IBM Training

Student Exercises

Lab-1: Develop Socioeconomic Annotators for COVID-19

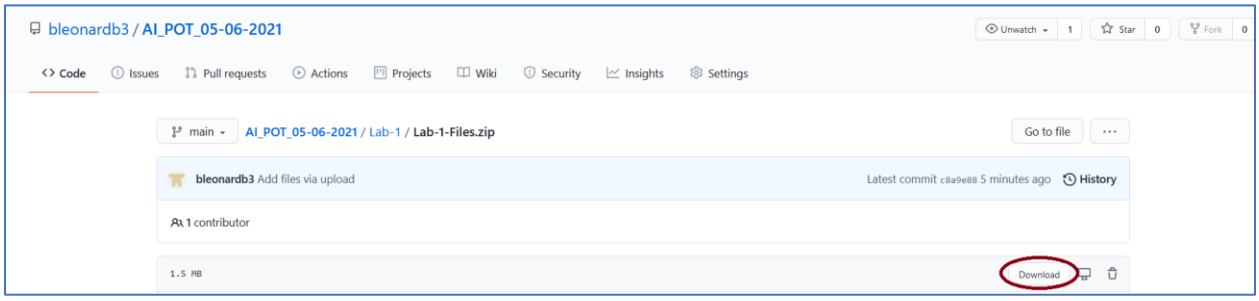
Hands-On Lab

Table of Contents

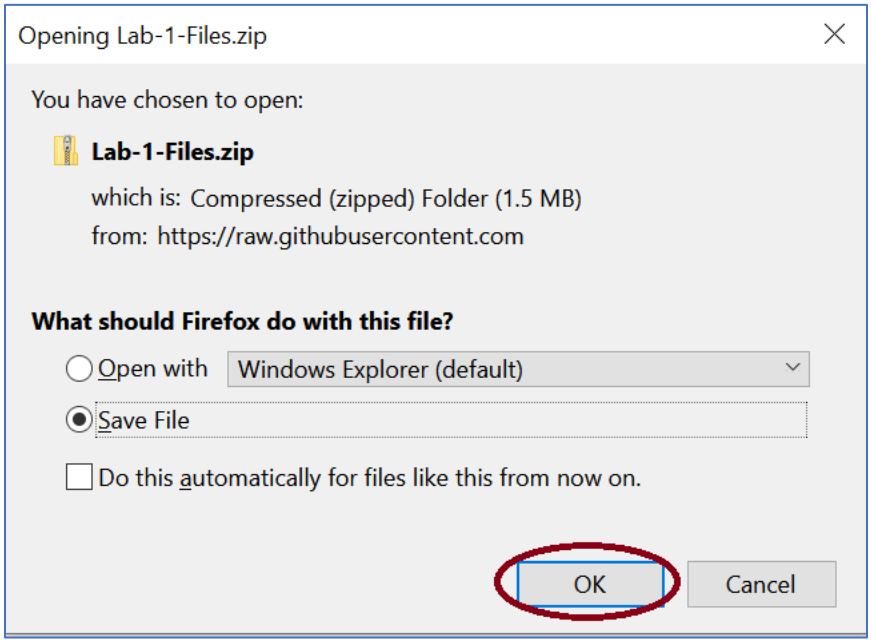
Prerequisites

Download the data files to the Desktop

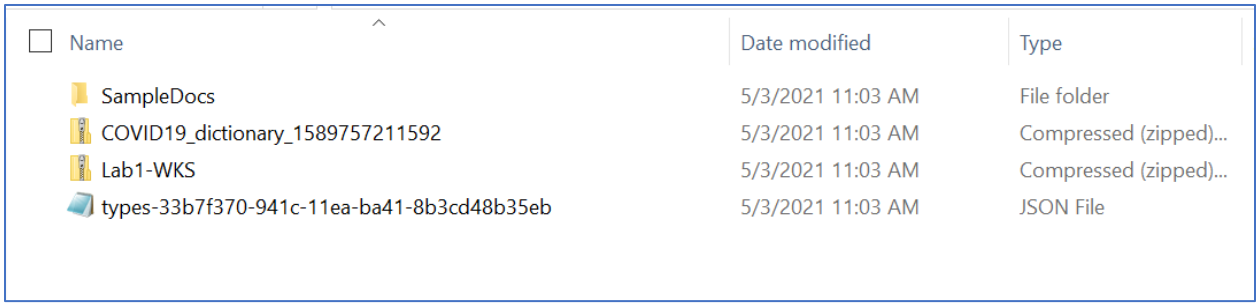
1. Click on https://github.com/bleonardb3/AI_POT_07-29-2021/blob/main/Lab-1/Lab-1-Files.zip
2. Click on the **Download** button.



3. Click **OK**.



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



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.

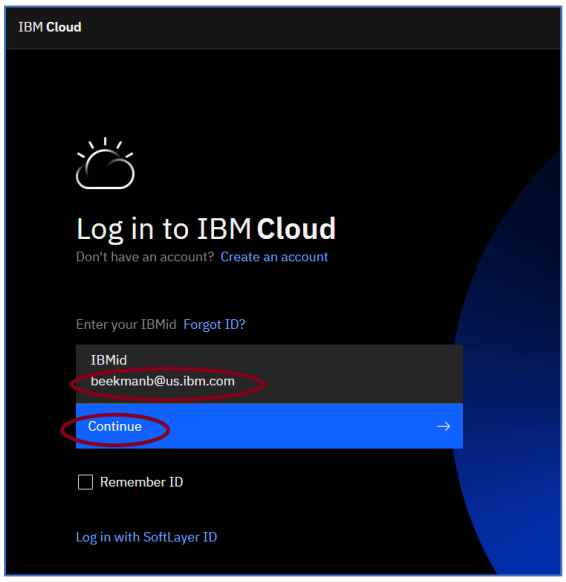
Watson Discovery is an enterprise AI search technology that leverages machine learning, including natural language processing, to retrieve specific answers to your questions and can analyze trends and relationships buried in enterprise data. By integrating a machine learning annotator from Watson Knowledge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

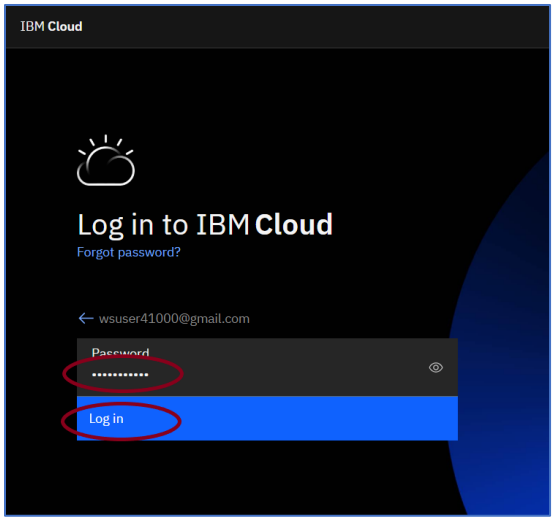
1. Provision an instance of Watson Knowledge Studio
2. Provision an instance of Watson Discovery
3. Create a type system
4. Create a dictionary
5. Upload a corpus of documents
6. Perform manual annotation
7. Train and create a machine learning (ML) annotator
8. Save and deploy the ML annotator to Watson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

Exercise 1: Create a Watson Knowledge Studio Instance

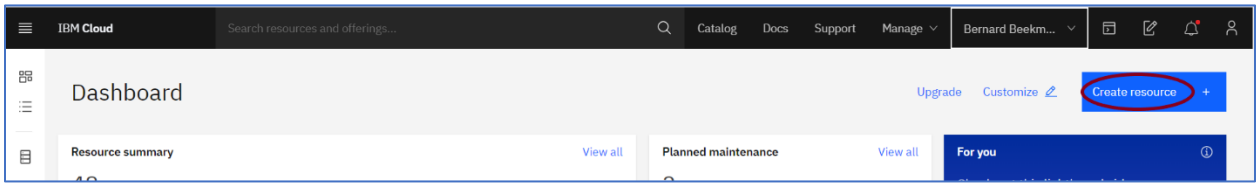
1. Log into your IBM Cloud account by typing **cloud.ibm.com** into the URL address bar of your Firefox or Chrome browser.
2. Enter your **IBMid** and click **Continue.**



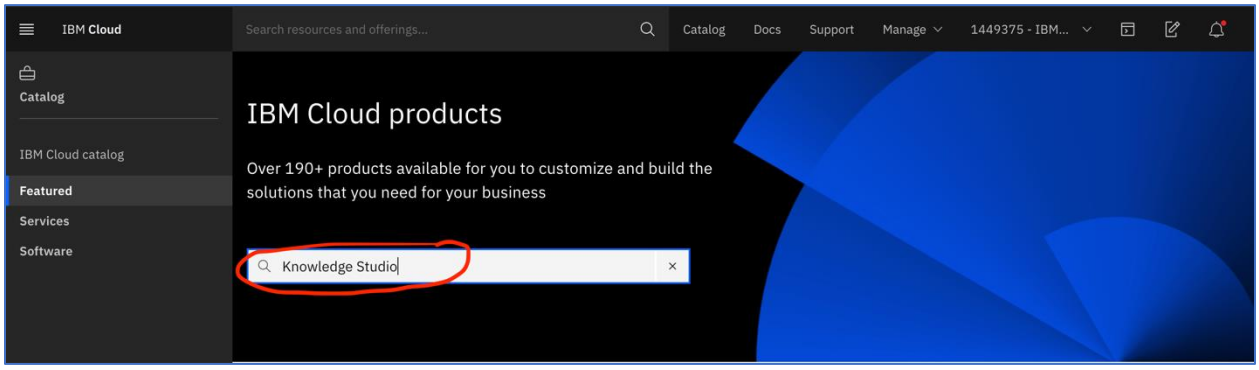
3. Enter your **Password** and click **Log in**.



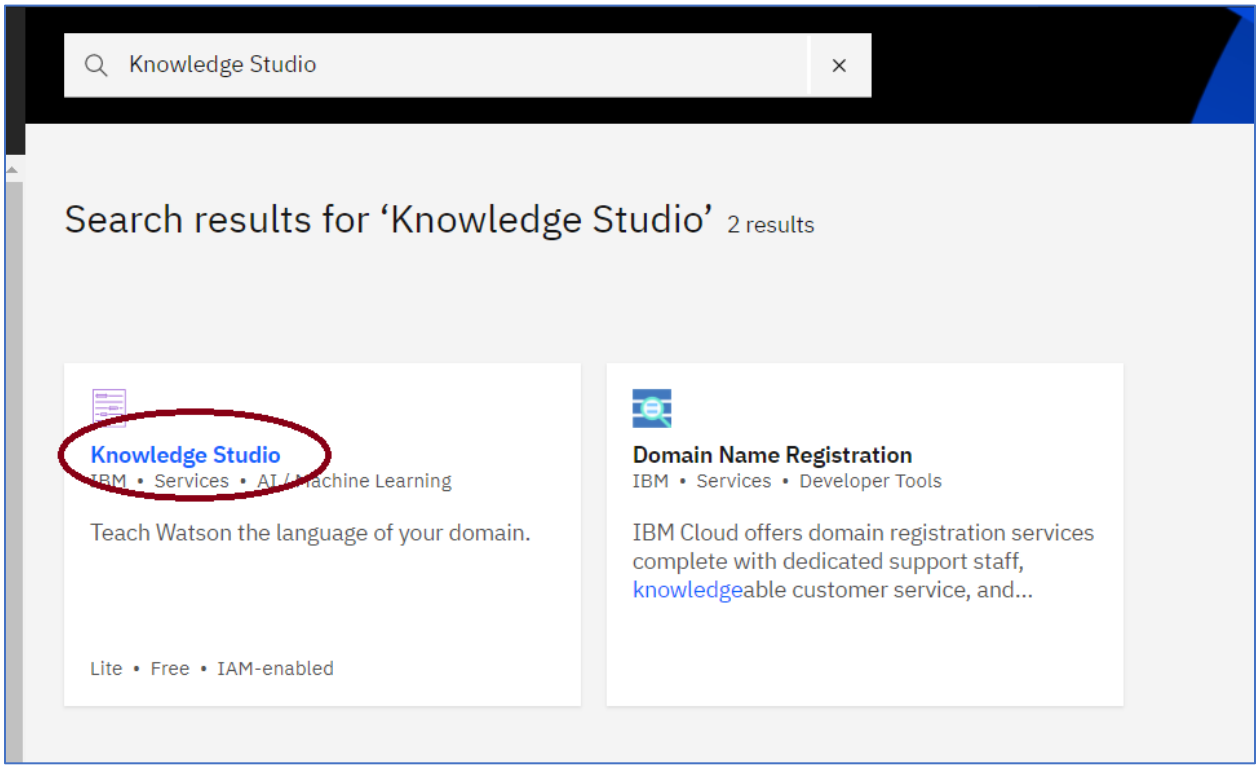
4. Click **Create Resource.**



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



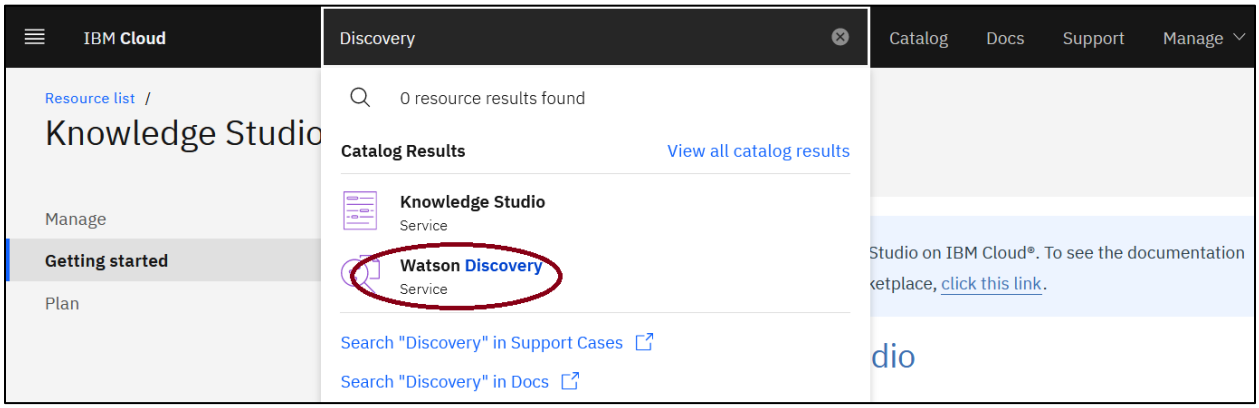
6. Click on **Knowledge Studio**.



