (which is selected and highlighted in grey), 'Relation Types', 'Dictionaries', 'Rule-based Model', 'Machine Learning Model', 'Settings', and 'Help'. The main area is titled 'Entity Types' and has tabs for 'Entity Types' (selected), 'Mention Classes', and 'Mention Types'. Below the tabs, there are buttons for 'Add Entity Type' (blue), 'Upload' (highlighted with a red circle), and 'Create a type system'. To the right of these buttons is a dark grey box containing the text 'Upload an existing type system.' There's also a 'Download Types' button and a search bar with placeholder text 'Enter text to filter'. At the bottom, there are sections for 'Entity Type Name', 'Roles', 'Subtypes', and 'Action'.

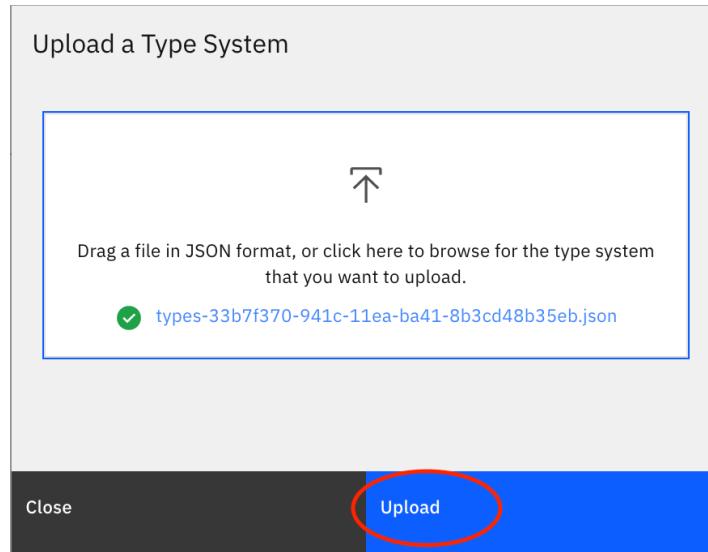
- Click on the upload icon and select the **types-33b7f370-941c-11ea-ba41-8b3cd48b35eb.json**.



The screenshot shows a file explorer window with a sidebar containing 'Favorites' (Recents, Desktop, Downloads, Documents, Applications, Creative Cloud...), 'iCloud' (iCloud Drive), 'Locations' (Docker, Remote Disc, Network), 'Media' (Music, Photos), and a main pane showing a list of files. One file, 'types-33b7f370-941c-11ea-ba41-8b3cd48b35eb.json', is highlighted with a red circle at the bottom of the list. The file details are as follows:

Name	Date Modified	Size	Kind
Lab 1 - Watson Knowledge Studio	Today at 6:08 PM	1.8 MB	Micros...(docx)
Home_Owner_1589693231974.csv	Today at 1:27 AM	170 bytes	CSV Document
Car_Owner_1589693226813.csv	Today at 1:27 AM	224 bytes	CSV Document
No_Vehicle_1589693223234.csv	Today at 1:27 AM	146 bytes	CSV Document
Mobile_Home_Owner_1589693216310.csv	Today at 1:26 AM	116 bytes	CSV Document
Crowded_Living_1589693209295.csv	Today at 1:26 AM	191 bytes	CSV Document
Apartment_Renter_1589693204808.csv	Today at 1:26 AM	211 bytes	CSV Document
ESL_Speaker_1589693200465.csv	Today at 1:26 AM	212 bytes	CSV Document
Minority_1589693196515.csv	Today at 1:26 AM	324 bytes	CSV Document
Single_Parent_1589693192264.csv	Today at 1:26 AM	196 bytes	CSV Document
University_Student_1589693188087.csv	Today at 1:26 AM	304 bytes	CSV Document
No_High_School_Diploma_1589693182889.csv	Today at 1:26 AM	225 bytes	CSV Document
High_School_Student_1589693176460.csv	Today at 1:26 AM	217 bytes	CSV Document
Disabled_1589693170302.csv	Today at 1:26 AM	134 bytes	CSV Document
Minor_1589693163782.csv	Today at 1:26 AM	169 bytes	CSV Document
Senior_Citizen_1589693158519.csv	Today at 1:26 AM	264 bytes	CSV Document
Full_Time_Employment_1589693149530.csv	Today at 1:26 AM	161 bytes	CSV Document
Medically_Insured_1589693143233.csv	Today at 1:26 AM	225 bytes	CSV Document
Hourly_Wage_Employment_1589693134260.csv	Today at 1:25 AM	337 bytes	CSV Document
No_Health_Insurance_1589693125605.csv	Today at 1:25 AM	286 bytes	CSV Document
Unemployed_1589693113791.csv	Today at 1:25 AM	486 bytes	CSV Document
types-33b7f370-941c-11ea-ba41-8b3cd48b35eb.json	Today at 1:23 AM	12 KB	JSON Document

- Click on **Upload**.



You should now see 20 entity types on your screen. These entity types directly pertain to social vulnerability to COVID-19 and will be used to annotate a corpus of social media posts from citizens living in New York City, Washington DC, Los Angeles, Seattle and Chicago – 5 cities that are among the most populous in the U.S. and most affected by the COVID-19 pandemic.

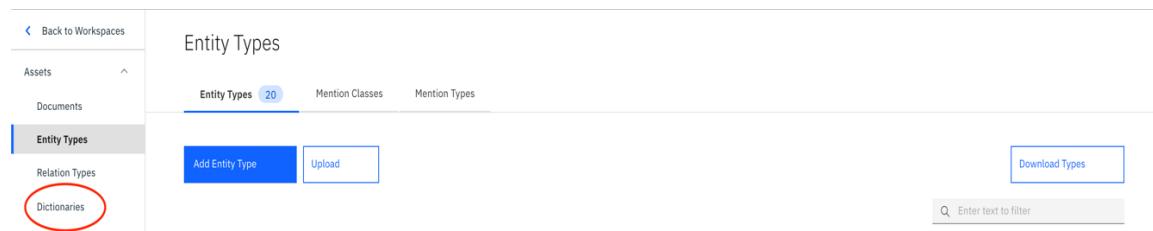
Entity Type Name	Roles	Subtypes	Action
No_Health_Insurance	No_Health_Insurance		Edit Delete
Medically_Insured	Medically_Insured		Edit Delete
No_Vehicle	No_Vehicle		Edit Delete
Car_Owner	Car_Owner		Edit Delete
Mobile_Home_Owner	Mobile_Home_Owner		Edit Delete
Crowded_Living	Crowded_Living		Edit Delete
Apartment_Renter	Apartment_Renter		Edit Delete
Home_Owner	Home_Owner		Edit Delete
ESL_Speaker	ESL_Speaker		Edit Delete
Minority	Minority		Edit Delete

Exercise 4: Create a Dictionary

To help with manual annotation (which we will tackle in the next exercise), we will create a dictionary for each of the entity types in our type system. Each dictionary will contain a list of terms and key phrases pertaining to each entity type.

