

IBM Training

Student Exercises

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

Hands-On Lab

Legal Copyright: © *Copyright IBM Corp. 2021*

Course materials may not be reproduced in whole or in part without the prior written permission of IBM

Table of Contents

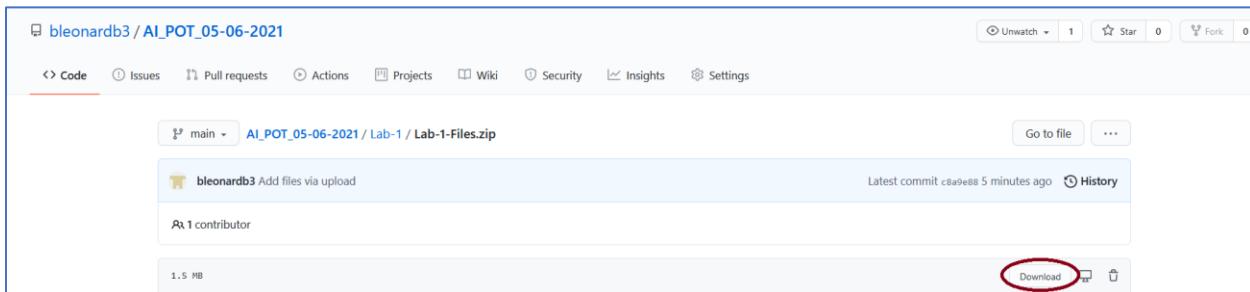
<i>Prerequisites</i>	4
Download the data files to the Desktop	4
<i>Introduction</i>	5
<i>Objectives</i>	5
<i>Exercise 1: Create a Watson Knowledge Studio Instance</i>	5
<i>Exercise 2: Create a Watson Discovery Instance</i>	8
<i>Exercise 3: Create a Type System</i>	9
Mentions	9
Entity Types	9
Relationship Types	10
Steps to create the type system	10
<i>Exercise 4: Create a Dictionary</i>	14
<i>Exercise 5: Upload a corpus of documents</i>	18
<i>Exercise 6: Perform Manual Annotation</i>	21
<i>Exercise 7: Train and create a machine learning (ML) annotator</i>	30

Exercise 8: Save and Deploy the ML Annotator to Discovery..... 34

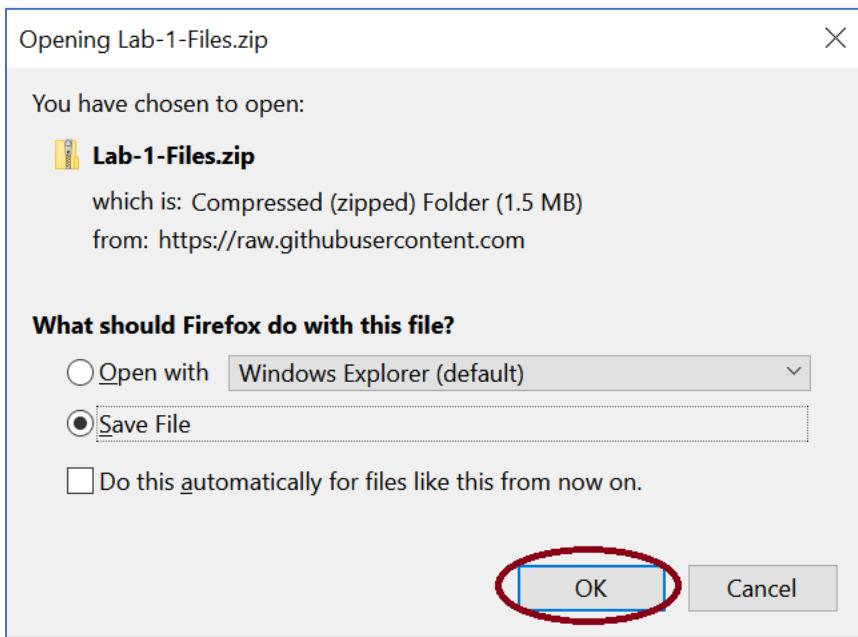
Prerequisites

Download the data files to the Desktop

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



3. Click **OK**.



4. Extract the file contents. You should have 3 files and a directory as shown below.

Name	Date modified	Type
SampleDocs	5/3/2021 11:03 AM	File folder
COVID19_dictionary_1589757211592	5/3/2021 11:03 AM	Compressed (zipped)...
Lab1-WKS	5/3/2021 11:03 AM	Compressed (zipped)...
types-33b7f370-941c-11ea-ba41-8b3cd48b35eb	5/3/2021 11:03 AM	JSON File

Introduction

This lab will cover the development of socioeconomic annotators for COVID-19 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.

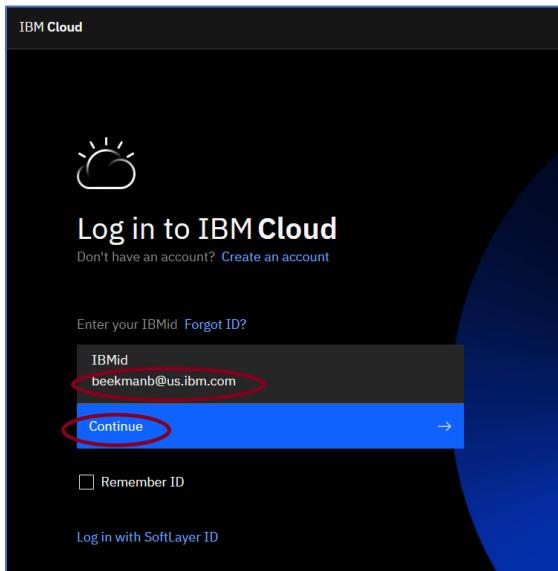
Watson Discovery is an enterprise AI search technology that leverages machine learning, including natural language processing, to retrieve specific answers to your questions and can analyze trends and relationships buried in enterprise data. By integrating a machine learning annotator from Watson Knowledge 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:

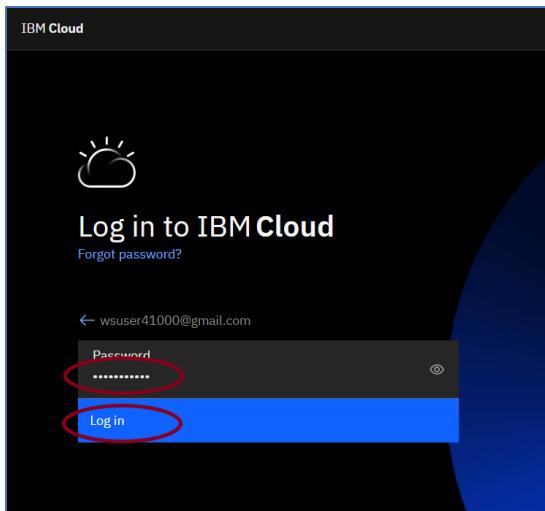
1. Provision an instance of Watson Knowledge Studio
2. Provision an instance of Watson Discovery
3. Create a type system
4. Create a dictionary
5. Upload a corpus of documents
6. Perform manual annotation
7. Train and create a machine learning (ML) annotator
8. Save and deploy the ML annotator to Watson Discovery

Exercise 1: Create a Watson Knowledge Studio Instance

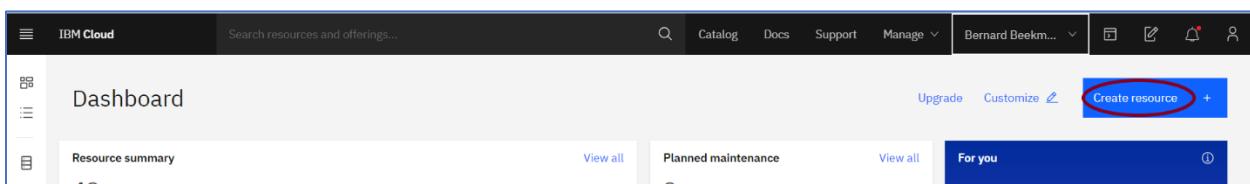
1. Log into your IBM Cloud account by typing **cloud.ibm.com** into the URL address bar of your Firefox or Chrome browser.
2. Enter your **IBMid** and click **Continue**.