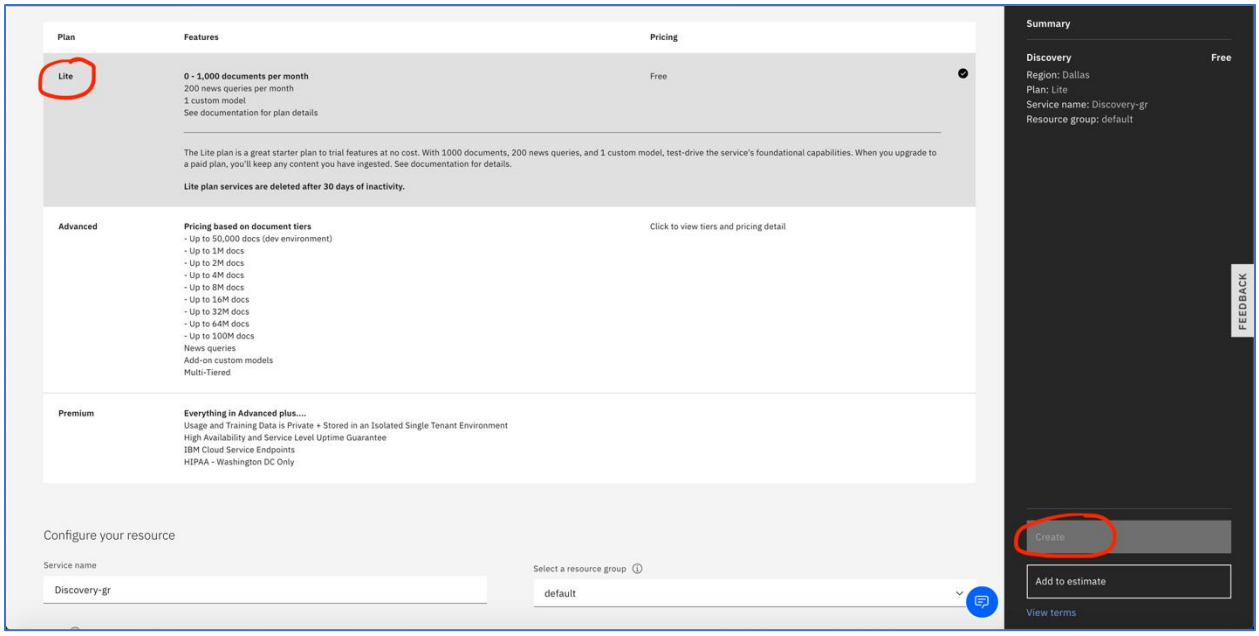
7. Click on the **Lite** plan and click **Create.**

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*.



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



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. Entites 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.

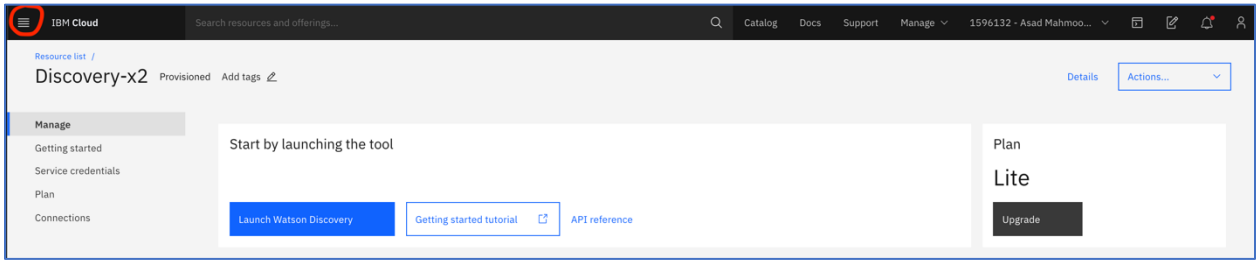
Relationship Types

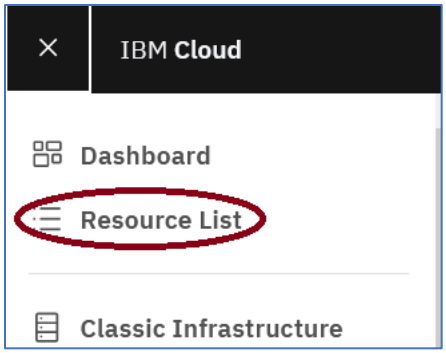
A relation type defines a binary, ordered relationship between two entities. For a relation mention to exist, text must explicitly define the relation and bind mentions of the two entities together and must do so within a single sentence. For example, the sentence **Mary** works for **IBM** is textual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

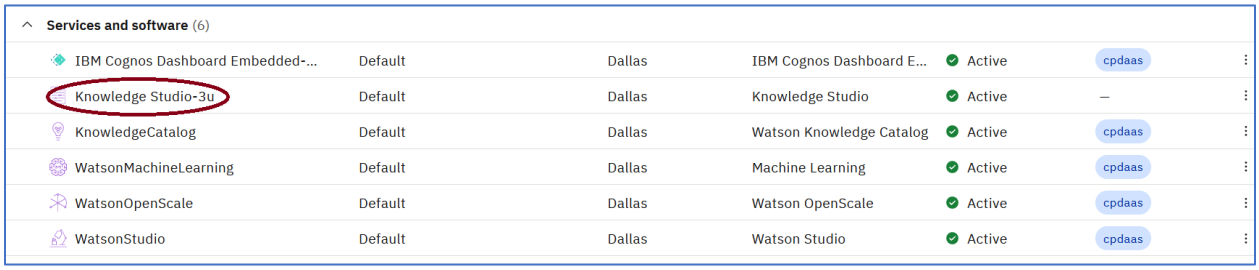
Steps to create the type system

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.

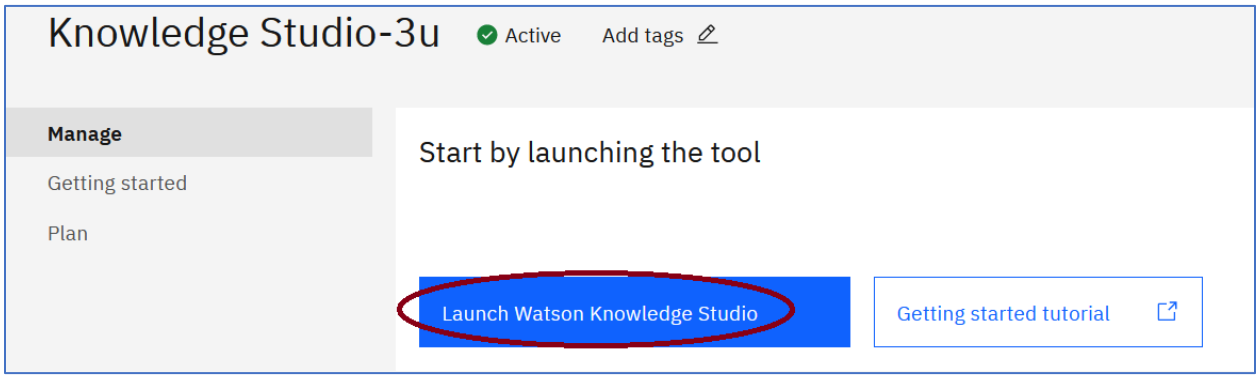




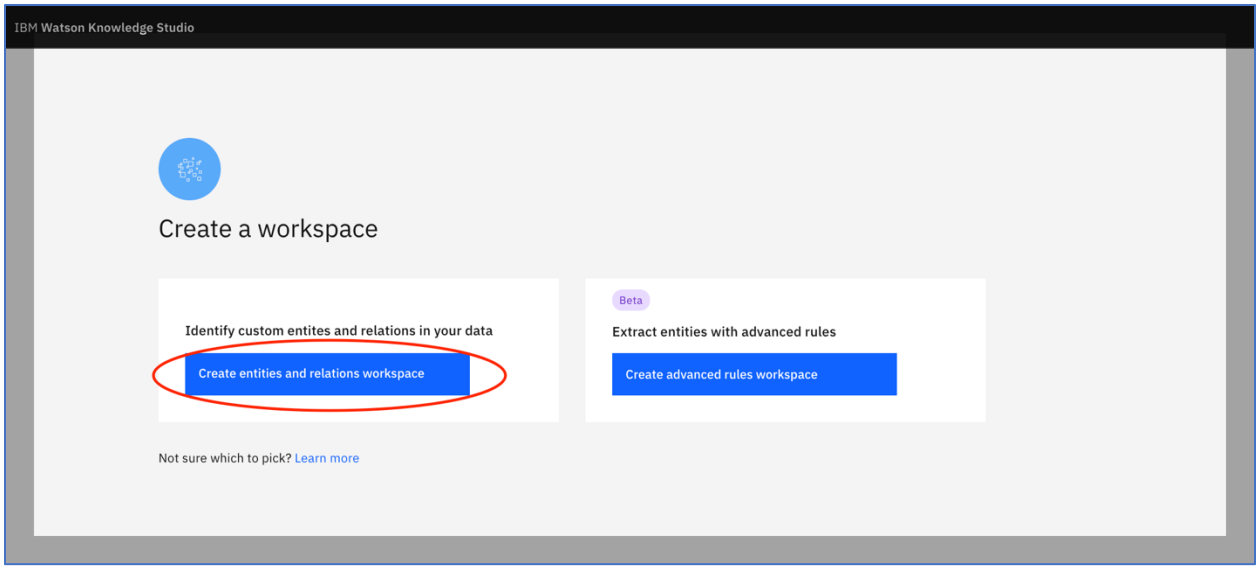
2. Under Services, click on your Knowledge Studio.



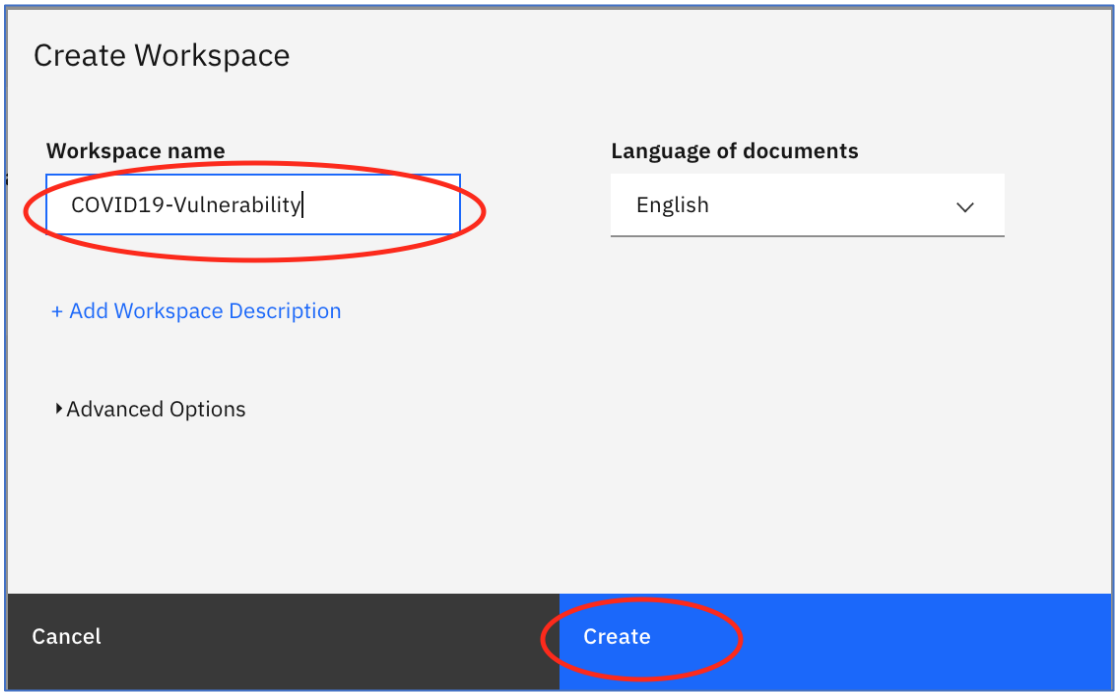
3. Click **Launch Knowledge Studio** to start your instance of Watson Knowledge Studio.