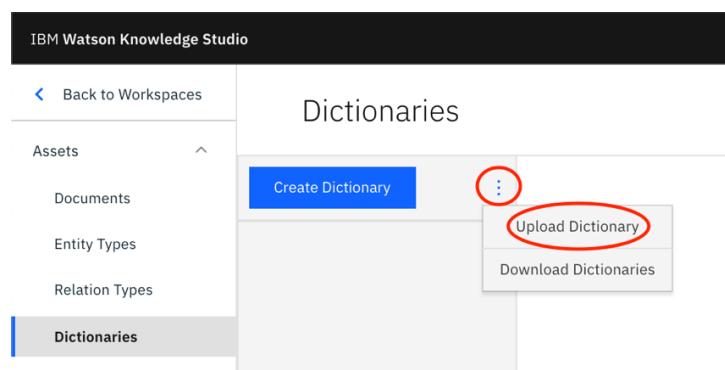
Although we can manually create a dictionary for each entity type, we will instead upload dictionary files for all 20 entity types in this exercise.

1. Under Assets, click **Dictionaries**.



The screenshot shows the 'Entity Types' page. The left sidebar has 'Assets' and 'Entity Types' selected. Under 'Entity Types', 'Relation Types' and 'Dictionaries' are listed, with 'Dictionaries' circled in red. The main area shows 'Entity Types 20', 'Add Entity Type', 'Upload', and 'Download Types' buttons. A search bar at the bottom right says 'Enter text to filter'.

2. On the Dictionaries page, we can upload the zip file containing dictionaries for all of our entity types. Click on the **vertical dots icon** and select **Upload Dictionary**.



The screenshot shows the 'Dictionaries' page. The left sidebar has 'Assets' and 'Dictionaries' selected. The main area has 'Create Dictionary', a vertical dots icon, 'Upload Dictionary' (circled in red), and 'Download Dictionaries' options.

3. Click on the **Upload icon** and select the **COVID19_dictionary_1589757211592.zip** file.

Upload Dictionaries

Drag a file here, or click to browse for the file that you want to upload.

You can upload a dictionary CSV file to create a read-only dictionary. The CSV file must be under 16MB. You can also upload a ZIP file that contains dictionaries downloaded from another workspace. The type system and dictionary ZIP file must originate from the same workspace. Upload the type system that was downloaded from the other workspace in JSON format before you upload the dictionary file. Dictionaries that you reuse from another workspace can be edited.

Close **Upload**

Favorites	Name	Date Modified	Size	Kind
Recents	▶ COVID19_dictionary_1589757211592	Today at 7:13 PM	--	Folder
Desktop	COVID19_dictionary_1589757211592.zip	Today at 7:13 PM	7 KB	ZIP archive

4. Click on **Upload**.

Upload Dictionaries

Drag a file here, or click to browse for the file that you want to upload.

You can upload a dictionary CSV file to create a read-only dictionary. The CSV file must be under 16MB. You can also upload a ZIP file that contains dictionaries downloaded from another workspace. The type system and dictionary ZIP file must originate from the same workspace. Upload the type system that was downloaded from the other workspace in JSON format before you upload the dictionary file. Dictionaries that you reuse from another workspace can be edited.

✓ COVID19_dictionary_1589757211592.zip

Close **Upload**

You should now be able to see dictionaries for each entity type. We will use these dictionaries to pre-annotate a sample set of the social media posts prior to manual annotation.

Dictionaries

Unemployed
Language: English | 9 entries

Entity type: None Rule class: None

Lemma	Surface Forms	Part of Speech	Action
to file for unemployment	to file for unemployment	Verb	Edit Delete
no job	no job	Noun	Edit Delete
weekly claims	weekly claims	Noun	Edit Delete
got laid off	got laid off, Got laid off	Verb	Edit Delete
unemployment office	unemployment office	Noun	Edit Delete
filed for unemployment	filed for unemployment, Filed for unemployment	Verb	Edit Delete
file for unemployment	file for unemployment, file for Unemployment	Verb	Edit Delete
unemployment	unemployment	Noun	Edit Delete
unemployment claim	unemployment claim	Noun	Edit Delete

Add Entry Upload Download Enter text to filter

In order to save these dictionaries as a pre-annotator, we need to remember to match each dictionary with its corresponding entity type. For example, the Unemployed dictionary, which currently has an entity type of None needs to be matched to the Unemployed entity type.

Entity type: **None**

To fix this, we will have to update the Entity type for each dictionary. For the Unemployed dictionary:

- Click the drop-down menu under Entity type and select **Unemployed**.

Dictionaries

Create Dictionary	⋮	
Unemployed 9	⋮	
No_Health_Insura... 5	⋮	
Hourly_Wage_Emp... 9	⋮	
Medically_Insured 6	⋮	
Full_Time_Employ... 4	⋮	
Senior_Citizen 7	⋮	
Minor 4	⋮	
Disabled 3	⋮	

Unemployed
Language: English | 9 entries

Entity type:

Select an entity type

- High_School_Student
- No_High_School_Diploma
- University_Student
- Minor
- Senior_Citizen
- Full_Time_Employment
- Hourly_Wage_Employment
- Unemployed**

Rule cl Non

Repeat the same process for each dictionary until all 20 dictionaries are matched to their corresponding entity type (none of the dictionaries should have an Entity type of None).