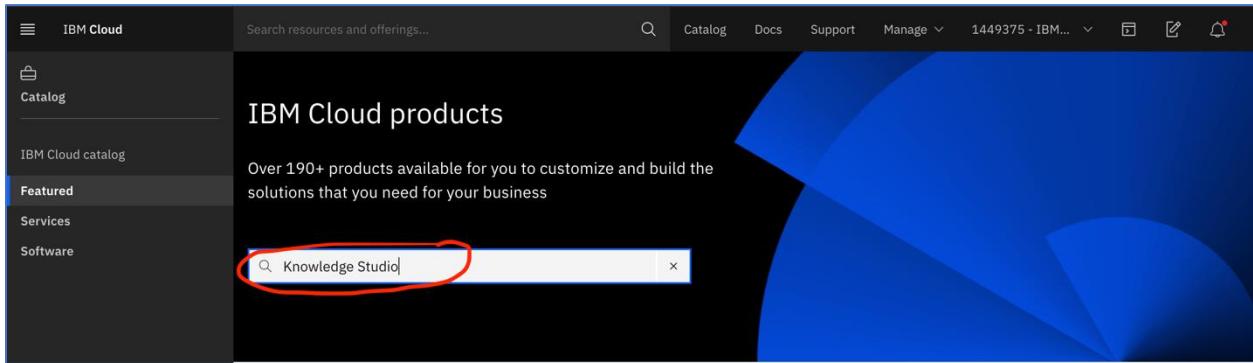
3. Enter your **Password** and click **Log in**.



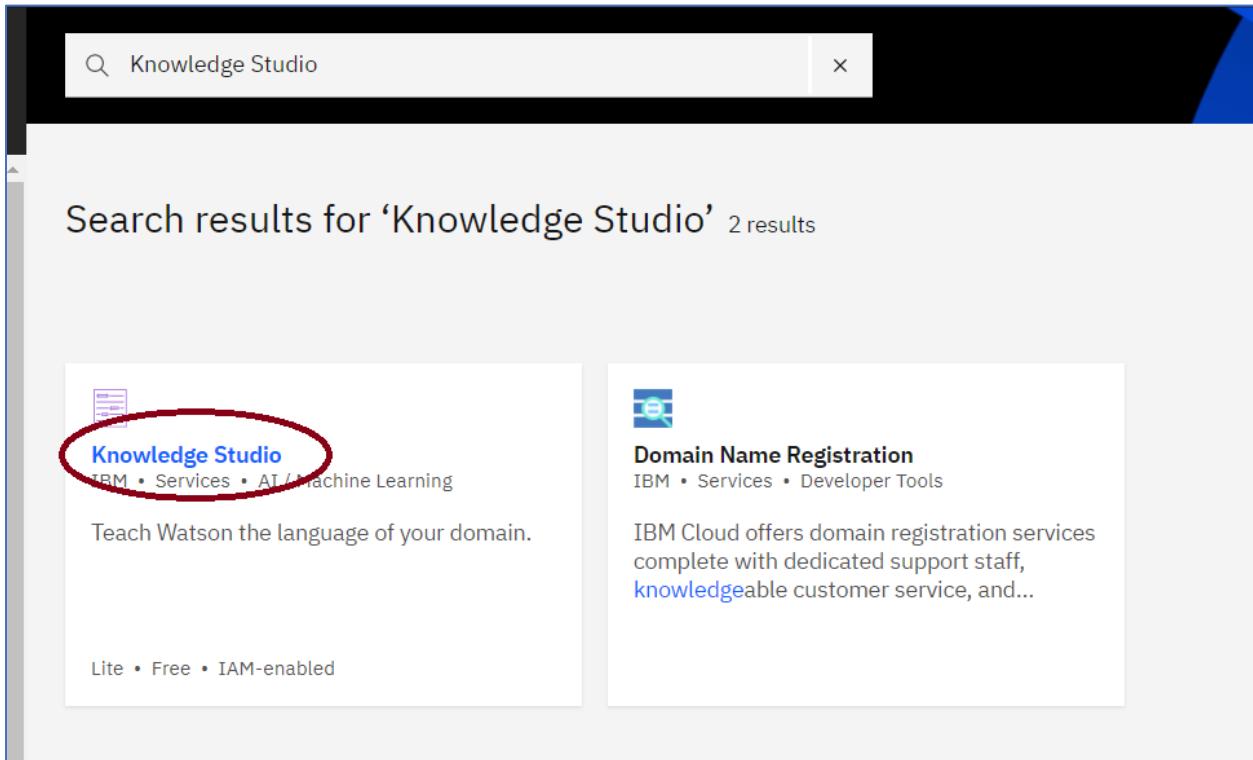
4. Click **Create Resource**.



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



6. Click on **Knowledge Studio**.



7. Click on the **Lite** plan and click **Create**.

The screenshot shows the IBM Cloud Catalog interface for creating a Knowledge Studio instance. The 'Create' tab is selected. A red circle highlights the 'Lite' plan in the pricing table. Another red circle highlights the 'Create' button at the bottom right of the page.

Exercise 2: Create a Watson Discovery Instance

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

The screenshot shows the IBM Cloud Catalog search results for 'Discovery'. The 'Watson Discovery' service is listed under 'Catalog Results' and is highlighted with a red circle. The 'Getting started' tab is selected in the sidebar.

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

Plan

Plan	Features	Pricing
Lite	0 - 1,000 documents per month 200 news queries per month 1 custom model See documentation for plan details	Free
Advanced	Pricing based on document tiers - Up to 50,000 docs (dev environment) - Up to 1M docs - Up to 10M docs - Up to 4M docs - Up to 8M docs - Up to 16M docs - Up to 32M docs - Up to 64M docs - Up to 100M docs News queries Add-on custom models Multi-Tiered...	Click to view tiers and pricing detail
Premium	Everything in Advanced plus.... Usage and Training Data is Private + Stored in an Isolated Single Tenant Environment High Availability and Service Level Uptime Guarantee IBM Cloud Service Endpoints HIPAA - Washington DC Only	

Configure your resource

Service name: Select a resource group:

[View terms](#)

Summary

Discovery

- Region: Dallas
- Plan: Lite
- Service name: Discovery-gr
- Resource group: default

FEEDBACK

Create

[Add to estimate](#)

An instance of the Watson Discovery service will be created. We will link this instance to the machine learning annotator that we create and deploy in this lab. Watson Discovery will use this annotator to perform entity extraction in Lab-2.