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

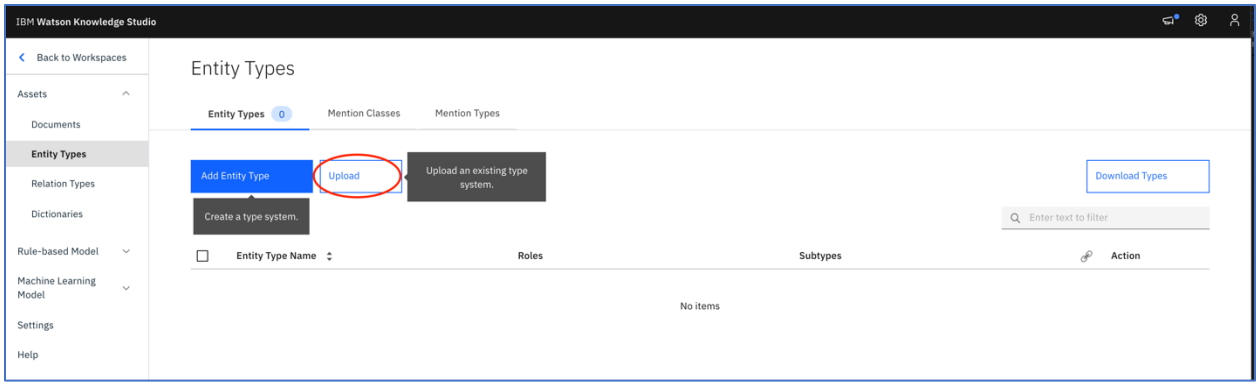


5. Type **COVID19-Vulnerability** for the Workspace name and click **Create**.

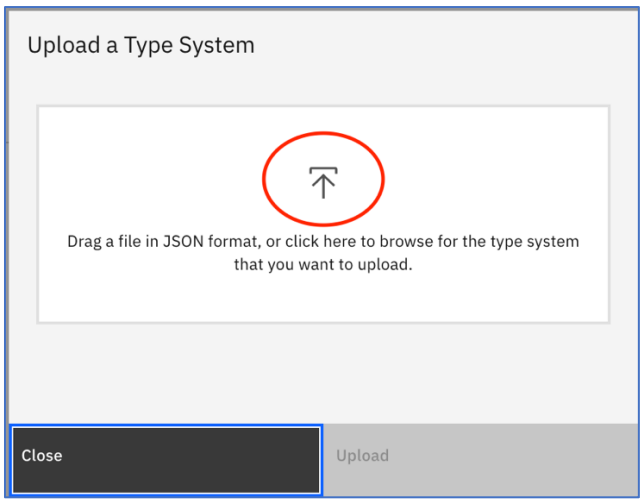


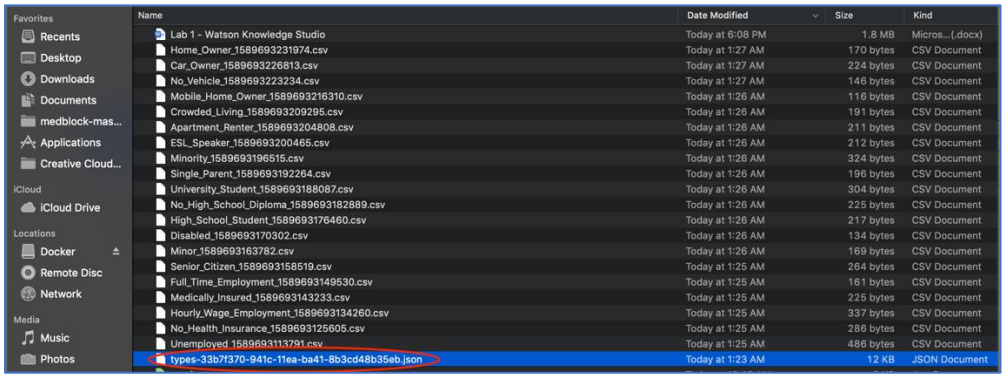
Inside of this workspace, we will create a type system consisting of the custom entities of the COVID-19 vulnerability index, create a dictionary, perform manual annotation and upload a training corpus for the development of the entity recognition machine learning model.

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**.

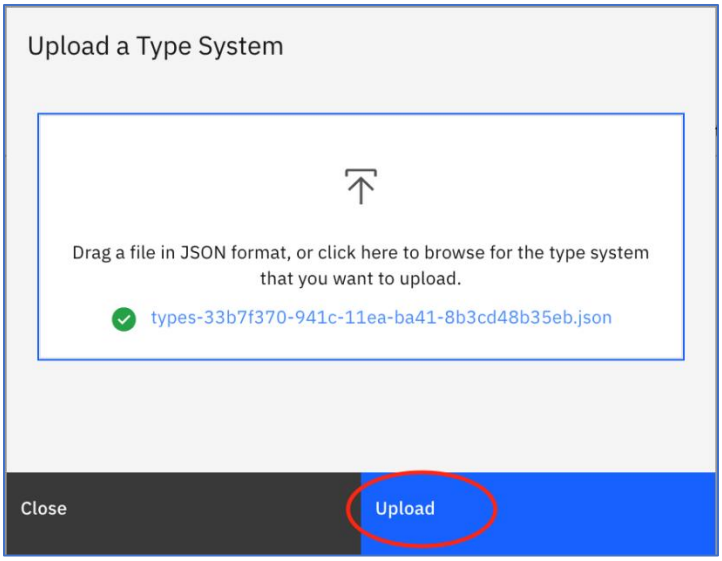


7. Click on the upload icon and navigate to the folder where you extracted the zip file downloaded from the GitHub repository. Select **types-33b7f370-941c-11ea-ba41-8b3cd48b35eb.json**.

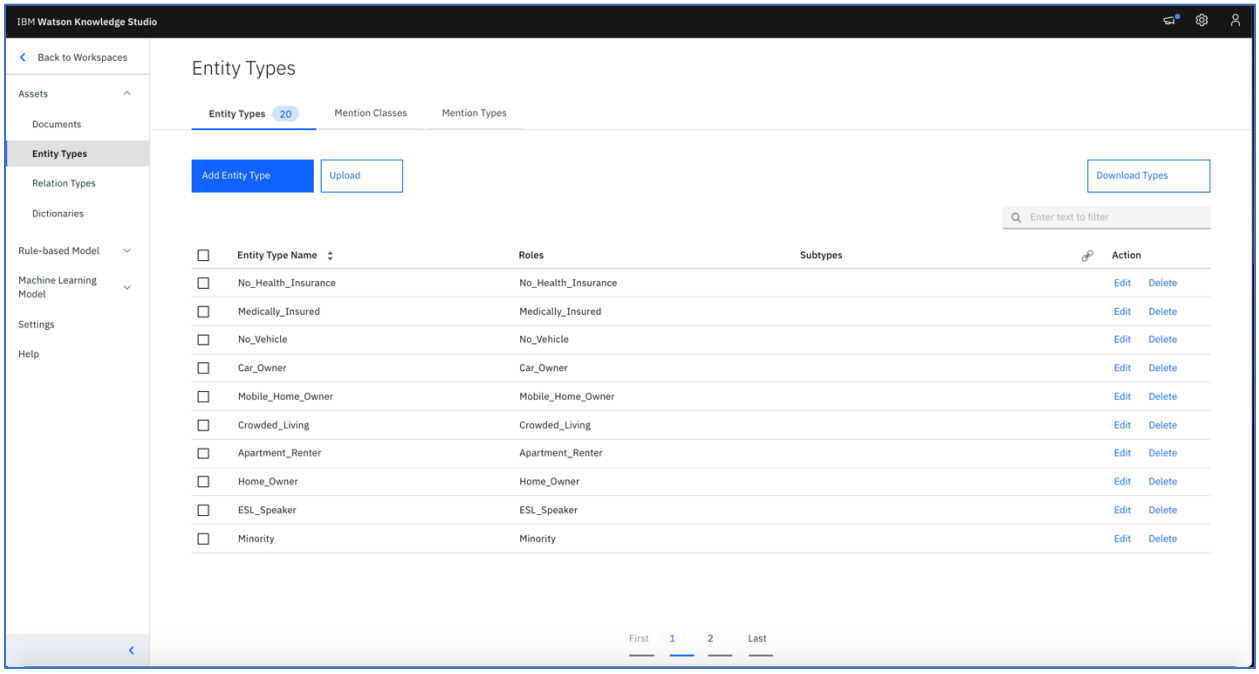




8. Click on Upload.



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.

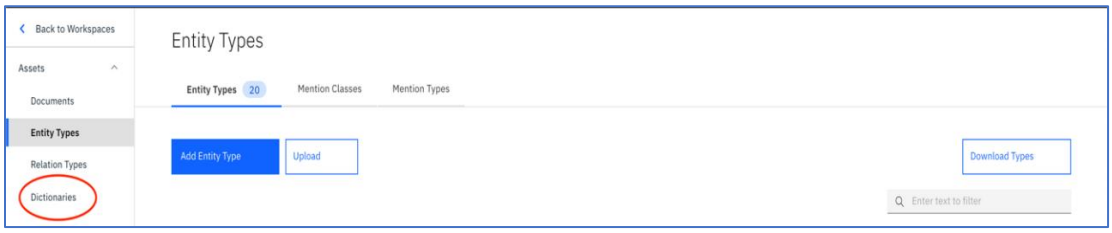


Exercise 4: Create a Dictionary

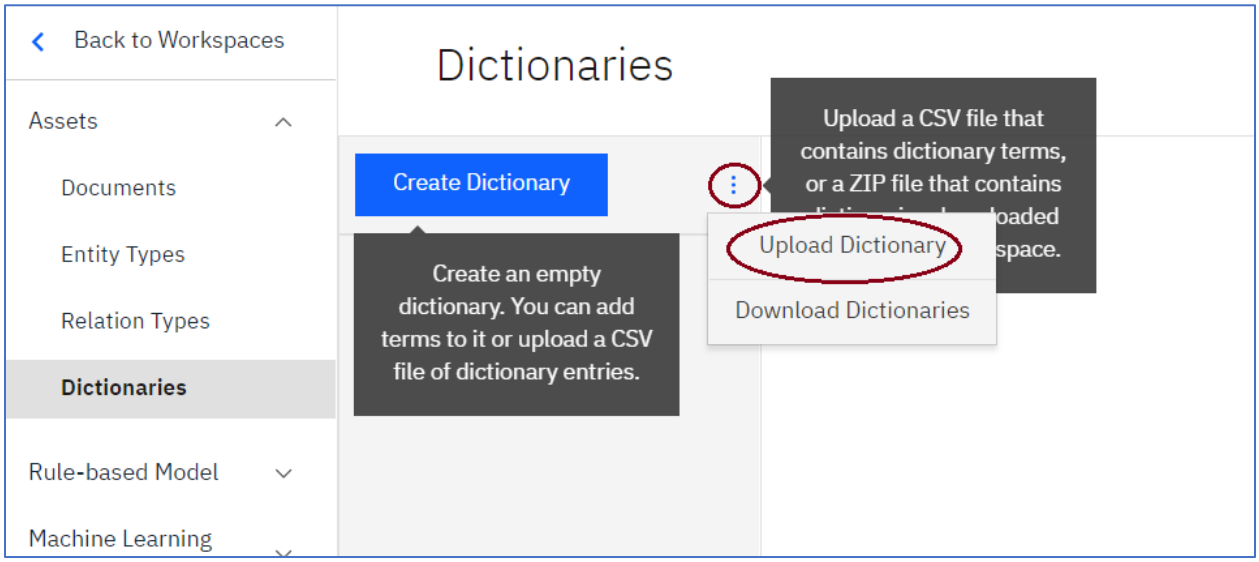
To help with manual annotation (which we will tackle in the next exercise), we will create a dictionary for each of the entity types in our type system. A dictionary is a list of words or phrases that are equivalent for information-extraction purposes, meaning that they are

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.

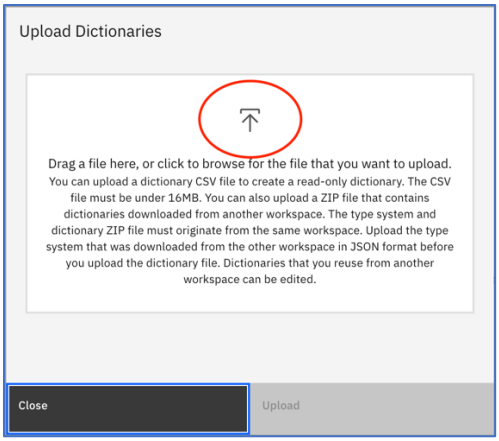
1. Under Assets, click **Dictionaries**.