Exercise 5: Upload a corpus of documents

In this exercise, we will upload a corpus of social media posts to which we will apply a dictionary pre-annotator and perform manual annotation. This is a small set of social media posts containing first-hand narratives from citizens living in New York City, Washington D.C., Los Angeles, Seattle and Chicago.

- Under Assets, select **Documents**.

IBM Watson Knowledge Studio

Back to Workspaces

Assets

- Documents**
- Entity Types
- Relation Types
- Dictionaries**
- Rule-based Model
- Machine Learning Model

Dictionaries

Create Dictionary	⋮	
Unemployed 9	⋮	
No_Health_Insura... 5	⋮	
Hourly_Wage_Emp... 9	⋮	
Medically_Insured 6	⋮	

Unemployed
Language: English | 9 entries

Entity type:

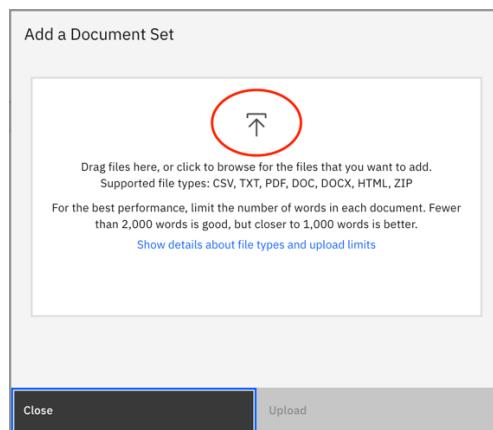
None

Add Entry | Upload

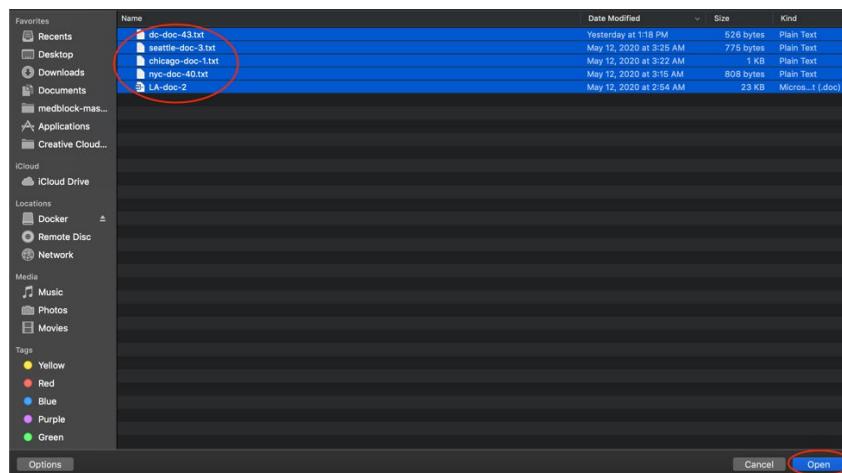
- Click **Upload Document Sets**.

The screenshot shows the 'Documents' section of the IBM Watson Knowledge Studio. On the left, there's a sidebar with 'Assets' expanded, showing 'Documents' selected. Below it are 'Entity Types', 'Relation Types', 'Dictionaries', and 'Rule-based Model'. The main area has tabs for 'Document Sets (1)' and 'Documents (All, 0)'. A red circle highlights the 'Upload Document Sets' button, which is part of a larger callout box with the text 'Add documents to the corpus'.

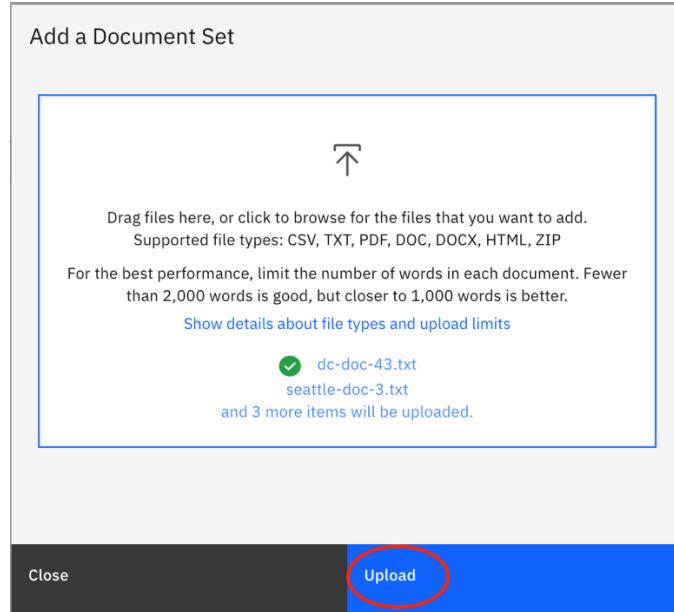
3. Click on the **Upload icon** and double-click on the **SampleDocs** folder.



4. Shift select all 5 documents in the folder and click **Open**.



5. Click **Upload**.



You should now be able to see a set of five documents named dc-doc-43.txt_set to which we will apply a dictionary pre-annotator as well as manually annotate in the next exercise.

Name	Documents	Last Modified	Action
All	5	-	
dc-doc-43.txt_set	5	05/17/2020	Rename Delete

Exercise 6: Perform Manual Annotation

In order to create an entity recognition model, we will need to teach Watson about our custom entity types by manually annotating a sample corpus of documents.

We will start by pre-annotating the document set with our dictionaries. This will allow Watson to quickly annotate our documents using the terms defined in each entity type dictionary.

1. Under Machine Learning Model, click **Pre-annotation**.

The screenshot shows the 'Documents' section of the IBM Watson Knowledge Studio. On the left, there's a sidebar with various options like 'Assets', 'Entity Types', 'Relation Types', 'Dictionaries', 'Rule-based Model', 'Machine Learning Model', 'Pre-annotation' (which is circled in red), 'Annotations', 'Performance', and 'Versions'. The main area shows 'Document Sets (2)' and an 'Upload Document Sets' button. Below it is a search bar with 'Name' and dropdown options 'All' and 'dc-doc-43.txt_set'.