Exercise 3: Create a Type System

A type system defines entities that are interesting in your domain content that you want to label with an annotation. The type system controls how content can be annotated by defining the types of entities that can be labeled and how relationships among different entities can be labeled.

In Knowledge Studio, you can create a type system from scratch or upload an existing type system. To jump-start a workspace, you might want to upload a type system that was created for a similar domain. You can then edit the type system to add or remove entity types or redefine the relationship types.

You must create or upload a type system before you begin any annotation tasks. More details about the Watson Knowledge Studio type system are discussed below.

Mentions

A mention is any span of text that you consider relevant in your domain data. For example, in a type system about automotive vehicles, occurrences of terms like **airbag**, **Ford Explorer**, and **child restraint system** might be relevant mentions.

Entity Types

An entity type is how you categorize a real-world thing. An entity mention is an example of a thing of that type. For example, the mention President Obama can be annotated as a PERSON entity type. The mention IBM can be annotated as an ORGANIZATION entity type. Entities are often nouns, but can also be verbs, as long as the verb is important to capture for the purposes of the application that will use the type system. For example, EVENT_CRASH might be a valid entity type for a type system about automotive vehicles, so that the word hit in the sentence, “The car hit the barrier.” can be annotated.

The goal of your annotation workspace is to annotate each mention in a document with the type of thing that it is. After a mention is classified by entity type, the labeled span of text is referred to as an entity.

A best practice is to keep the entity type names sorted and representative, so human annotators can remember them easily. In addition, try to define enough entity types to capture the key concepts that you want to annotate, but not so many entity types that it becomes cumbersome for human annotators to apply the labels accurately.

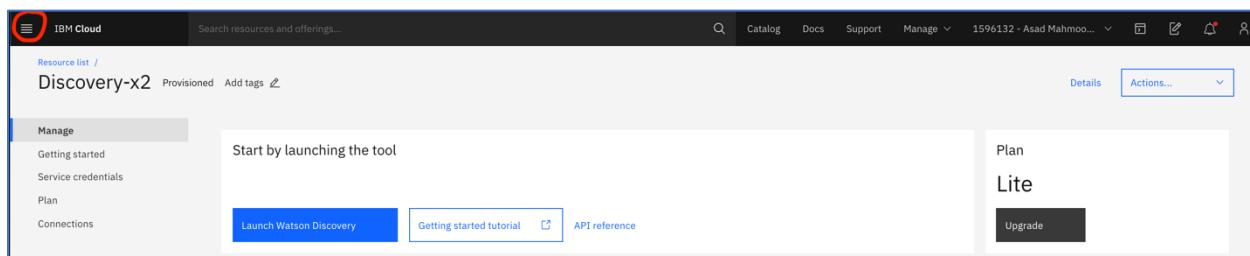
Relationship Types

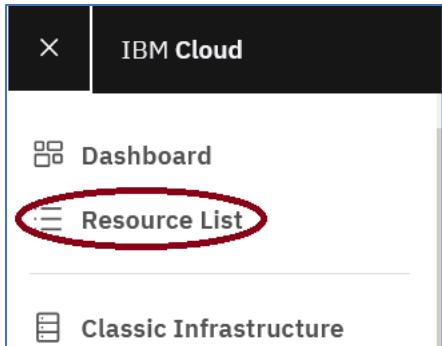
A relation type defines a binary, ordered relationship between two entities. For a relation mention to exist, text must explicitly define the relation and bind mentions of the two entities together and must do so within a single sentence. For example, the sentence **Mary** works for **IBM** is textual evidence of the **employedBy** relation type.

For some relation types, the order of entity mentions matters. For example, the **employedBy** relation type allows the entity type PERSON or PEOPLE as the first mention in the relationship, and ORGANIZATION or GPE as the second mention, but not the other way around. Mary **employedBy** IBM is a valid relationship. IBM **employedBy** Mary is not. For some relation types, such as **spouseOf**, **colleague**, or **sibling**, order does not matter. When you define a relation type where order is not important, a best practice is to add information to the annotation guidelines to regularize how the relation type is used. A convention for noting such symmetrical relations is to say that the entity mention that occurs first in the text should be the first one in the relation.

Steps to create the 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.





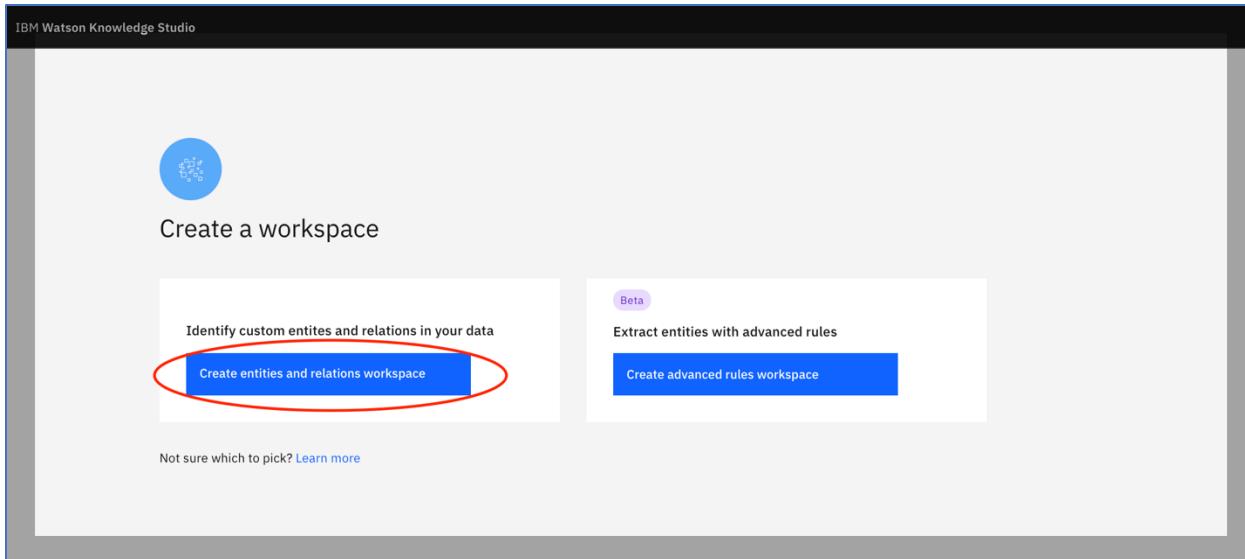
2. Under Services, click on your Knowledge Studio.

A screenshot of the "Services and software" section in the IBM Cloud dashboard. It lists six services: IBM Cognos Dashboard Embedded..., Knowledge Studio-3u (circled in red), KnowledgeCatalog, WatsonMachineLearning, WatsonOpenScale, and WatsonStudio. Each service has columns for name, location, provider, status, and tags (cpdaas). The "Knowledge Studio-3u" row is highlighted with a red circle around its name.

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

A screenshot of the "Knowledge Studio-3u" service details page. The title is "Knowledge Studio-3u" with a green "Active" status and an "Add tags" link. On the left, there's a "Manage" sidebar with "Getting started" and "Plan" options. The main area says "Start by launching the tool" and features a large blue button with white text that reads "Launch Watson Knowledge Studio" (circled in red). To the right of the button is a "Getting started tutorial" link with a help icon.