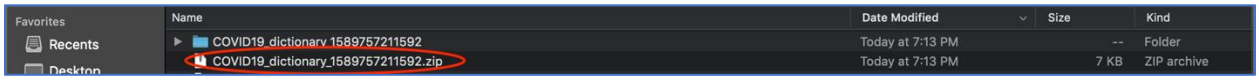


2. On the Dictionaries page, we can upload the zip file containing dictionaries for all our entity types. Click on the **vertical dots icon** and select **Upload Dictionary**.

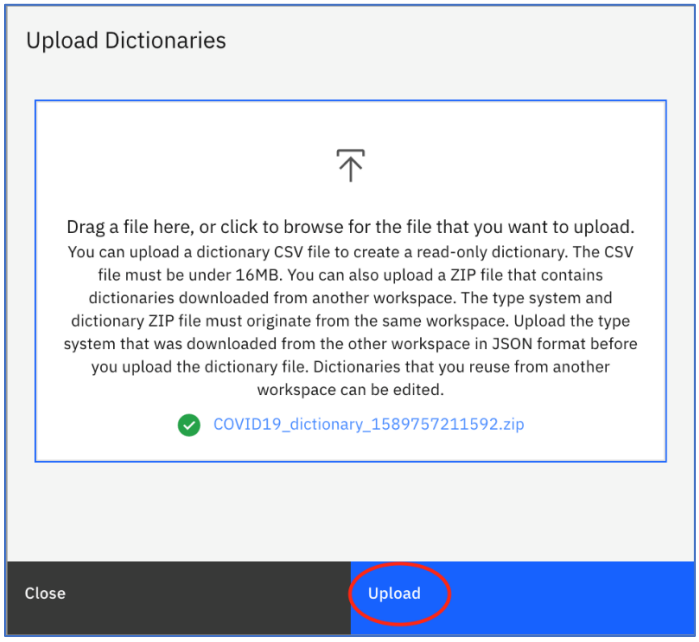


3. Click on the **Upload icon** and navigate to the folder where you extracted the zip file downloaded from the GitHub repository. Select the **COVID19_dictionary_1589757211592.zip** file.

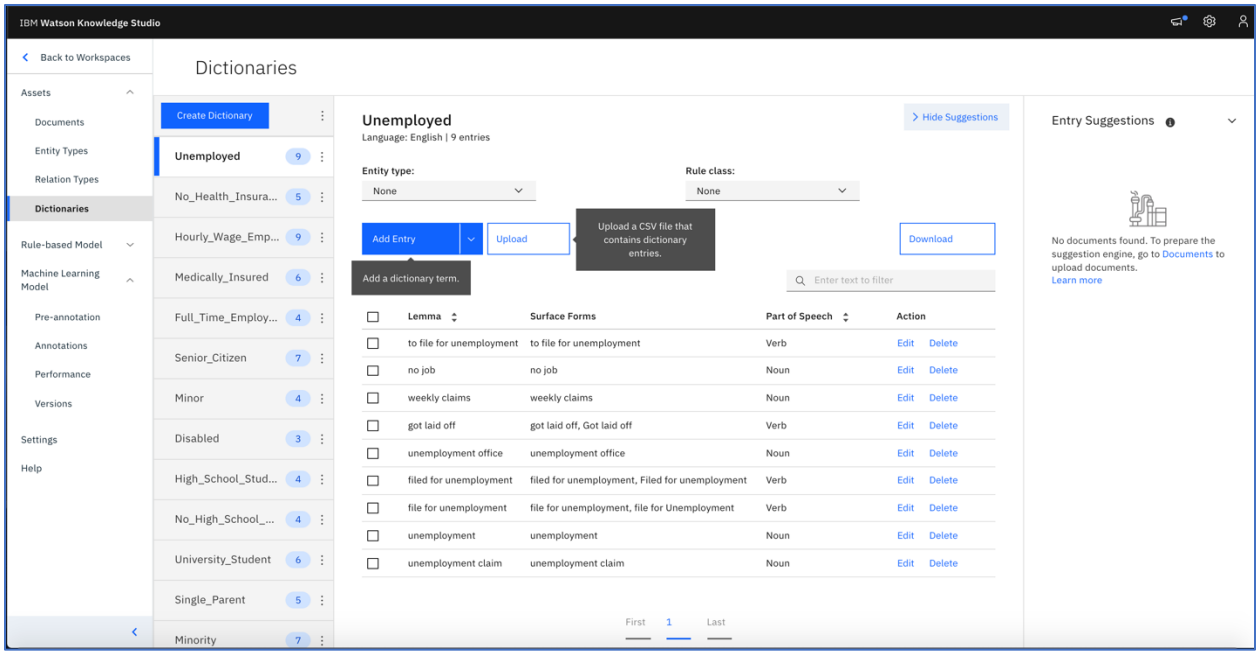




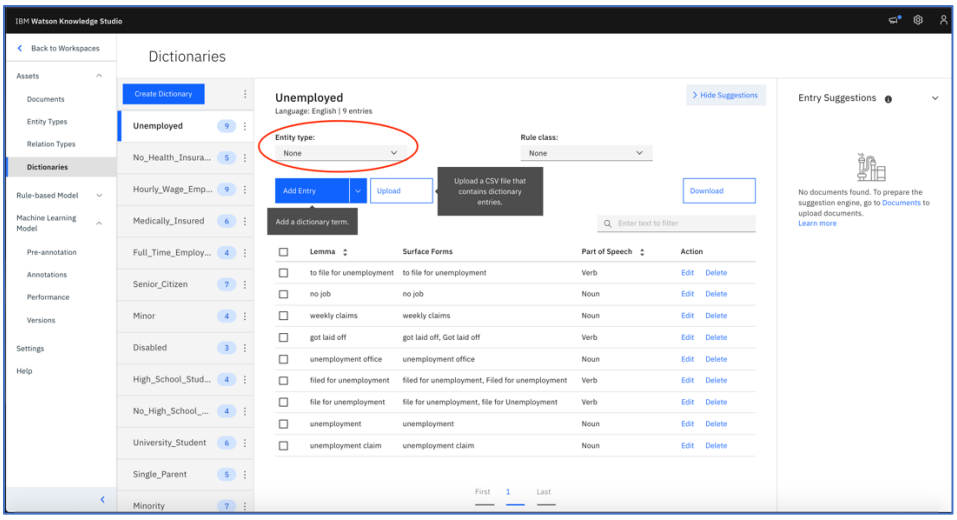
4. Click on **Upload**.



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.

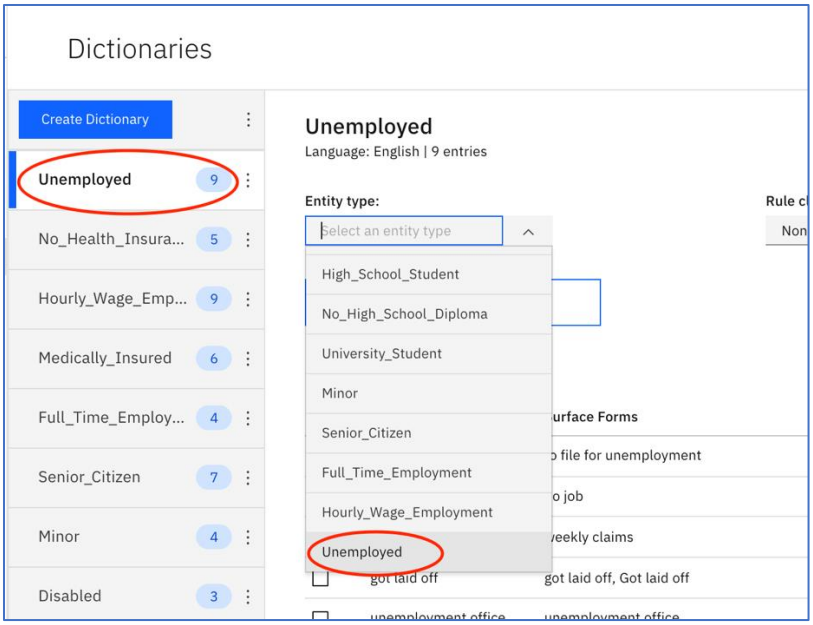


In order to save these dictionaries as a pre-annotator, we need to match each dictionary with its corresponding entity type. For example, the Unemployed dictionary, which currently has an entity type of None needs to be matched to the Unemployed entity type.



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

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

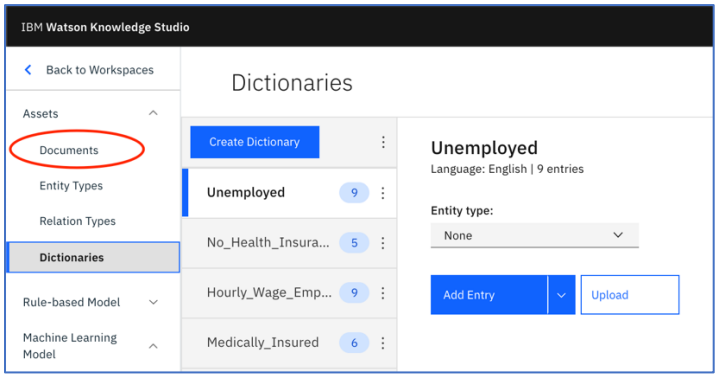


Repeat the same process for each dictionary until all 20 dictionaries are matched to their corresponding entity type (none of the dictionaries should have an Entity type of None).

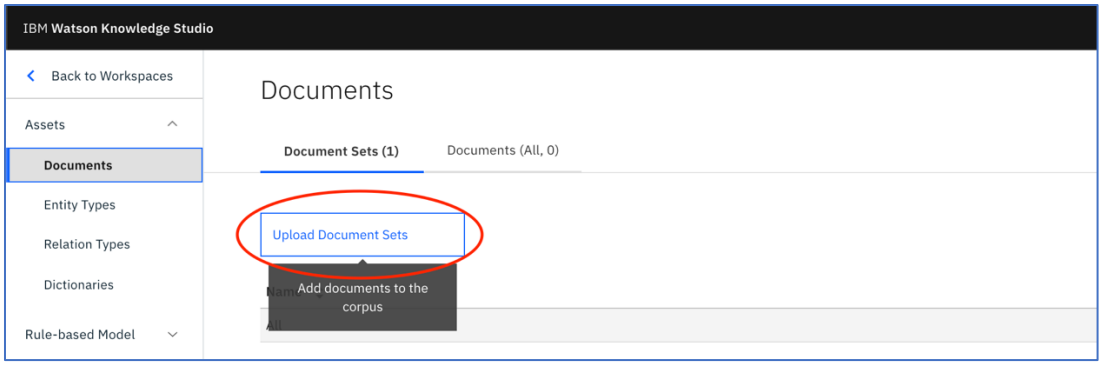
Exercise 5: Upload a corpus of documents

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.

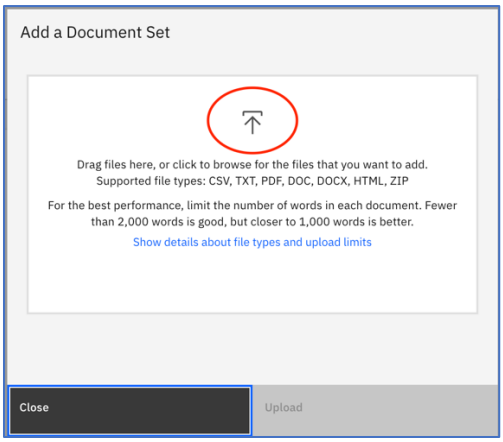
1. Under Assets, select **Documents**.



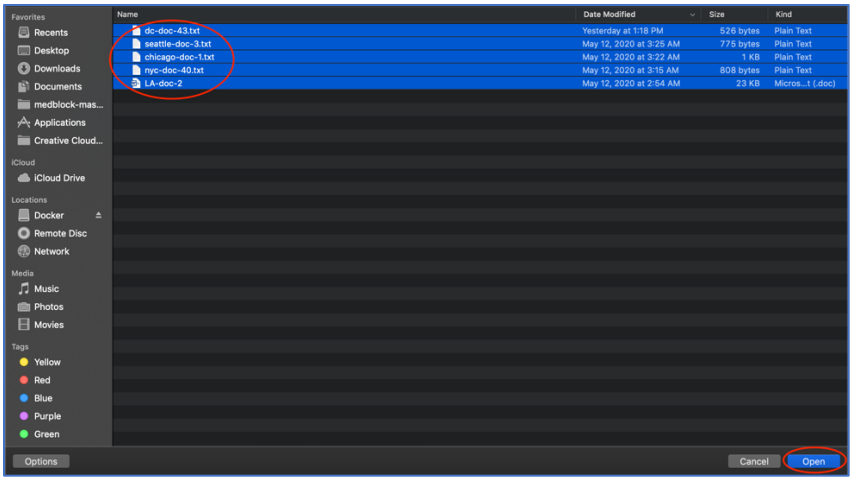
2. Click **Upload Document Sets**.