2. Click Run Pre-annotators.

This screenshot shows the 'Pre-annotation' wizard. At the top, it says 'You can run pre-annotators on document sets. Click Run Pre-annotators to start the pre-annotation wizard. If the pre-annotator you want to run is not available, open the menu and make the necessary changes to enable the pre-annotator.' It also says 'Click Order Settings to change the execution order of pre-annotators. Learn more'.

Order	Pre-annotator	Status	⋮
1	Rule-based Model	Not available	⋮
2	Dictionaries	Available	⋮
3	Machine Learning Model	Not available	⋮
4	Natural Language Understanding	Not available	⋮

At the bottom right, there's an 'Order Settings' link and a 'Run Pre-annotators' button, which is circled in red.

You should be able to see that Dictionaries is available as a pre-annotator. If you do not see any available pre-annotators in the table, please revisit Exercise 4, step 5 to match each dictionary with its corresponding entity type.

3. Under Select pre-annotators, click the checkbox next to Dictionaries and click Next.

This screenshot shows the 'Select pre-annotators' dialog. It has a header 'Run Pre-annotators' and a sub-header 'Select pre-annotators'. Below that, it says 'Select the pre-annotators that you want to use.' There's a table with a single row:

Pre-annotator
<input checked="" type="checkbox"/> Dictionaries

At the bottom right, there are 'Close' and 'Next' buttons, with the 'Next' button circled in red.

4. Under Select document sets, click the checkbox next to dc-doc-43.txt_set and click Run.

Run Pre-annotators

Select document sets

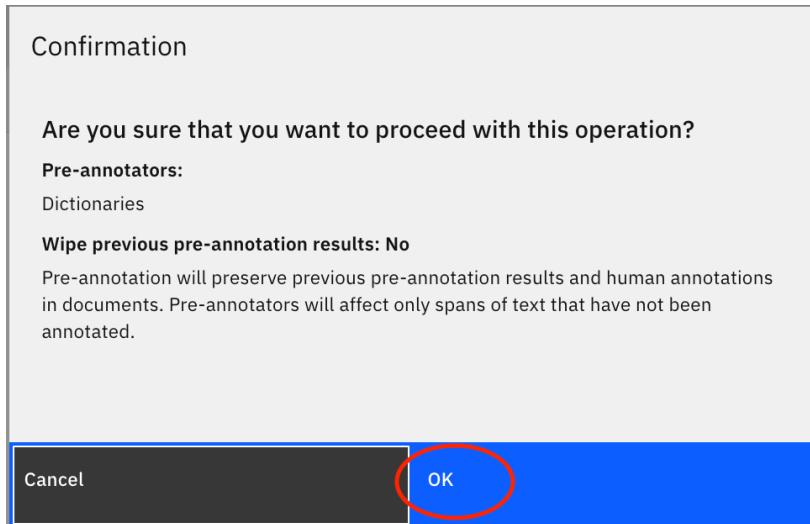
Check if you want to remove previous pre-annotation results from documents before running the pre-annotators. If not checked, all previous annotations are preserved.
 * Annotations made by humans outside of the pre-annotation process remain even if you check the wipe option.

Wipe previous pre-annotation results

Select the document sets or annotation sets that you want to pre-annotate.

Document set	Documents	Pre-annotated documents	Human annotated documents
<input checked="" type="checkbox"/> dc-doc-43.txt_set	5	0	0

5. Click OK.



After a few seconds, pre-annotation will be complete and you will see the following success message:

IBM Watson Knowledge Studio

Pre-annotation

You can run pre-annotators on document sets. Click Run Pre-annotators to start the pre-annotation wizard. If the pre-annotator you want to run is not available, open the menu and make the necessary changes to enable the pre-annotator.

Click Order Settings to change the execution order of pre-annotators. [Learn more](#)

Success: May 17, 2020 10:30:59 PM
Pre-annotation complete.

6. Under Machine Learning Model, click Annotations.

Machine Learning Model

Pre-annotation

Annotations (highlighted with a red circle)

1 Rule-based Model	Not available
2 Dictionaries	Available
3 Machine Learning Model	Not available
4 Natural Language Understanding	Not available

On the Annotations screen, you will see that Watson used the dictionary pre-annotator to annotate 3 of the 5 documents. We will now manually annotate all 5 documents. When annotating each document, we will highlight any mention of the custom entity types in each social media post.

7. Click the second **Annotate** link on the dc-doc-43.txt_set row.

Document Set	Last Modified	Documents (Annotated/Total)	Action
All	-	3 / 5	Annotate
dc-doc-43.txt_set	05/17/2020	3 / 5	Annotate

8. On the Select Document screen, click the first **Open** link (on the dc-doc-43.txt row).

Document Name	Status	Last Modified	Actions
dc-doc-43.txt		May 17, 2020 10:30:48 PM	Open

9. To annotate the dc-doc-43.txt, we will skim through the post and find any mention of the custom entity types. We see that this post mentions one entity type in particular – **University_Student**. The following sentences can be highlighted with this entity type: “**UDC incoming student**,” “**I will be in the speech program at UDC**” and “**off campus student housing**.”

To highlight a sentence with an entity type, click on the first word of the sentence and then the last word before clicking on the corresponding entity type on the right, which in this post is only University_Student.

Manually annotate the above sentences with the **University_Student** entity type so that you get the following annotated post:

The screenshot shows the annotation interface for the file 'dc-doc-43.txt'. On the left, the document content is displayed with several annotations highlighted in red. The right side features a detailed legend for entity types, organized into three columns: Entity, Subtype, and Role. The Entity column lists categories like Apartment_Renter, Cat_Owner, Crowded_Living, Disabled, ESL_Speaker, Full_Time_Employment, High_School_Student, Home_Owner, Hourly_Wage_Employment, Medically_Insured, Minor, Minority, Mobile_Home_Owner, No_Health_Insurance, No_High_School_Diploma, No_Vehicle, Senior_Citizen, Single_Parent, Unemployed, and University_Student. The Subtype and Role columns provide more specific details for each entity type.

10. Click **Save** and click **Open document list** to return to the list of documents.

This screenshot shows the same annotation interface as above, but with the 'Save' button and the 'Open document list' link both circled in red. This indicates the steps required to save the annotations and return to the main document list.