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



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

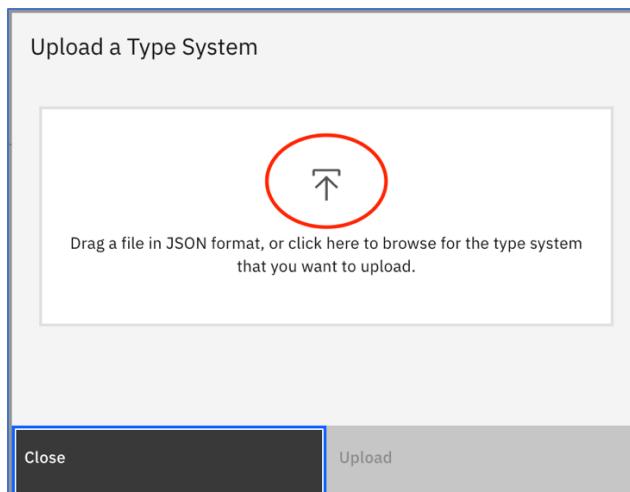
A screenshot of the "Create Workspace" dialog box. It has a title "Create Workspace". On the left, there's a "Workspace name" field containing "COVID19-Vulnerability", which is circled in red. To the right, there's a "Language of documents" dropdown set to "English". Below the name field is a link "+ Add Workspace Description" and an "Advanced Options" button. At the bottom, there are "Cancel" and "Create" buttons, with the "Create" button also circled in red.

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 categories like Assets, Documents, Entity Types (which is selected), Relation Types, Dictionaries, Rule-based Model, Machine Learning Model, Settings, and Help. The main area has tabs for 'Entity Types' (0), 'Mention Classes', and 'Mention Types'. Below these are buttons for 'Add Entity Type', 'Upload' (circled in red), 'Create a type system.', and 'Download Types'. A search bar at the top right says 'Enter text to filter'. The central part of the screen shows a table with columns for 'Entity Type Name', 'Roles', 'Subtypes', and 'Action'. A message 'No items' is displayed.

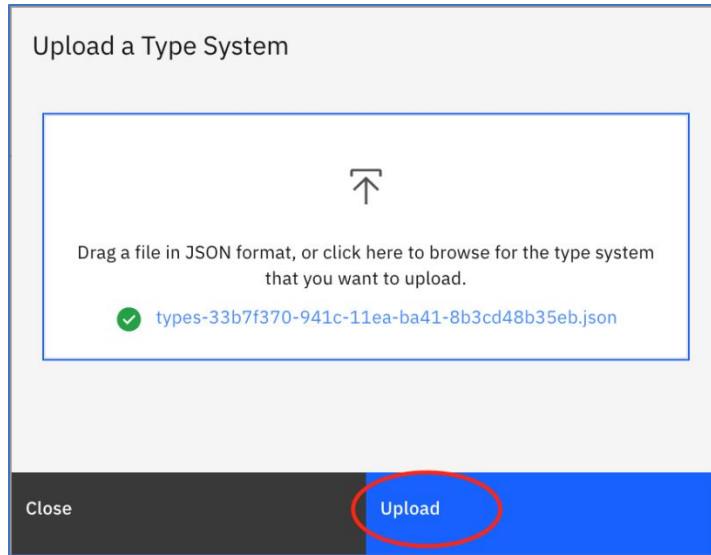
- Click on the upload icon and navigate to the folder where you extracted the zip file downloaded from the GitHub repository. Select **types-33b7f370-941c-11ea-ba41-8b3cd48b35eb.json**.



The screenshot shows a file explorer window with a sidebar containing 'Favorites' (Recents, Desktop, Downloads, Applications, Creative Cloud), 'iCloud' (iCloud Drive), 'Locations' (Docker, Remote Disc, Network), 'Media' (Music, Photos), and a list of CSV files. The 'types-33b7f370-941c-11ea-ba41-8b3cd48b35eb.json' file is highlighted with a red rectangle. The table below lists the files and their details:

Name	Date Modified	Size	Kind
Lab 1 - Watson Knowledge Studio	Today at 6:08 PM	1.8 MB	Micro... (.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 1:25 AM	264 bytes	CSV Document
Full_Time_Employment_1589693149530.csv	Today at 1:25 AM	161 bytes	CSV Document
Medically_Insured_1589693143233.csv	Today at 1:25 AM	225 bytes	CSV Document
Hourly_Wage_Employment_1589693142620.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 were 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. A dictionary is a list of words or phrases that are equivalent for information-extraction purposes, meaning that they are

interchangeable for the purposes of identifying entity and relation mentions. Each dictionary will contain a list of terms and key phrases pertaining to each entity type. Dictionaries help the Knowledge Studio machine learning models to understand the language of the domain. You can create dictionaries in Knowledge Studio by manually adding individual entries. Knowledge Studio also supports the ability to upload several types of dictionary files. We will use this capability to upload dictionary files for all 20 entity types in this exercise.

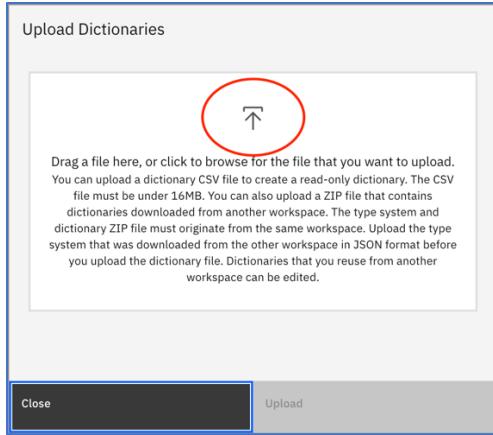
1. Under Assets, click **Dictionaries**.

The screenshot shows the 'Entity Types' page in Knowledge Studio. On the left, there's a sidebar with 'Assets' expanded, showing 'Documents', 'Entity Types' (which is highlighted in blue), 'Relation Types', and 'Dictionaries'. A red circle highlights the 'Dictionaries' link. The main area has tabs for 'Entity Types' (20), 'Mention Classes', and 'Mention Types'. There are buttons for 'Add Entity Type', 'Upload', and 'Download Types'. A search bar at the bottom says 'Enter text to filter'.

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

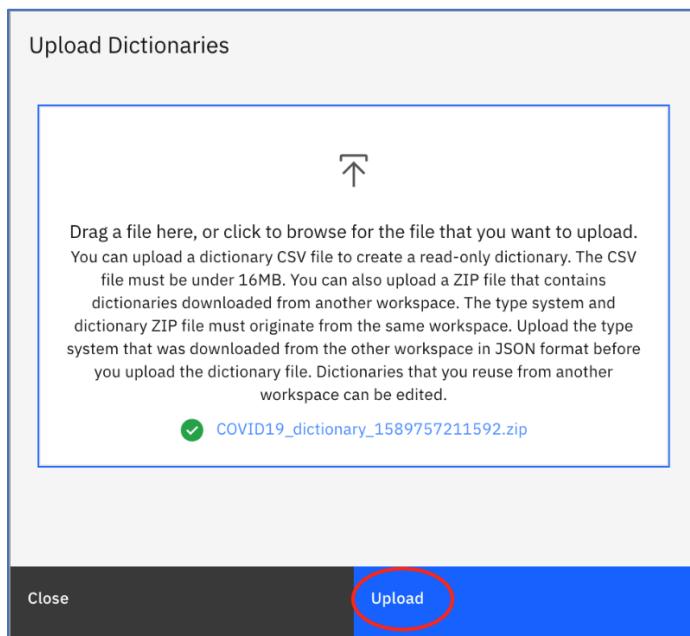
The screenshot shows the 'Dictionaries' page in Knowledge Studio. The left sidebar shows 'Assets' expanded, with 'Dictionaries' selected. A red circle highlights the 'Upload Dictionary' option in a dropdown menu. The menu also includes 'Create Dictionary' (highlighted in blue) and 'Download Dictionaries'. A tooltip explains: 'Upload a CSV file that contains dictionary terms, or a ZIP file that contains one or more CSV files. Up to 100MB of space.' The main area has a large 'Create Dictionary' button and a text box for creating an empty dictionary.