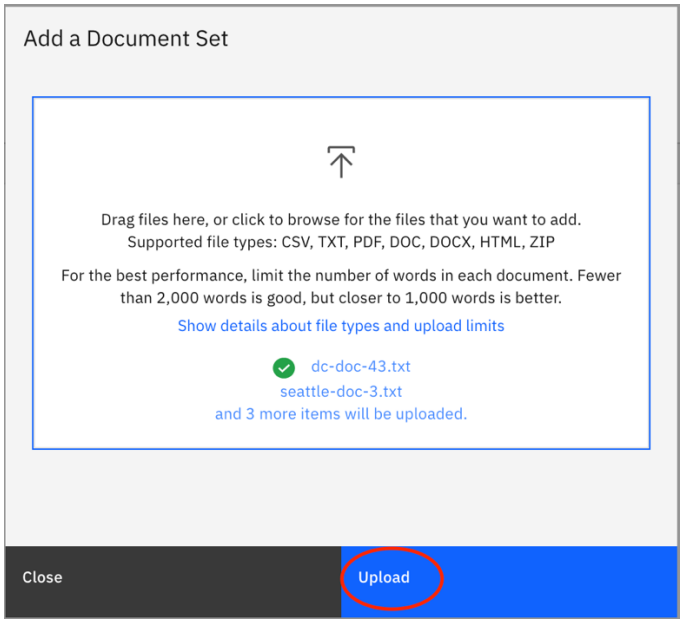
3. Click on the **Upload icon** and navigate to the folder where you extracted the zip file downloaded from the GitHub repository. Double-click on the **SampleDocs** folder**.**



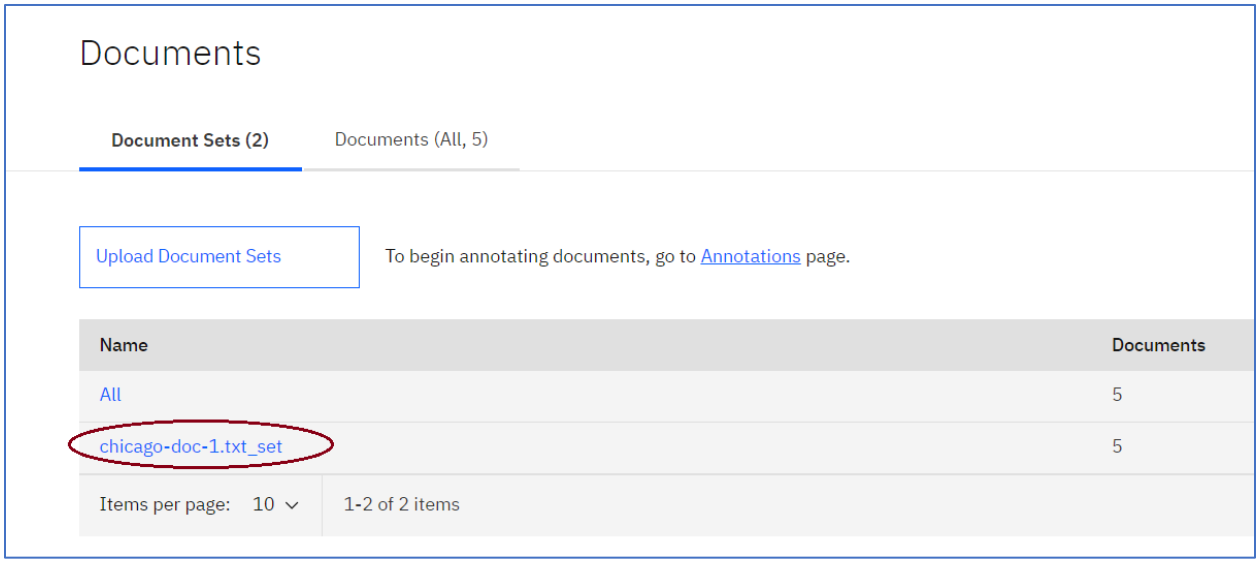
4. **Shift select** all 5 documents in the folder (don't select the README file) and click **Open**.



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.

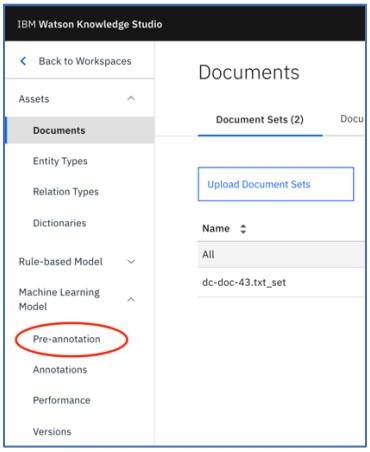


Exercise 6: Perform Manual Annotation

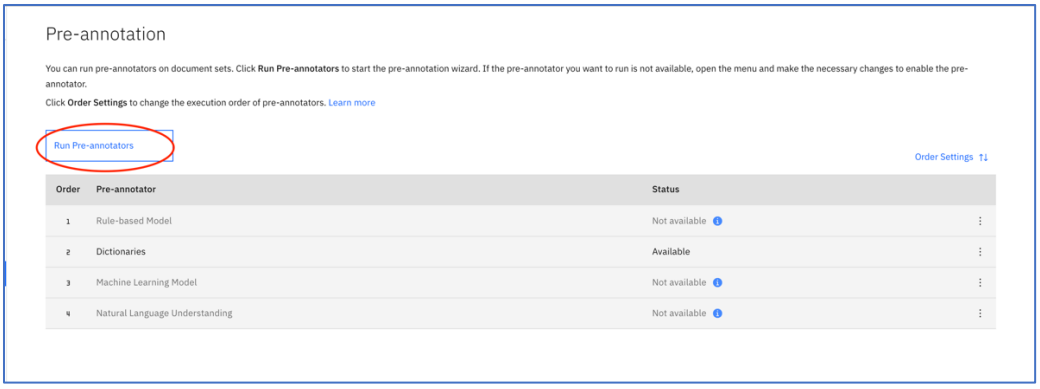
To create an entity recognition model, we will need to teach Watson about our custom entity types by manually annotating a sample corpus of documents.

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**.



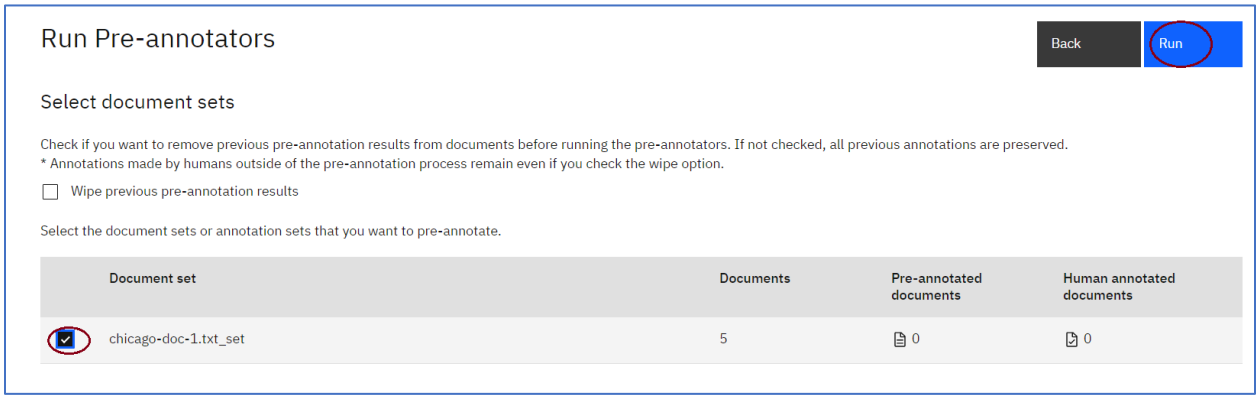
2. Click **Run Pre-annotators.**



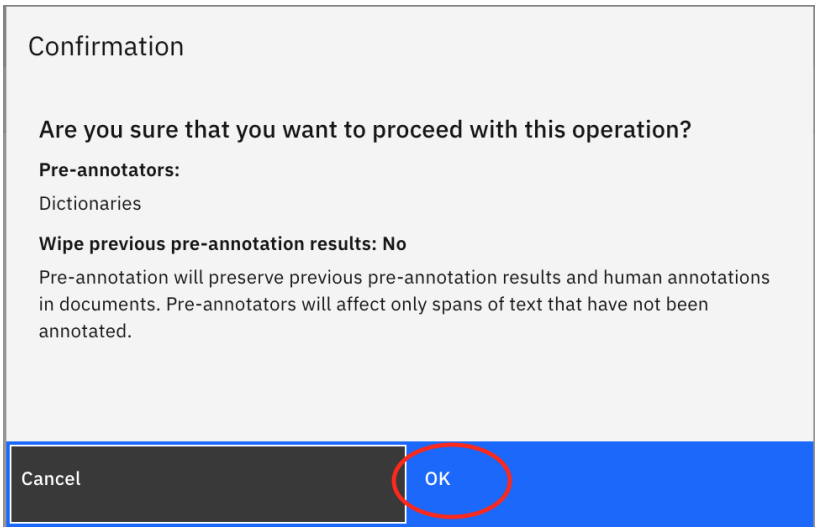
You should be able to see that Dictionaries is available as a pre-annotator. If you do not see any available pre-annotators in the table, please revisit Exercise 4, step 5 to match each dictionary with its corresponding entity type.

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

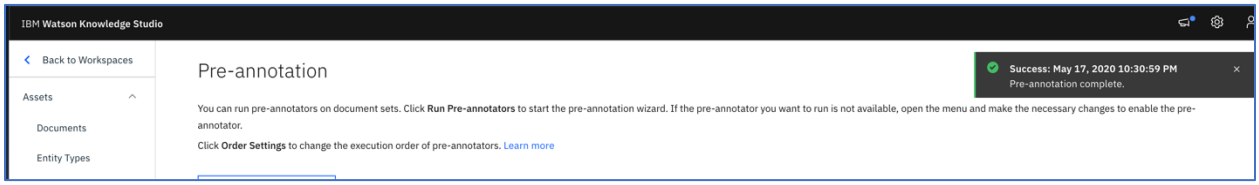
4. Under Select document sets, click the **checkbox** next to chicago-doc-1.txt_set and click **Run.**



5. Click **OK**.



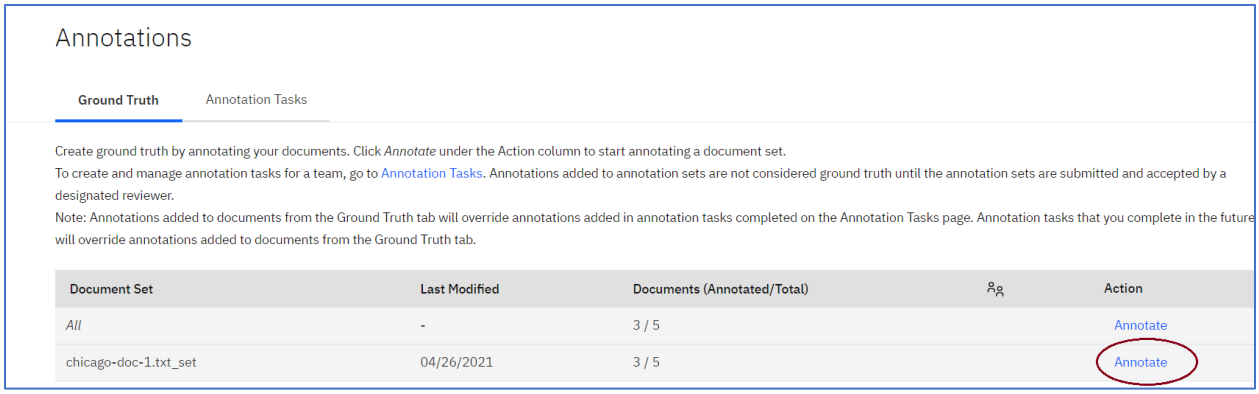
After a few seconds, pre-annotation will be complete and you will see the following success message:



6. Under Machine Learning Model, click **Annotations**.

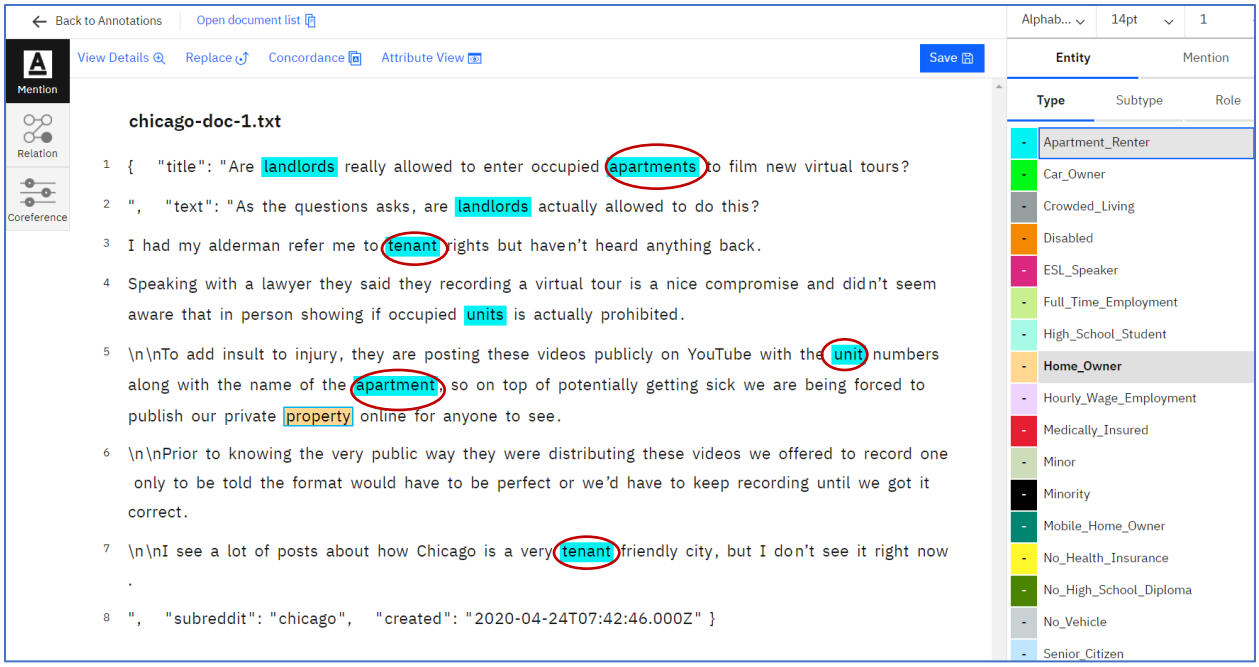
On the Annotations screen, you will see that Watson used the dictionary pre-annotator to annotate 3 of the 5 documents. We will now manually annotate all 5 documents. When annotating each document, we will highlight any mention of the custom entity types in each social media post.