11. Click the **Open** link on the seattle-doc-3.txt row.

This screenshot shows a 'Select Document' dialog box. It displays a list of documents under the heading 'Document Set: dc-doc-43.txt_set'. The list includes 'seattle-doc-3.txt', 'chicago-doc-1.txt', 'nyc-doc-40.txt', 'LA-doc-2.doc', and 'dc-doc-43.txt'. Each row shows the document name, status (last modified date), and an 'Actions' column with an 'Open' link. The 'Open' link for the 'seattle-doc-3.txt' row is circled in red.

12. Annotate the seattle-doc-3.txt and compare your result with the annotated post below.
Don't forget to **Save** your annotation and click **Open document list** when you're done
annotating this post.

Annotations

[Open document list](#)

[View Details](#) [Replace](#) [Concordance](#) [Attribute View](#)

Mention

Relation

Coreference

seattle-doc-3.txt

```

1 { "title": "Etiquette Question - Sheltering in Place, Apartments & Music", "text": "Just getting a read of general feelings on this."
2 Given that many of us are staying at home/working from home during the pandemic, there's a lot more opportunity for grating on each other unintentionally.
3 I live in an apartment complex with fairly thin walls.
4 At what point, in your personal opinion, is it reasonable to start playing music, watch action movies, or other entertainment activities that involve a degree of noise?
5 Personally I don't turn on music my neighbors may hear before 9:00am (I start work early each morning), but is this a good rule of thumb?
6 Too early?
7 Curious to hear folks' thoughts.
8 ", " subreddit": "Seattle", "created": "2020-04-25T01:45:48.000Z" }

```

Save

13. Click the **Open** link on the chicago-doc-1.txt row.

Select Document

Document Set: dc-doc-43.txt_set

Showing 1-5 of 5

Document Name	Status	Last Modified	Actions
chicago-doc-1.txt		May 17, 2020 10:30:48 PM	Open

14. Annotate the chicago-doc-1.txt post and compare your result with the annotated post below. Don't forget to **Save** your annotation and click **Open document list** when you're done annotating this post.

Annotations

[Open document list](#)

[View Details](#) [Replace](#) [Concordance](#) [Attribute View](#)

Mention

Relation

Coreference

chicago-doc-1.txt

```

1 { "title": "Are landlords really allowed to enter occupied apartments to film new virtual tours?"
2 ", "text": "As the questions asks, are landlords actually allowed to do this?
3 I had my alderman refer me to tenant rights but haven't heard anything back.
4 Speaking with a lawyer they said they recording a virtual tour is a nice compromise and didn't seem aware that in person showing if occupied units is actually prohibited.
5 \n\nTo add insult to injury, they are posting these videos publicly on YouTube with the unit numbers along with the name of the apartment, so on top of potentially getting sick we are being forced to publish our private property online for anyone to see.
6 \n\nPrior to knowing the very public way they were distributing these videos we offered to record one only to be told the format would have to be perfect or we'd have to keep recording until we got it correct.
7 \n\nI see a lot of posts about how Chicago is a very tenant friendly city, but I don't see it right now.
8 ", " subreddit": "chicago", "created": "2020-04-24T07:42:46.000Z" }

```

Save

Entity

Type Subtype Role

- Apartment_Renter
- Car_Owner
- Crowded_Living
- Disabled
- ESL_Speaker
- Full_Time_Employment
- High_School_Student
- Home_Owner
- Hourly_Wage_Employment
- Medically_Insured
- Minor
- Minority
- Mobile_Home_Owner
- No_Health_Insurance
- No_High_School_Diploma
- No_Vehicle
- Senior_Citizen
- Single_Parent
- Unemployed
- University_Student

15. Click the **Open** link on the nyc-doc-40.txt row.

Document Name	Status	Last Modified	Actions
nyc-doc-40.txt		May 17, 2020 10:30:48 PM	Open

16. Annotate the nyc-doc-40.txt post and compare your result with the annotated post below.
Don't forget to **Save** your annotation and click **Open document list** when you're done
annotating this post.

The screenshot shows the annotation interface for the file 'nyc-doc-40.txt'. On the left, the document content is displayed with several annotations highlighted in yellow. On the right, there is a sidebar with a tree view of entity types and a 'Save' button. The 'Unemployed' entity type is selected in the tree view, and the 'Save' button is circled in red.

17. Click the **Open** link on the LA-doc-2.doc row.

Document Name	Status	Last Modified	Actions
LA-doc-2.doc		May 17, 2020 10:30:48 PM	Open

18. Annotate the nyc-doc-40.txt post and compare your result with the annotated post below.
Don't forget to **Save** your annotation and click **Open document list** when you're done
annotating this post.

The screenshot shows the annotation interface for a document named 'LA-doc-2.doc'. The document content is as follows:

```

1 {
2 "title": "Isolated outdoors spot?
3 ";
4 "text": "It's my wife's birthday this weekend and was wondering if anyone knew of any places within and around the city that are isolated where you could drive and park your car to enjoy outdoors for a picnic or something?
5 I don't want to endanger anyone or break any county rules but was just hoping to get us outside of the house for a couple hours to make it at least a little memorable.
6 ";
7 " subreddit": "LosAngeles",
8 "created": "2020-04-22T02:49:34.000Z"
9 }

```

The interface includes a sidebar with icons for Assets, Documents, Entity Types, Relation Types, Dictionaries, Rule-based Model, Machine Learning Model, Pre-annotation, Annotations (which is circled in red), Performance, and Versions. At the top, there are buttons for Back to Annotations, Open document list, View Details, Replace, Concordance, Attribute View, and Save (also circled in red). A font size dropdown is set to 14pt. On the right, a list of entity types is shown with their subtypes and roles, with 'Home_Owner' highlighted.

All 5 documents have now been manually annotated. However, we will need a much larger set of documents in order to train and create a machine learning model. In the next exercise, we will upload the complete corpus of documents and create an entity recognition model.