3. Click on the **Upload icon** and navigate to the folder where you extracted the zip file downloaded from the GitHub repository. Select the **COVID19_dictionary_1589757211592.zip** file.



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.



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

In order to save these dictionaries as a pre-annotator, we need 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 c
Non

Surface Forms
File for unemployment
Job
Weekly claims
Got laid off
Unemployment office
Unemployment office

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.

1. 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	⋮

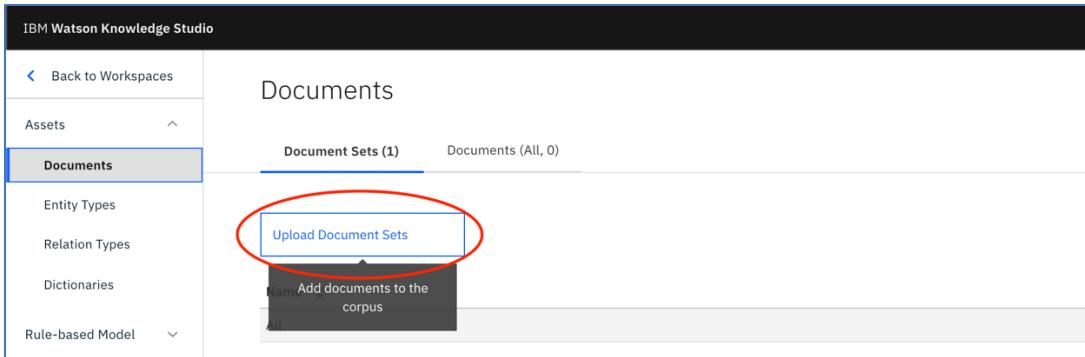
Unemployed
Language: English | 9 entries

Entity type:

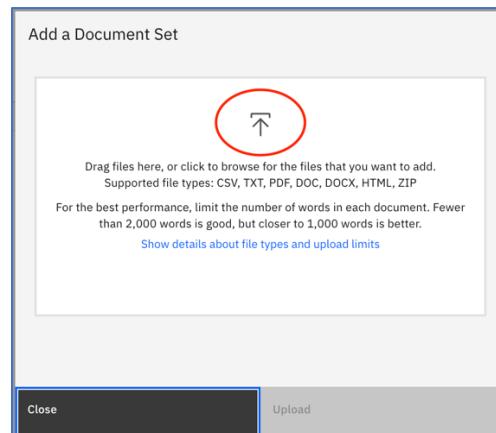
None

Add Entry Upload

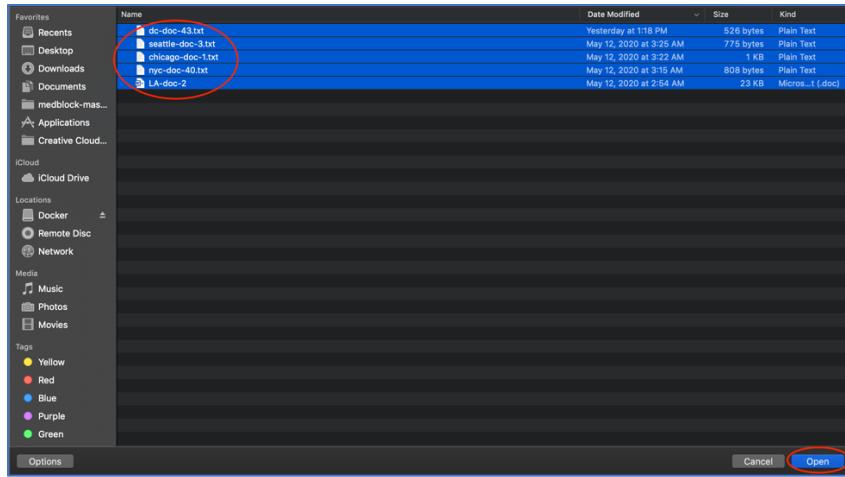
2. Click **Upload Document Sets**.



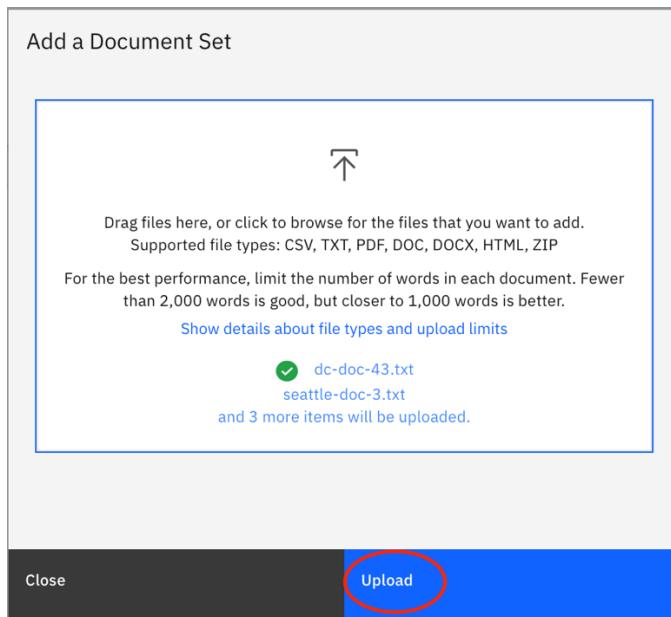
- Click on the **Upload icon** and navigate to the folder where you extracted the zip file downloaded from the GitHub repository. Double-click on the **SampleDocs** folder.



- Shift select** all 5 documents in the folder (don't select the README file) and click **Open**.



5. Click Upload.



You should now be able to see a set of five documents named Chicago-doc-1.txt_set to which we will apply a dictionary pre-annotator as well as manually annotate in the next exercise.

Documents

Document Sets (2) Documents (All, 5)

[Upload Document Sets](#)

To begin annotating documents, go to [Annotations](#) page.

Name	Documents
All	5
chicago-doc-1.txt_set	5

Items per page: 10 ▾ 1-2 of 2 items

Exercise 6: Perform Manual Annotation

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 'Machine Learning Model' section of the Watson Knowledge Studio interface. The 'Pre-annotation' option is circled in red. Other visible options include Entity Types, Relation Types, Dictionaries, Rule-based Model, Annotations, Performance, and Versions.

2. Click **Run Pre-annotators**.

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](#)

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 ⓘ	⋮

Order Settings ⓘ

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.

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

Run Pre-annotators

Select pre-annotators

Select the pre-annotators that you want to use.