7. Click the **Annotate** link on the chicago-doc-1.txt_set row.

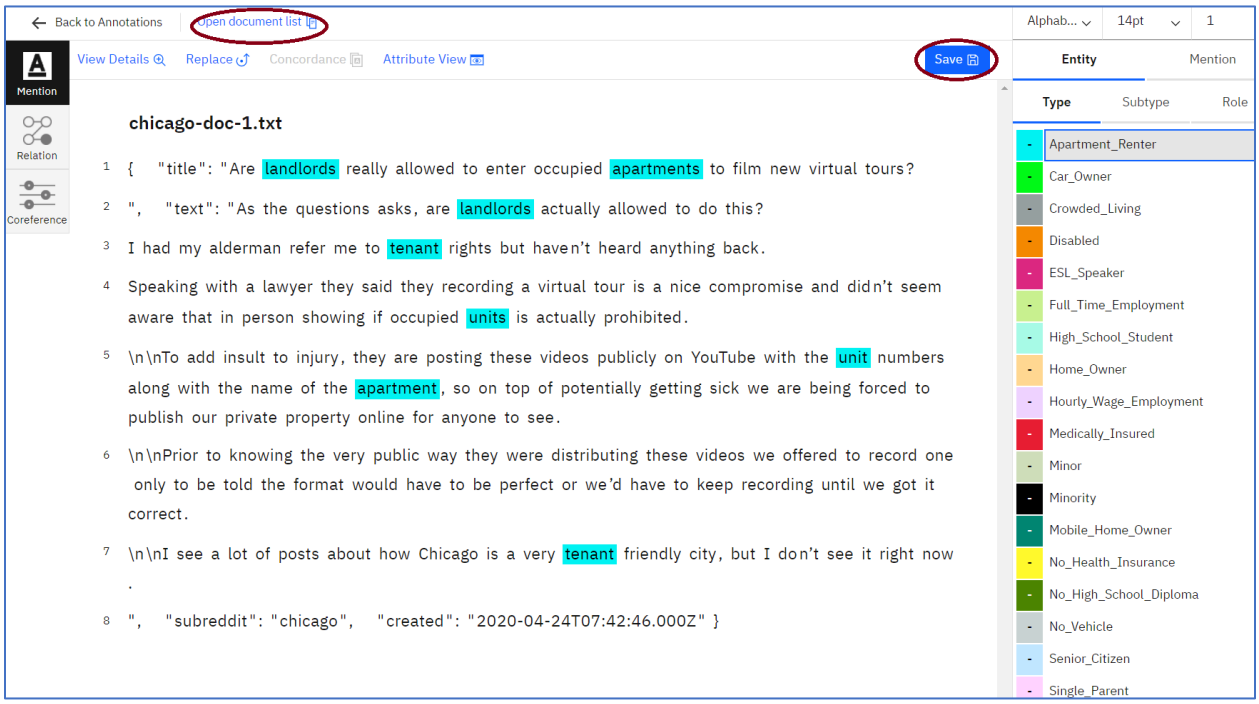


8. On the Select Document screen, click on chicago-doc-1.txt.

9. 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.



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.



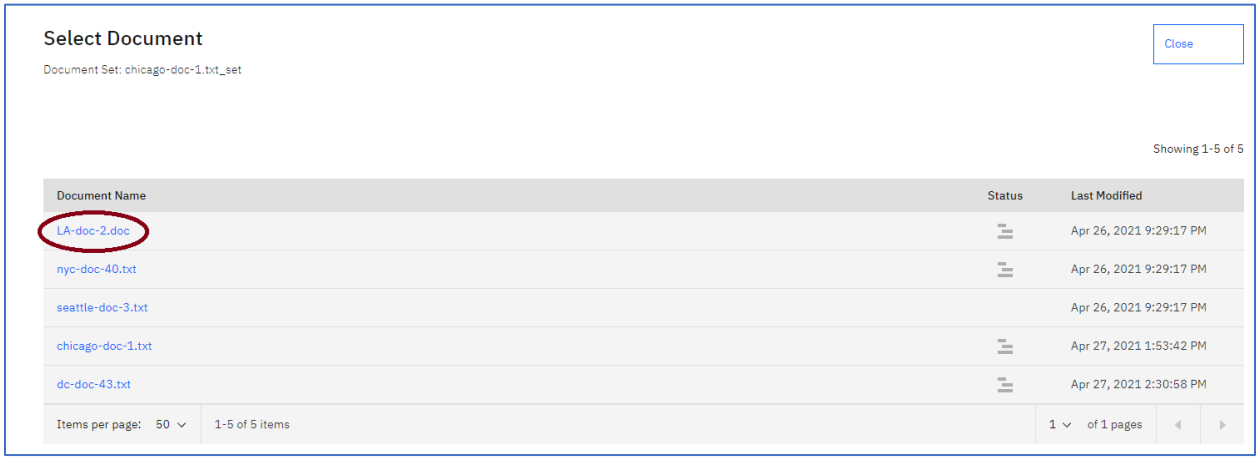
11. On the **Select Document** panel, click on **dc-doc-43.txt**.

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:

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

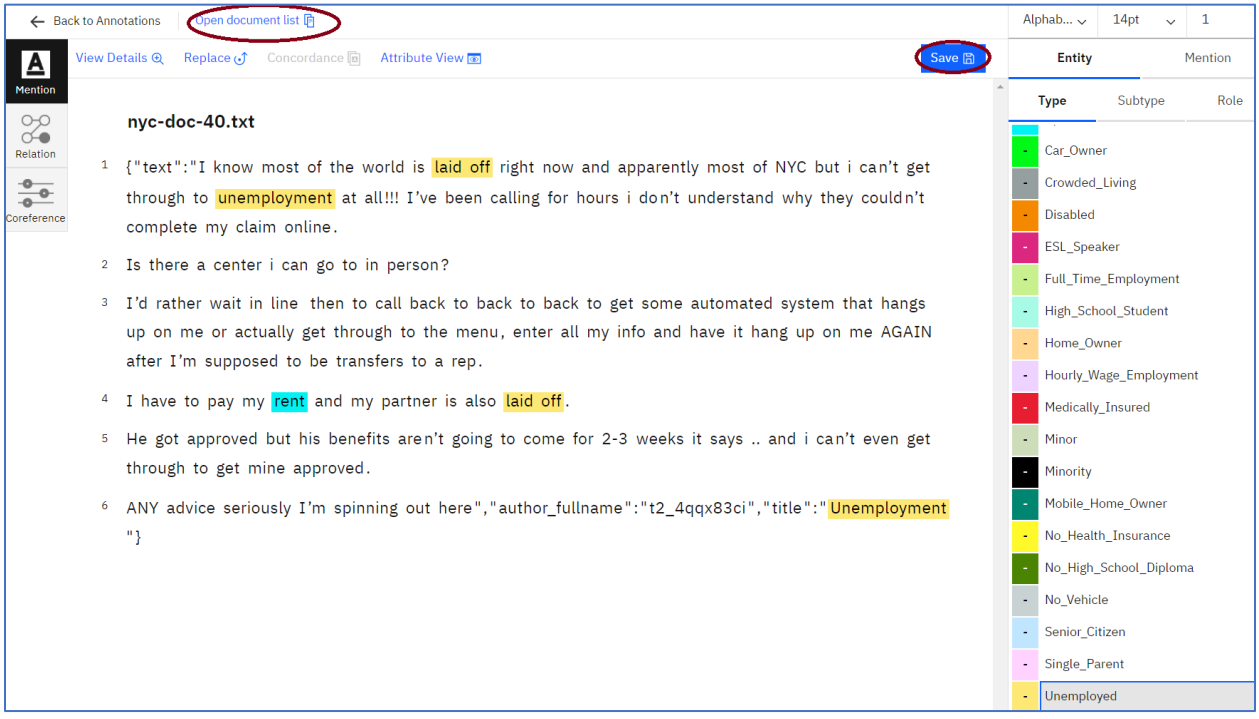


15. Annotate as shown below, and then click the **Save**  **icon** and **Open document list**.

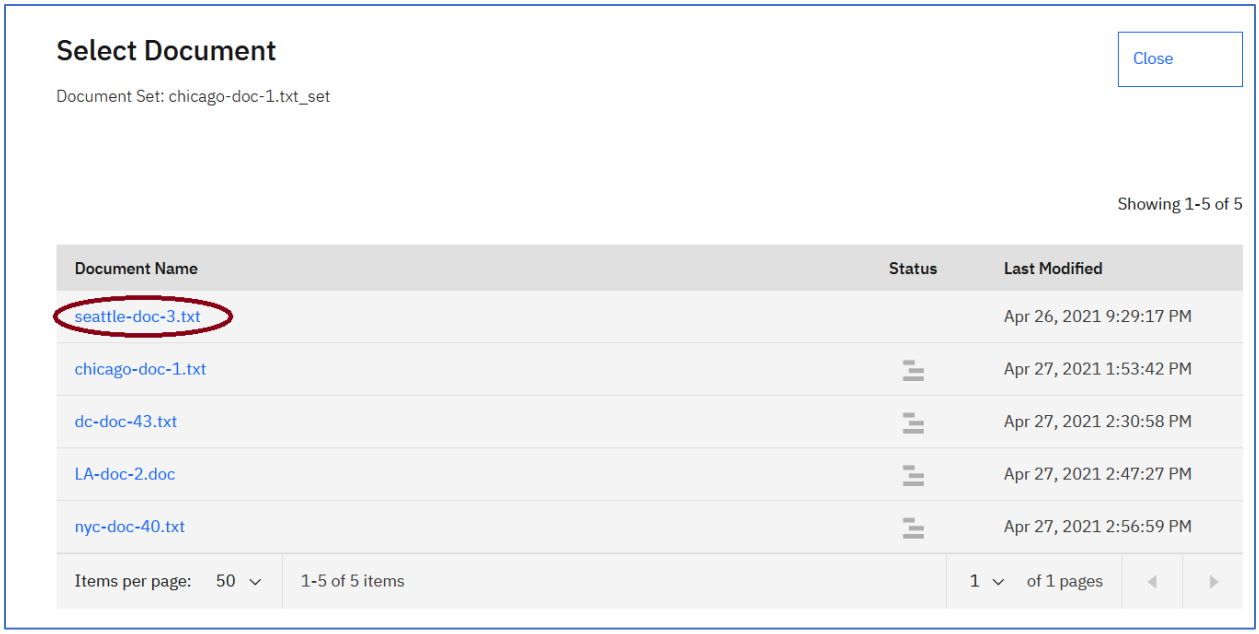
16. On the **Select Document** panel, click on **nyc-doc-40.txt**.

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.