19. Click **Annotations** to return to the Annotations screen.

The screenshot shows the Annotations screen with the following interface elements:

- Left Sidebar:** Back to Workspaces, Assets, Documents, Entity Types, Relation Types, Dictionaries, Rule-based Model, Machine Learning Model, Pre-annotation, **Annotations** (circled in red), Performance, and Versions.
- Header:** Select Document, Document Set: dc-doc-43.txt_set.
- Table:** A list of documents with columns for Document Name, Status, and Last Modified.

Document Name	Status	Last Modified
dc-doc-43.txt	Normal	May 18, 2020 12:02:21 AM
seattle-doc-3.txt	Normal	May 18, 2020 12:33:04 AM
chicago-doc-1.txt	Normal	May 18, 2020 12:40:47 AM
nyc-doc-40.txt	Normal	May 18, 2020 12:47:26 AM
LA-doc-2.doc	Normal	May 18, 2020 1:01:13 AM

Exercise 7: Train and create a machine learning (ML) annotator

As stated above, we will require a much larger set of documents in order to create a machine learning annotator. Although we can provide a folder with all of the social media posts extracted for each of the 5 cities and instruct you annotate each post one by one, we have already done all of the hard work for you and have prepared a zip file containing the entire corpus of documents called Lab1-WKS.zip. Let's upload this zip file to our workspace.

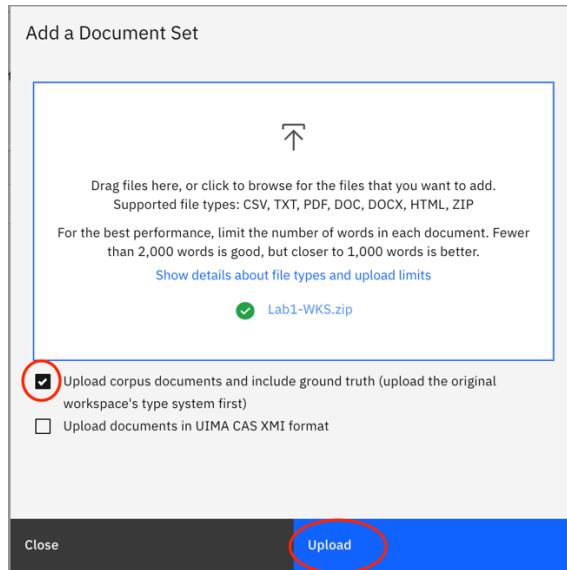
- Under Assets, click **Documents**.

- On the Documents screen, click **Upload Document Sets**.

- Click on the **Upload icon** and select the **Lab1-WKS.zip** file and click **Open**.

	Lab1-WKS.zip	Today at 1:26 AM	5.4 MB	ZIP archive
	SampleDocs.zip	Yesterday at 9:06 PM	10 KB	ZIP archive
▶	SampleDocs	Yesterday at 7:59 PM	--	Folder
▶	nyc-docs	Yesterday at 7:55 PM	--	Folder
▶	LA-docs	Yesterday at 7:53 PM	--	Folder
▶	COVID19_dictionary_1589757211592	Yesterday at 7:13 PM	--	Folder
▶	COVID19_dictionary_1589757211592.zip	Yesterday at 7:13 PM	7 KB	ZIP archive
	Dictionaries.zip	Yesterday at 6:58 PM	19 KB	ZIP archive
▶	Dictionaries	Yesterday at 6:57 PM	--	Folder

4. Click the box next to **Upload corpus documents and include ground truth (upload the original workspace's type system first)** and click **Upload**.



You should now see several new document sets on the Documents screen including an Import document set consisting of 368 posts that were just now added to the workspace. We will be using these newly uploaded documents to train and create a ML annotator.

Name	Documents	Last Modified	Action
All	373	-	Rename Delete
LA-doc-6.doc_set	32	05/12/2020	Rename Delete
nyc-doc-1.txt_set	96	05/12/2020	Rename Delete
chicago-doc-1.txt_set	36	05/12/2020	Rename Delete
seattle-doc-1.txt_set	113	05/12/2020	Rename Delete
dc-doc-1.txt_set	91	05/13/2020	Rename Delete
dc-doc-43.txt_set	5	05/18/2020	Rename Delete
Import	368	05/18/2020	Rename Delete

5. Under Machine Learning Model, click on **Performance**.

The screenshot shows the 'IBM Watson Knowledge Studio' interface. On the left, there is a navigation sidebar under 'Assets' with the following options: Back to Workspaces, Documents (highlighted), Entity Types, Relation Types, Dictionaries, Rule-based Model, Machine Learning Model (expanded), Pre-annotation, Annotations, Performance (circled in red), Versions, Settings, and Help. The main content area is titled 'Documents' and shows 'Document Sets (8)' and 'Documents (All, 373)'. A button labeled 'Upload Document Sets' is present. Below it is a list of document sets: All, LA-doc-6.doc_set, nyc-doc-1.txt_set, chicago-doc-1.txt_set, seattle-doc-1.txt_set, dc-doc-1.txt_set, dc-doc-43.txt_set, and Import.

6. On the Performance screen, click on **Train and evaluate**.

The screenshot shows the 'Performance' screen in IBM Watson Knowledge Studio. The left sidebar has the 'Performance' option selected (highlighted). The main area is titled 'COVID19-Vulnerability' and shows 'Number of documents per set' (0 Training Set, 0 Test Set, 0 Blind Set). It includes sections for 'Training Set' and 'Test Set' with 'View Ground Truth' and 'View Decoding Results' links. There are also 'Last trained on:' and 'Last evaluated on:' fields. A prominent blue button labeled 'Train and evaluate' is circled in red. A message below it says: 'You have not trained the machine learning model. Click here to train and evaluate it.' At the bottom, there is a 'Document set evaluation' section with a 'Model over time' chart, a 'View Log' link, and performance metrics: Precision: -- and Recall: --.

7. On the Select Training/Test/Blind Sets screen, choose **Import**, change the **Training Set** percentage to 85%, **Test Set** to 10% and **Blind Set** to 5%. Click **Train & Evaluate**.