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

Close **Next**

- Under Select document sets, click the **checkbox** next to chicago-doc-1.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.

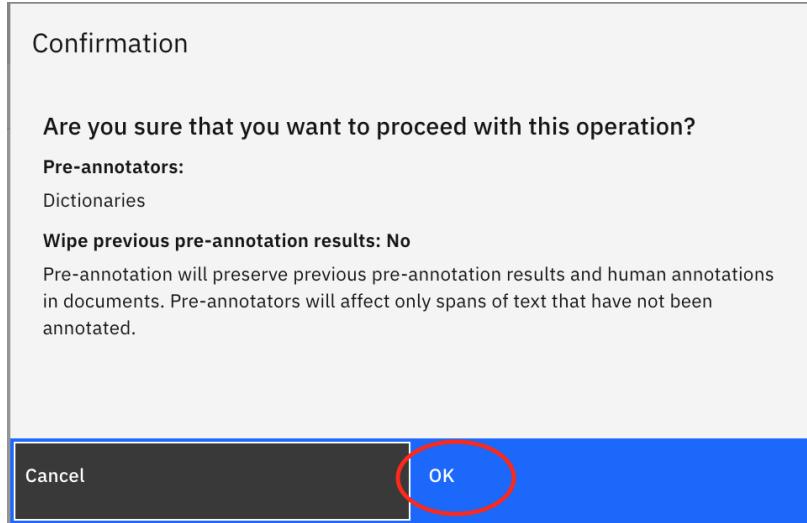
Wipe previous pre-annotation results

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"/> chicago-doc-1.txt_set	5	0	0

Back **Run**

- Click **OK**.



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

IBM Watson Knowledge Studio

Back to Workspaces

Assets

Documents

Entity Types

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.

Annotation Type	Status	Actions
Rule-based Model	Not available	⋮
Dictionaries	Available	⋮
Machine Learning Model	Not available	⋮
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 **Annotate** link on the chicago-doc-1.txt_set row.

Annotations

Document Set	Last Modified	Documents (Annotated/Total)	Action
All	-	3 / 5	Annotate
chicago-doc-1.txt_set	04/26/2021	3 / 5	Annotate

- On the Select Document screen, click on chicago-doc-1.txt.

Select Document

Document Set: chicago-doc-1.txt_set

Showing 1-5 of 5

Document Name	Status	Last Modified
chicago-doc-1.txt		Apr 26, 2021 9:29:17 PM
dc-doc-43.txt		Apr 26, 2021 9:29:17 PM
LA-doc-2.doc		Apr 26, 2021 9:29:17 PM
nyc-doc-40.txt		Apr 26, 2021 9:29:17 PM
seattle-doc-3.txt		Apr 26, 2021 9:29:17 PM

Items per page: 50 ▾ 1-5 of 5 items 1 ▾ of 1 pages ◀ ▶

- To annotate the chicago-doc-1.txt, we will skim through the post and find any mention of the custom entity types. Note that several mentions have already been pre-annotated. The annotation color matches the entity type on the right. When we find a word or phrase that we want to annotate, we click on the first word of the phrase and then the last word of the phrase, and then click on the corresponding entity type on the right. Annotate the chicago-doc-1.txt post and compare your result with the annotated post below. The annotations circled in red need to be manually added.

Back to Annotations | Open document list

Mention

View Details | Replace | Concordance | Attribute View

Save

Entity Mention

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	

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" }

```

10. The **property** pre-annotation is not accurate in this case. Select the property annotation and click the delete button on the keyboard to remove this annotation. Click the **Save** icon and then **Open document list** to annotate the next document.

Back to Annotations | **Open document list**

Mention

View Details | Replace | Concordance | Attribute View

Save

Entity Mention

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	

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" }

```

11. On the **Select Document** panel, click on **dc-doc-43.txt**.

The screenshot shows a 'Select Document' interface. At the top, it says 'Document Set: chicago-doc-1.txt_set'. On the right, there's a 'Close' button. Below that, it says 'Showing 1-5 of 5'. A table lists five documents:

Document Name	Status	Last Modified
dc-doc-43.txt		Apr 26, 2021 9:29:17 PM
LA-doc-2.doc		Apr 26, 2021 9:29:17 PM
nyc-doc-40.txt		Apr 26, 2021 9:29:17 PM
seattle-doc-3.txt		Apr 26, 2021 9:29:17 PM
chicago-doc-1.txt		Apr 27, 2021 1:53:42 PM

At the bottom, it says 'Items per page: 50 ▾' and '1-5 of 5 items'.

12. 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**.”

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

The screenshot shows an annotation interface for the document 'dc-doc-43.txt'. The left side shows the document content with some text highlighted in pink. The right side shows a list of entity types:

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		

13. Click the Save  icon and click Open document list to return to the list of documents.



14. On the Select Document panel, click on LA-doc-2.doc

Document Name	Status	Last Modified
LA-doc-2.doc		Apr 26, 2021 9:29:17 PM
nyc-doc-40.txt		Apr 26, 2021 9:29:17 PM
seattle-doc-3.txt		Apr 26, 2021 9:29:17 PM
chicago-doc-1.txt		Apr 27, 2021 1:53:42 PM
dc-doc-43.txt		Apr 27, 2021 2:30:58 PM

15. Annotate as shown below, and then click the Save  icon and Open document list.

```

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 }

```

16. On the **Select Document** panel, click on **nyc-doc-40.txt**.

The screenshot shows a 'Select Document' panel with a title bar 'Select Document' and a 'Close' button. Below the title bar, it says 'Document Set: chicago-doc-1.txt_set'. A message 'Showing 1-5 of 5' is displayed. The main area is a table with columns 'Document Name', 'Status', and 'Last Modified'. The table contains five rows:

Document Name	Status	Last Modified
nyc-doc-40.txt		Apr 26, 2021 9:29:17 PM
seattle-doc-3.txt		Apr 26, 2021 9:29:17 PM
chicago-doc-1.txt		Apr 27, 2021 1:53:42 PM
dc-doc-43.txt		Apr 27, 2021 2:30:58 PM
LA-doc-2.doc		Apr 27, 2021 2:47:27 PM

At the bottom, there are pagination controls: 'Items per page: 50 ▾', '1-5 of 5 items', '1 ▾ of 1 pages', and navigation arrows.

17. Annotate the nyc-doc-40.txt post and compare your result with the annotated post below.



Click the **Save** icon to save your annotation and click **Open document list** when you're done annotating this post to move to the next document.