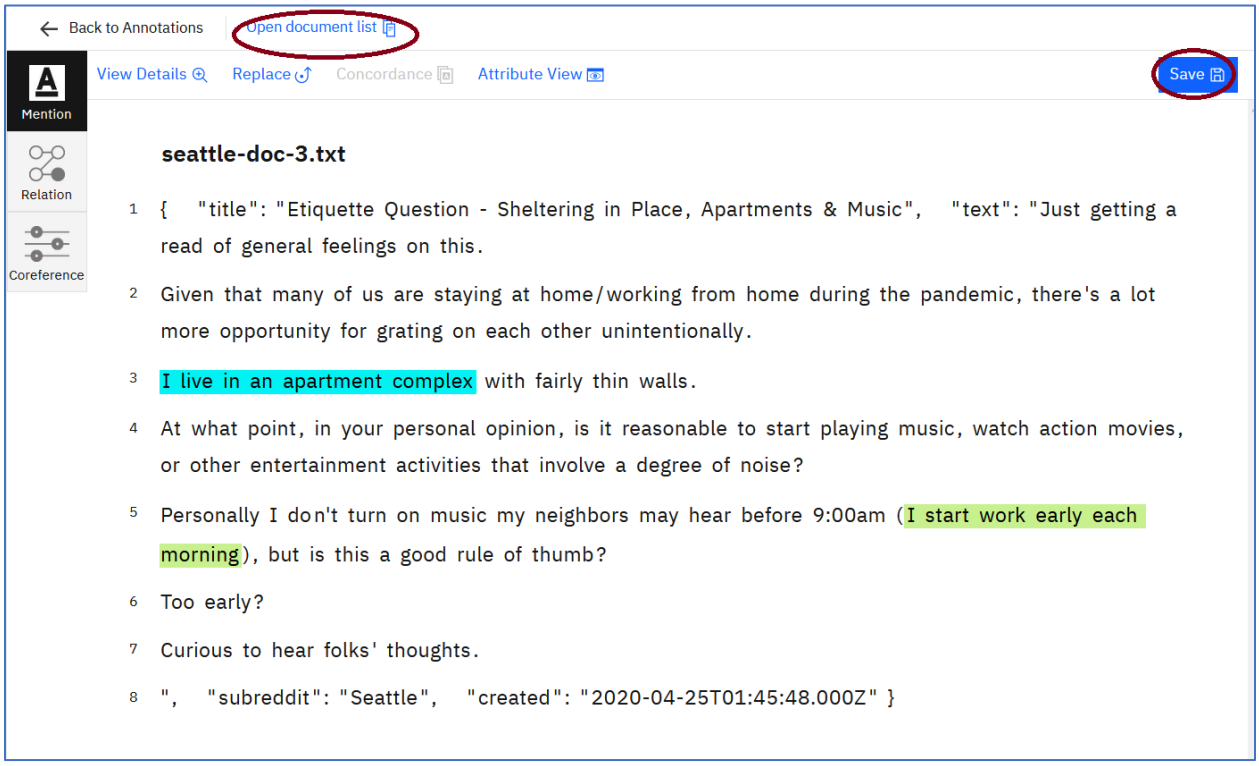


18. On the **Select Document** panel, click on **seattle-doc-3.txt**



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.



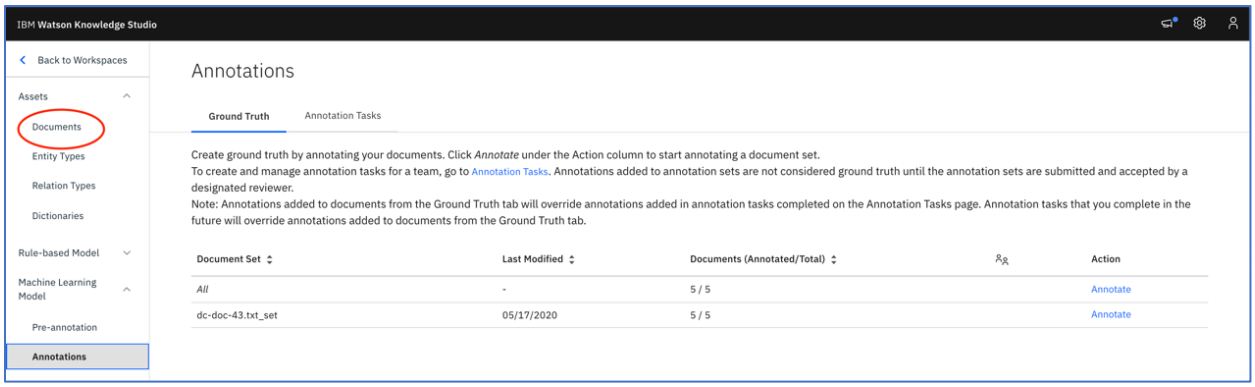
All 5 documents have now been manually annotated. However, we will need a much larger set of documents in order to train and create a machine learning model. In the next exercise, we will upload the complete corpus of documents and create an entity recognition model.

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

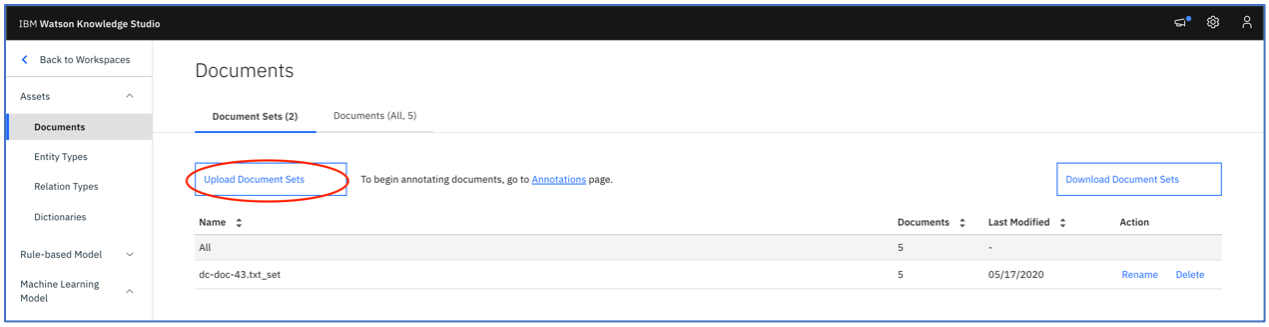
Exercise 7: Train and create a machine learning (ML) annotator

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.

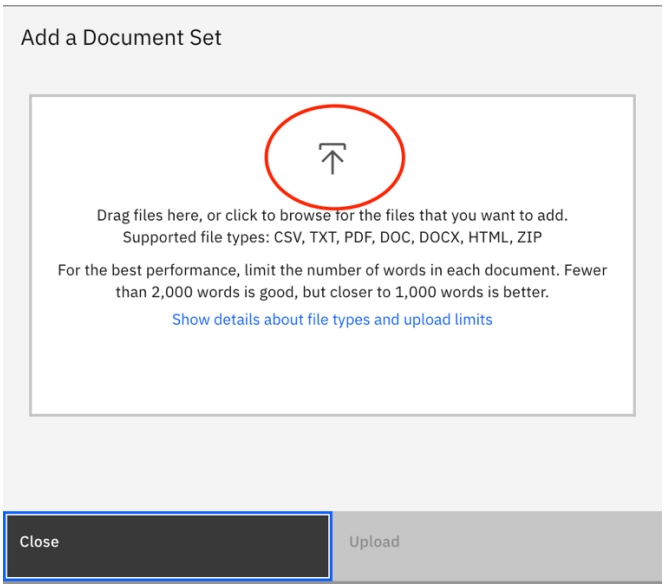
1. Under Assets, click **Documents**.



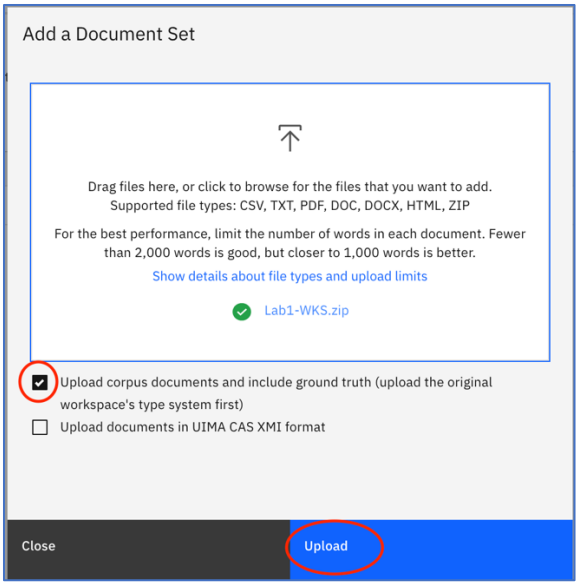
2. On the Documents screen, click **Upload Document Sets**.



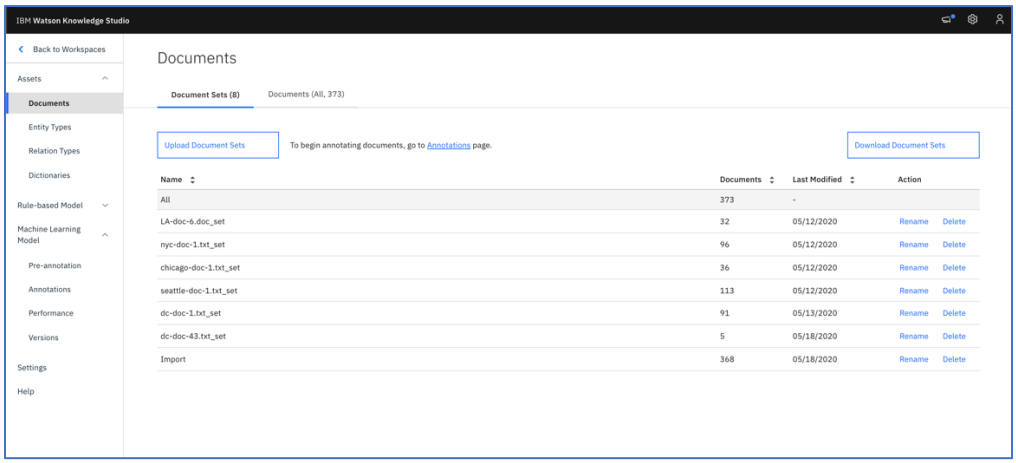
3. Click on the **Upload icon** and navigate to the folder where you extracted the zip file downloaded from the GitHub repository. Select the **Lab1-WKS.zip** file and click **Open**.



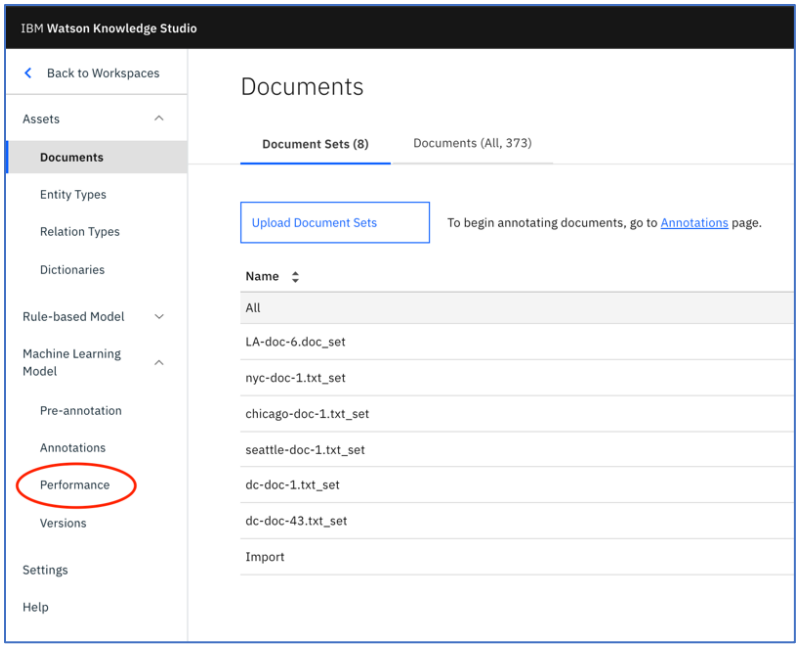
4. Click the box next to **Upload corpus documents and include ground truth (upload the original workspace's type system first)** and click **Upload**.



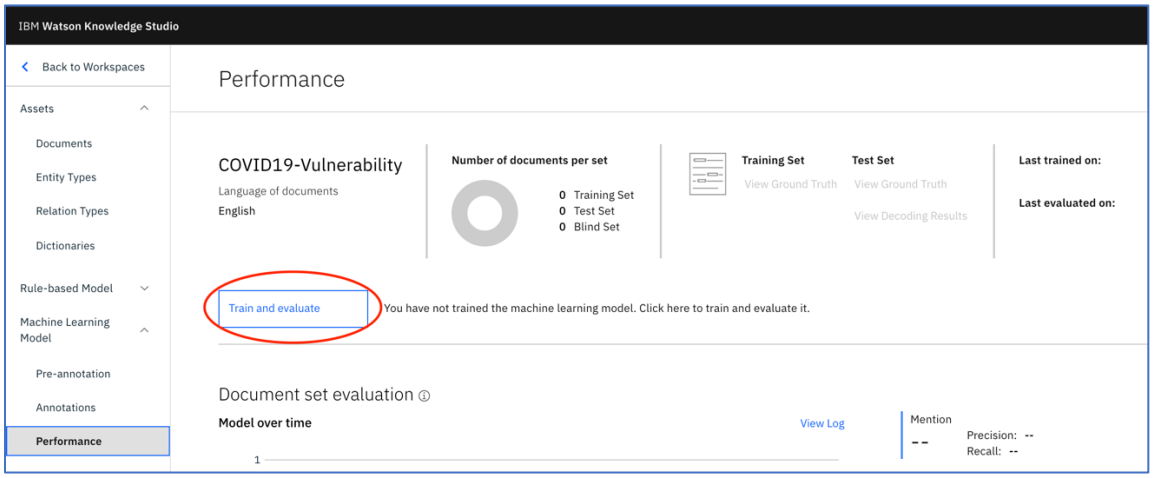
You should now see several new document sets on the Documents screen including an Import document set consisting of 368 posts that were just now added to the workspace. We will be using these newly uploaded documents to train and create a ML annotator.