← | Training / Test / Blind Sets

Select Training/ Test/ Blind Sets

Document Set	Task Status
<input type="checkbox"/> All	
<input type="checkbox"/> LA-doc-6.doc_set	
<input type="checkbox"/> nyc-doc-1.txt_set	
<input type="checkbox"/> chicago-doc-1.txt_set	
<input type="checkbox"/> seattle-doc-1.txt_set	
<input type="checkbox"/> dc-doc-1.txt_set	
<input type="checkbox"/> dc-doc-43.txt_set	
<input checked="" type="checkbox"/> Import	

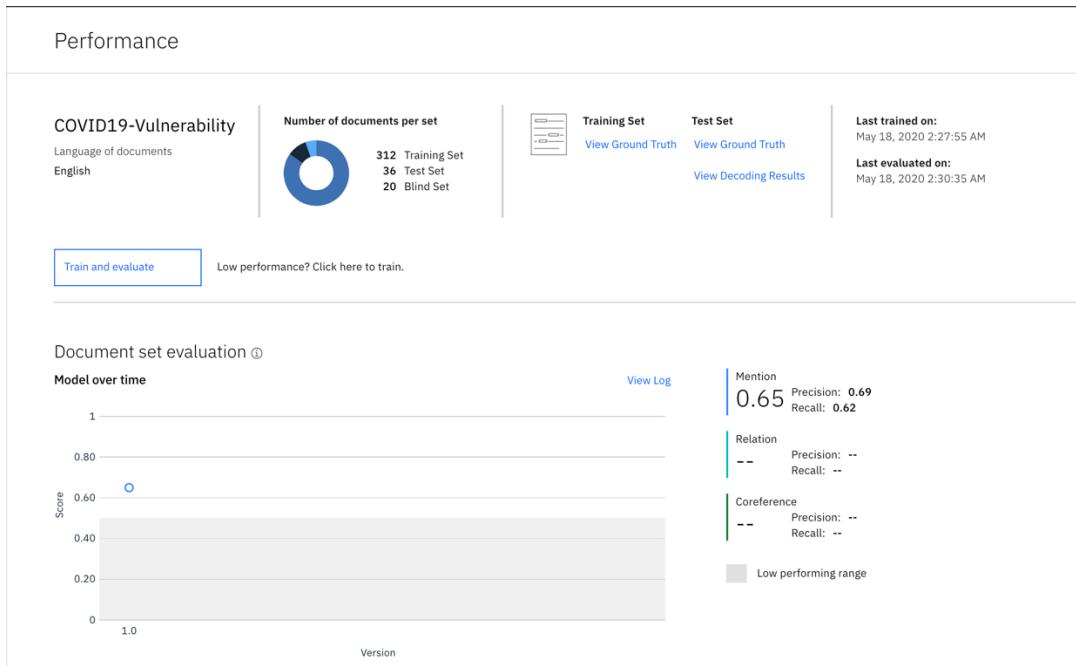
Create new sets by splitting the selected document sets

Ratio
Enter the percentage of documents to include in each set.

85	Training Set (70% Recommended)
10	Test Set (23% Recommended)
5	Blind Set (7% Recommended)

This will start the process of training and evaluating a machine learning annotator, which should take approximately 14 minutes to complete. You will see a progress message on the top right corner of the screen detailing the current phase – training or evaluation – and the amount of time elapsed.

Once the model is created, you should see the following on your Performance screen:

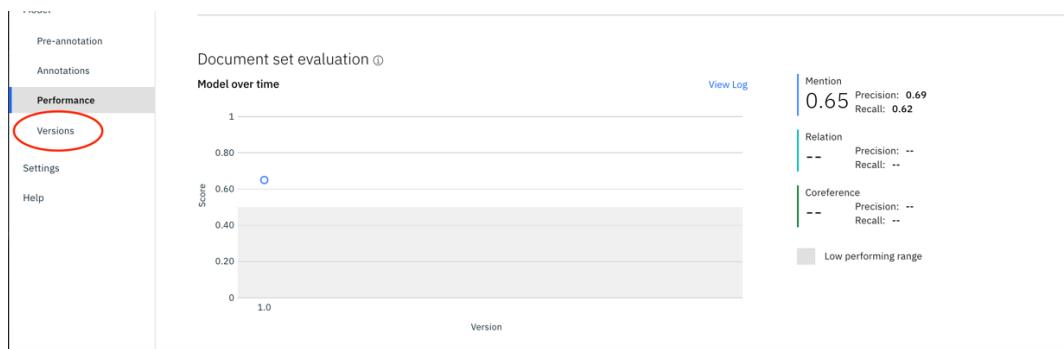


Exercise 8: Save and Deploy the ML Annotator to Discovery

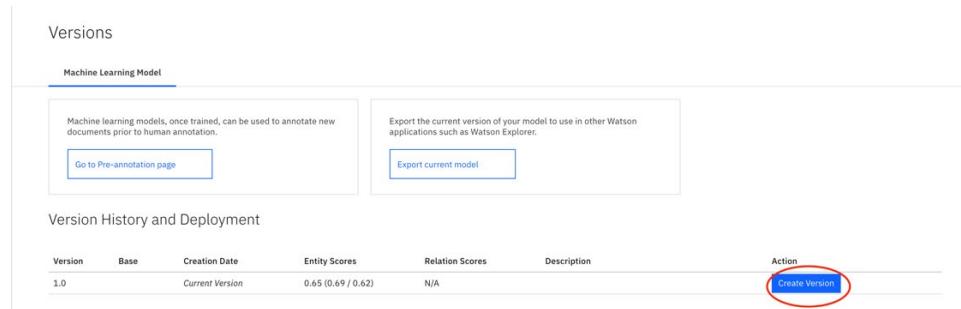
Now that we have a machine learning annotator, we can use it to automatically perform entity extraction inside of Watson Discovery. The automated entity extraction of social media posts for all 5 cities will get us closer to determining the social vulnerability index of each city.

Let's save this machine learning model and deploy it to the Discovery instance that we created at the beginning of this lab.