The screenshot shows the annotation interface for the 'nyc-doc-40.txt' post. At the top, there are buttons for 'Back to Annotations', 'Open document list' (circled in red), 'View Details', 'Replace', 'Concordance', and 'Attribute View'. On the right, there are buttons for 'Save' (circled in red) and 'Entity', 'Mention', 'Type', 'Subtype', and 'Role'. A sidebar on the left shows 'Mention', 'Relation', and 'Coreference' buttons. The main area displays the post content with annotations:

nyc-doc-40.txt

- 1 {"text":"I know most of the world is **laid off** right now and apparently most of NYC but i can't get through to **unemployment** at all!!! I've been calling for hours i don't understand why they couldn't complete my claim online.
- 2 Is there a center i can go to in person?
- 3 I'd rather wait in line then to call back to back to back to get some automated system that hangs up on me or actually get through to the menu, enter all my info and have it hang up on me AGAIN after I'm supposed to be transfers to a rep.
- 4 I have to pay my **rent** and my partner is also **laid off**.
- 5 He got approved but his benefits aren't going to come for 2-3 weeks it says .. and i can't even get through to get mine approved.
- 6 ANY advice seriously I'm spinning out here", "author_fullname":"t2_4qqx83ci", "title":"**Unemployment**"}

A sidebar on the right lists various entity types with color-coded squares:

- 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

18. On the **Select Document** panel, click on **seattle-doc-3.txt**

Select Document

Document Set: chicago-doc-1.txt_set

Close

Showing 1-5 of 5

Document Name	Status	Last Modified
seattle-doc-3.txt		Apr 26, 2021 9:29:17 PM
chicago-doc-1.txt		Apr 27, 2021 1:53:42 PM
dc-doc-43.txt		Apr 27, 2021 2:30:58 PM
LA-doc-2.doc		Apr 27, 2021 2:47:27 PM
nyc-doc-40.txt		Apr 27, 2021 2:56:59 PM

Items per page: 50 ▾ 1-5 of 5 items

19. Annotate the seattle-doc-3.txt and compare your result with the annotated post below.

Don't forget to click the Save  icon and click Open document list when you're done annotating this post.

[Back to Annotations](#) [Open document list](#) 

Mention 
 Relation 
 Coreference 

View Details  Replace  Concordance  Attribute View 

seattle-doc-3.txt

1 { "title": "Etiquette Question - Sheltering in Place, Apartments & Music", "text": "Just getting a read of general feelings on this." }
 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.
 I live in an apartment complex with fairly thin walls.
 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?
 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?
 Too early?
 Curious to hear folks' thoughts.
 ", " subreddit": "Seattle", "created": "2020-04-25T01:45:48.000Z" }

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.

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

Document Name	Status	Last Modified
chicago-doc-1.txt		Apr 27, 2021 1:53:42 PM
dc-doc-43.txt		Apr 27, 2021 2:30:58 PM
LA-doc-2.doc		Apr 27, 2021 2:47:27 PM
nyc-doc-40.txt		Apr 27, 2021 2:56:59 PM
seattle-doc-3.txt		Apr 28, 2021 2:03:47 PM

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

As stated above, we will require a much larger set of documents 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 to 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.

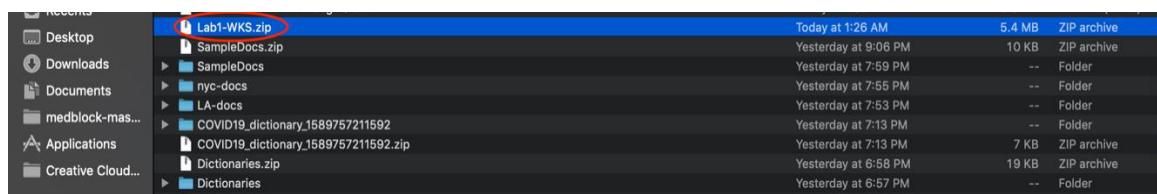
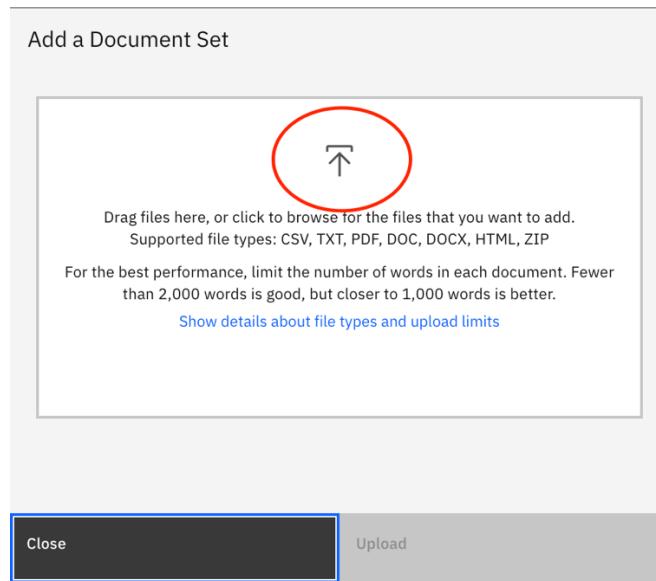
1. Under Assets, click **Documents**.

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

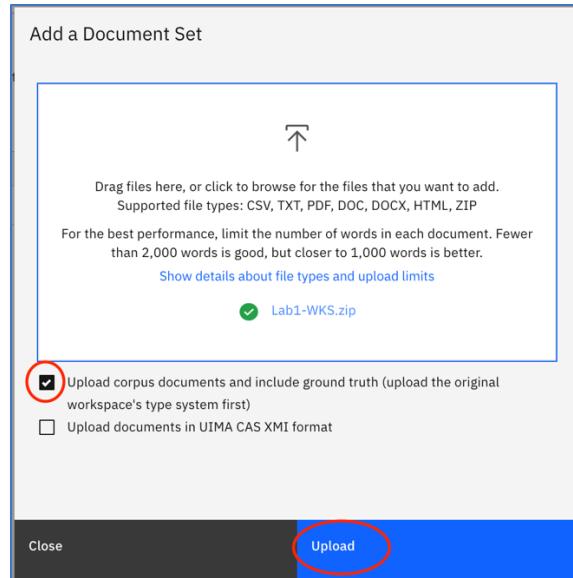
2. On the Documents screen, click **Upload Document Sets**.

The screenshot shows the 'Documents' section of the IBM Watson Knowledge Studio. On the left, there's a sidebar with categories like Assets, Documents (which is selected), Entity Types, Relation Types, Dictionaries, Rule-based Model, and Machine Learning Model. The main area shows 'Document Sets (2)' and 'Documents (All, 5)'. A red circle highlights the 'Upload Document Sets' button at the top left of the document list area. Below it, there's a message: 'To begin annotating documents, go to [Annotations](#) page.' To the right is a 'Download Document Sets' button.

- Click on the **Upload icon** and navigate to the folder where you extracted the zip file downloaded from the GitHub repository. Select the **Lab1-WKS.zip** file and click **Open**.



- 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's a sidebar with a tree view of assets: Entity Types, Relation Types, Dictionaries, Rule-based Model, Machine Learning Model, Pre-annotation, Annotations, **Performance** (which is circled in red), Versions, Settings, and Help. The main panel is titled 'Documents' and shows 'Document Sets (8)' and 'Documents (All, 373)'. It includes a 'Upload Document Sets' button and 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 the IBM Watson Knowledge Studio. The left sidebar shows the 'Performance' section selected. The main area displays 'COVID19-Vulnerability' details: Language of documents is English. It shows a donut chart for 'Number of documents per set' with 0 Training Set, 0 Test Set, and 0 Blind Set. There are buttons for 'Training Set' and 'Test Set' with links to 'View Ground Truth' and 'View Decoding Results'. On the right, it shows 'Last trained on:' and 'Last evaluated on:'. Below this, there's a section for 'Document set evaluation' with a 'Model over time' graph, a 'View Log' button, and mention statistics: Precision: -- and Recall: --. A red circle highlights the 'Train and evaluate' button at the bottom left of the main area.