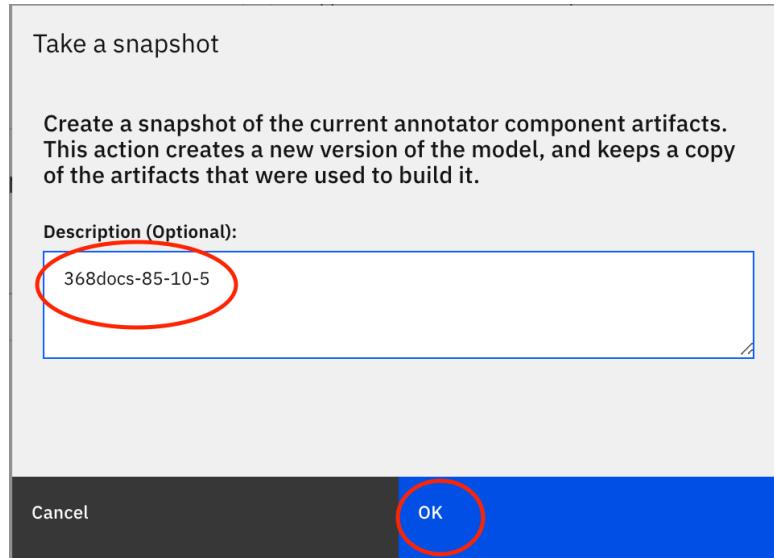
1. Under Machine Learning Model, click on **Versions**.



2. On the Versions page, click **Create Version**.



3. Type **368docs-85-10-5** (to distinguish this as an entity model using 368 docs with an 85-10-5 split) under Description and click **OK**.

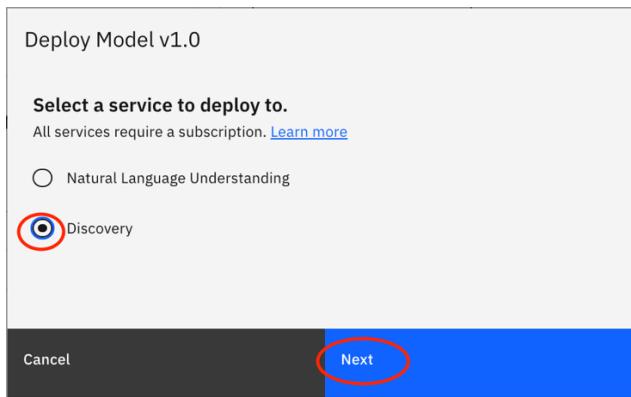


4. In the Version 1.0 row, click **Deploy**.

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		Create Version
1.0		05/18/2020	0.65 (0.69 / 0.62)	N/A	368docs-85-10-5	Promote Delete Deploy

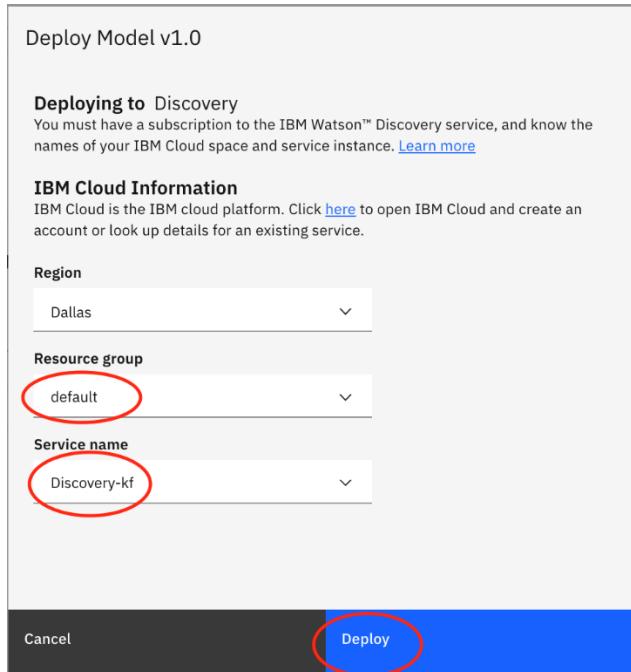
5. Select **Discovery** and click **Next**.



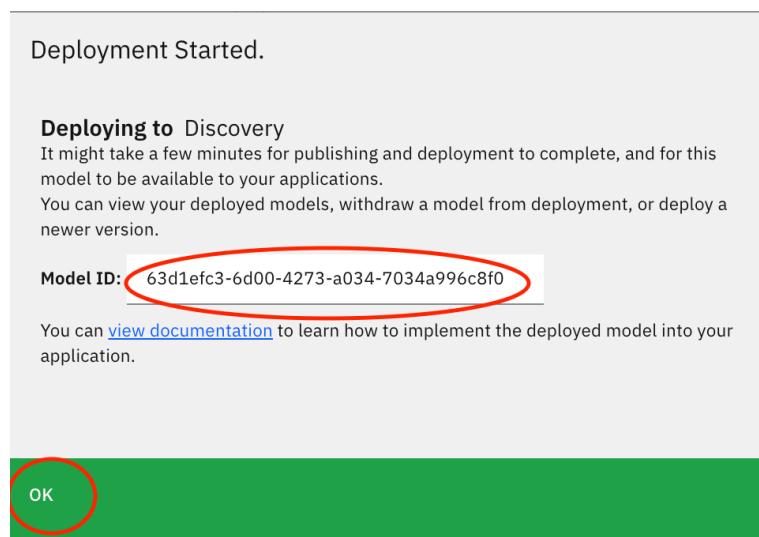
6. In order to deploy this model to your Discovery instance, you will need to select the resource group containing your instance as well as the Service name of the instance that you created.

If this is your first time working with the Watson APIs on the IBM Cloud, you should only have one instance of Discovery currently provisioned.

Select **default** from the drop-down menu under **Resource group** and the **name of the Discovery instance** under **Service name**.



7. Copy the **Model ID** displayed on the screen to use in the next lab and click **OK**.



8. Underneath Deployed Models, you should now see the Model ID number for your newly deployed model. This deployed model will be used to perform entity extraction within Watson Discovery in Lab 2.

The screenshot shows the Watson Studio interface. At the top, there's a navigation bar with icons for Home, Projects, Data, Machine Learning, and Watson Studio. Below that, a sidebar has 'Machine Learning Model' selected. The main content area has two sections: 'Versions' and 'Version History and Deployment'.

Versions

Under 'Machine Learning Model', there are two buttons: 'Go to Pre-annotation page' and 'Export current model'.

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		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		

You have completed Lab 1!