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

<input type="button" value="Train"/>	<input style="outline: 2px solid red; border-radius: 10px; padding: 2px 10px; border: none; color: inherit; background-color: inherit; font-size: inherit; font-weight: inherit; font-family: inherit; font-style: inherit; text-decoration: none; border-radius: 10px; border: 1px solid #ccc; padding: 5px; margin: 5px 10px;" type="button" value="Train & Evaluate"/>
--------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

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:

Performance

COVID-19 Vulnerability
Language of documents English

Number of documents per set

312	Training Set
36	Test Set
20	Blind Set

Training Set Test Set
[View Ground Truth](#) [View Decoding Results](#)

Last trained on: Apr 28, 2021 2:43:49 PM
Last evaluated on: Apr 28, 2021 2:45:50 PM

Low performance? Click here to train.

Document set evaluation ⓘ

Model over time

View Log

Mention
0.65 Precision: 0.69 Recall: 0.62

Relation
-- Precision: -- Recall: --

Coreference
-- Precision: -- Recall: --

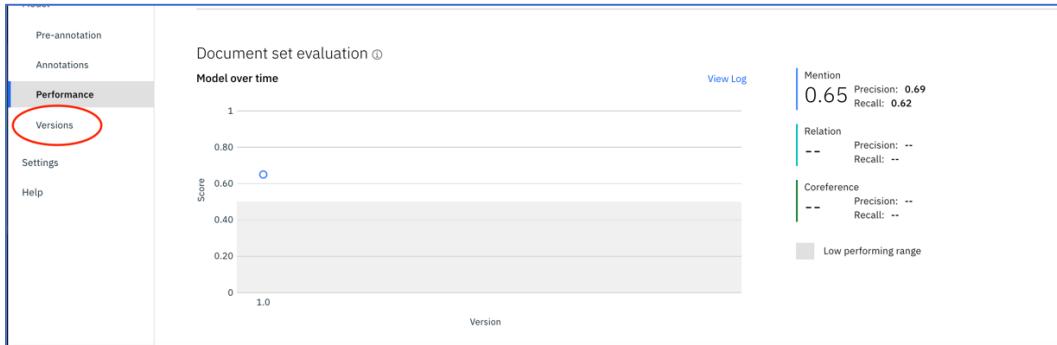
Low performing range

Exercise 8: Save and Deploy the ML Annotator to Discovery

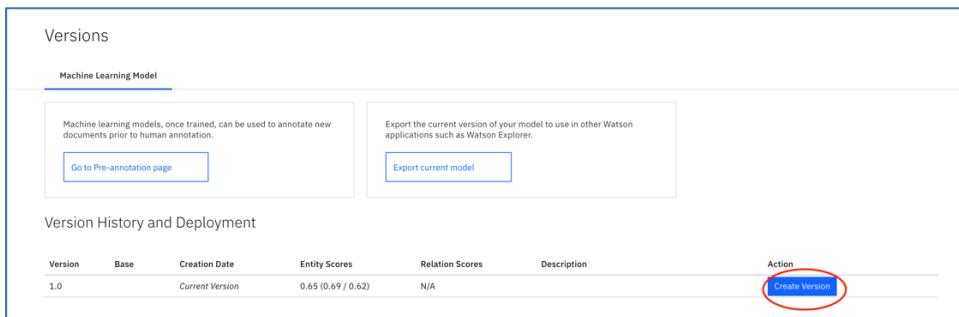
Now that we have a machine learning annotator, we can use to automatically perform entity extraction inside of Watson Discovery. The automated entity extraction of social media posts for all 5 cities will get us closer to determining the social vulnerability index of each city.

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

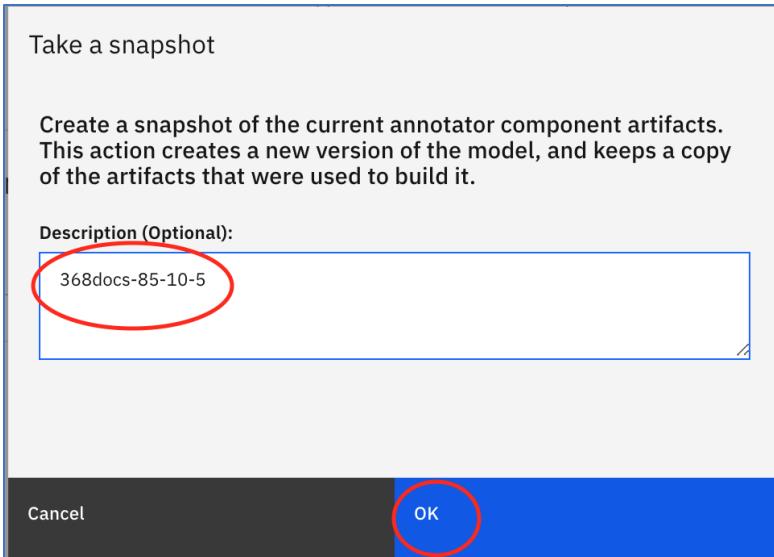
1. Under Machine Learning Model, click on **Versions**.



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



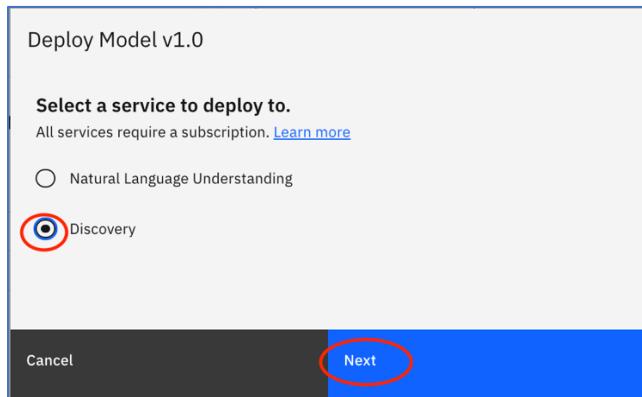
3. Type **368docs-85-10-5** (to distinguish this as an entity model using 368 docs with an 85-10-5 split) under Description and click **OK**.



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	05/18/2020	0.65 (0.69 / 0.62)	N/A		<button>Create Version</button>
1.0	05/18/2020	05/18/2020	0.65 (0.69 / 0.62)	N/A	368docs-85-10-5	<button>Promote</button> <button>Delete</button> 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**.

Deploy Model v1.0

Deploying to Discovery
You must have a subscription to the IBM Watson™ Discovery service, and know the names of your IBM Cloud space and service instance. [Learn more](#)

IBM Cloud Information
IBM Cloud is the IBM cloud platform. Click [here](#) to open IBM Cloud and create an account or look up details for an existing service.

Region
Dallas

Resource group
default

Service name
Discovery-kf

Cancel Deploy

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

Deployment Started.

Deploying to Discovery
It might take a few minutes for publishing and deployment to complete, and for this model to be available to your applications.
You can view your deployed models, withdraw a model from deployment, or deploy a newer version.

Model ID: 63d1efc3-6d00-4273-a034-7034a996c8f0

You can [view documentation](#) to learn how to implement the deployed model into your application.

OK

8. Click on the right arrow ➡ adjacent to **Deployed Models (1)**. You should 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.

Versions

Machine Learning Model

Machine learning models, once trained, can be used to annotate new documents prior to human annotation.	Export the current version of your model to use in other Watson applications such as Watson Explorer.
Go to Pre-annotation page	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!