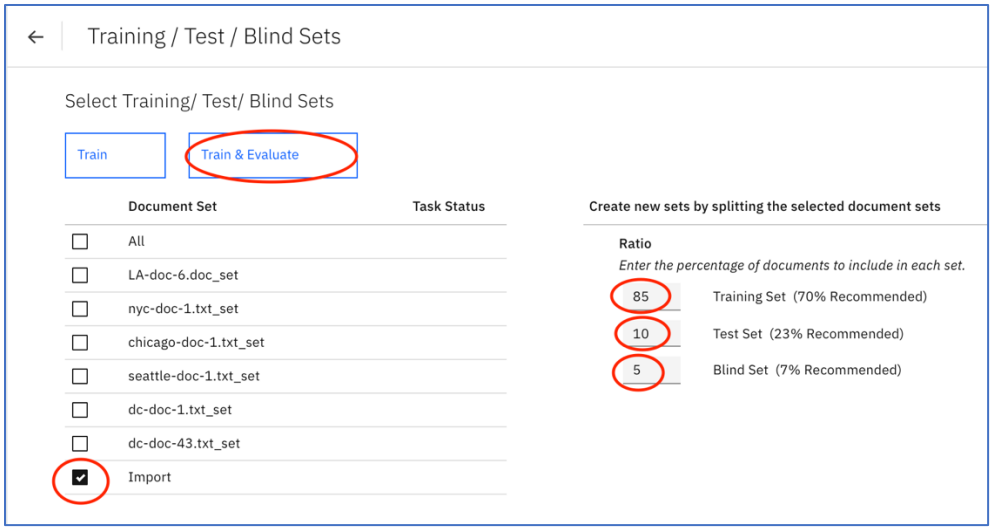
5. Under Machine Learning Model, click on **Performance**.



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

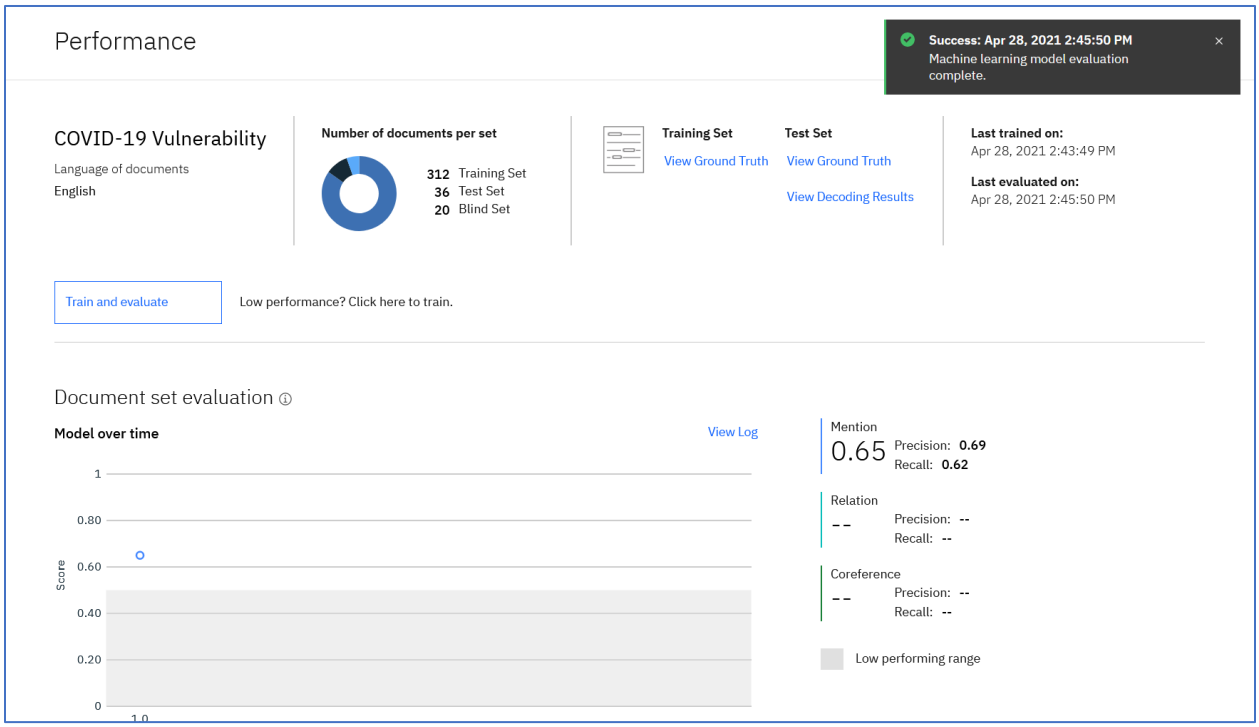


7. On the Select Training/Test/Blind Sets screen, choose **Import**, change the **Training Set** percentage to 85%, **Test Set** to 10% and **Blind Set** to 5%. Click **Train & Evaluate**.



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:

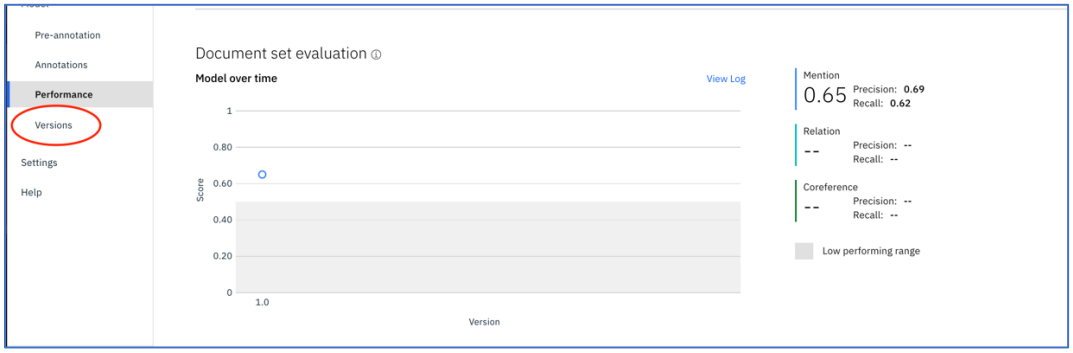


Exercise 8: Save and Deploy the ML Annotator to Discovery

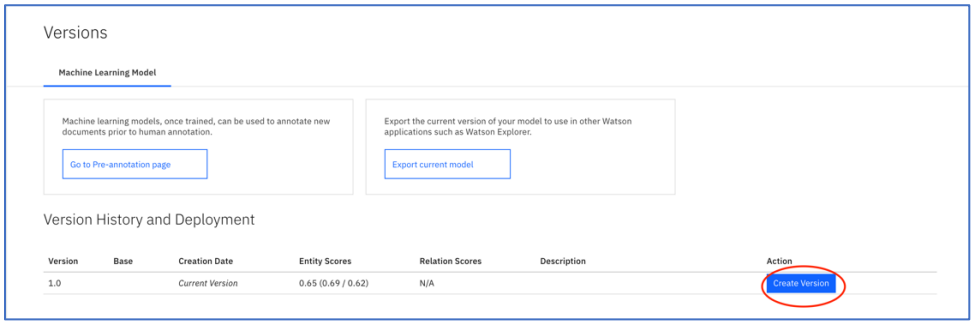
Now that we have a machine learning annotator, we can use to automatically perform entity extraction inside of Watson Discovery. The automated entity extraction of social media posts for all 5 cities will get us closer to determining the social vulnerability index of each city.

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

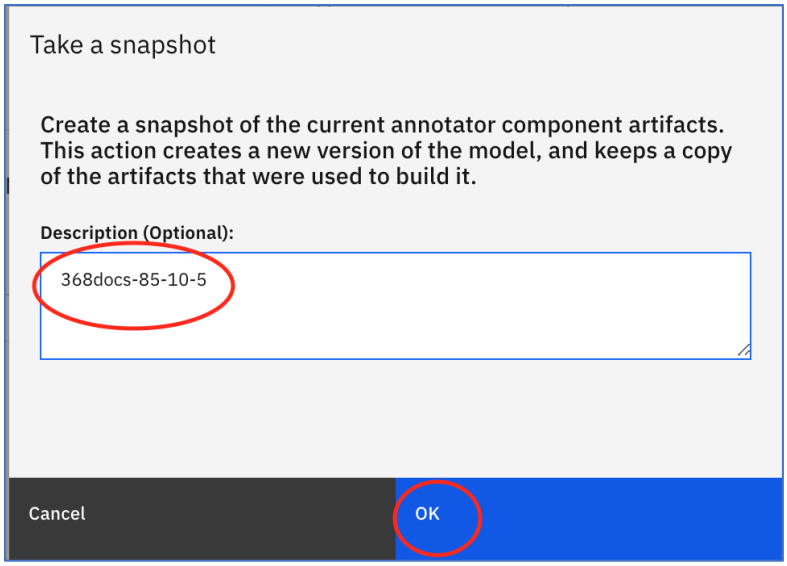
1. Under Machine Learning Model, click on **Versions**.



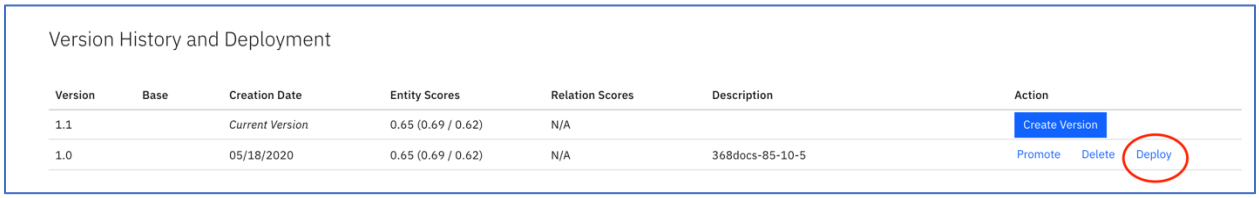
2. On the Versions page, click **Create Version**.



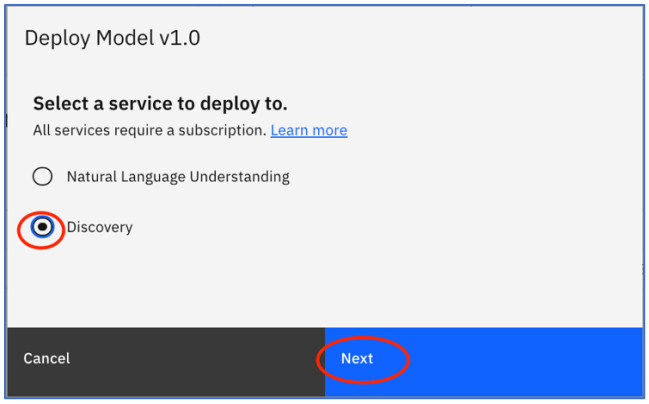
3. Type **368docs-85-10-5** (to distinguish this as an entity model using 368 docs with an 85-10-5 split) under Description and click **OK**.



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

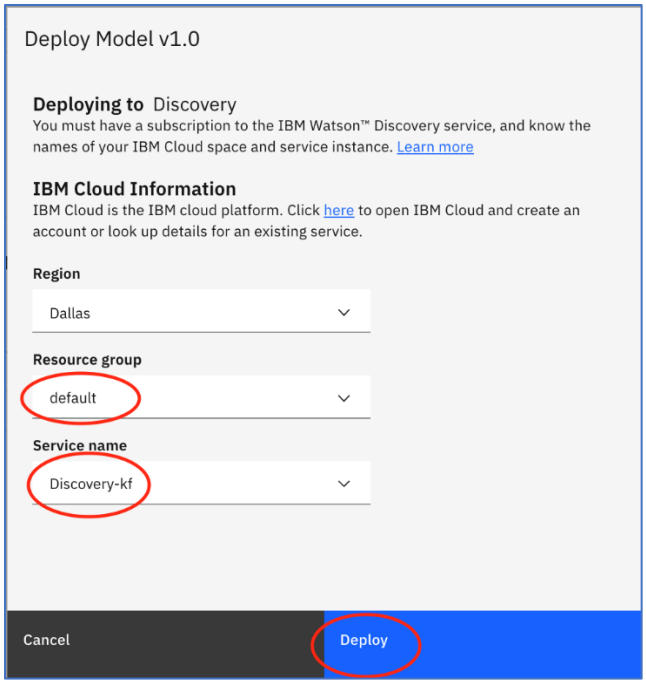


5. Select **Discovery** and click **Next**.

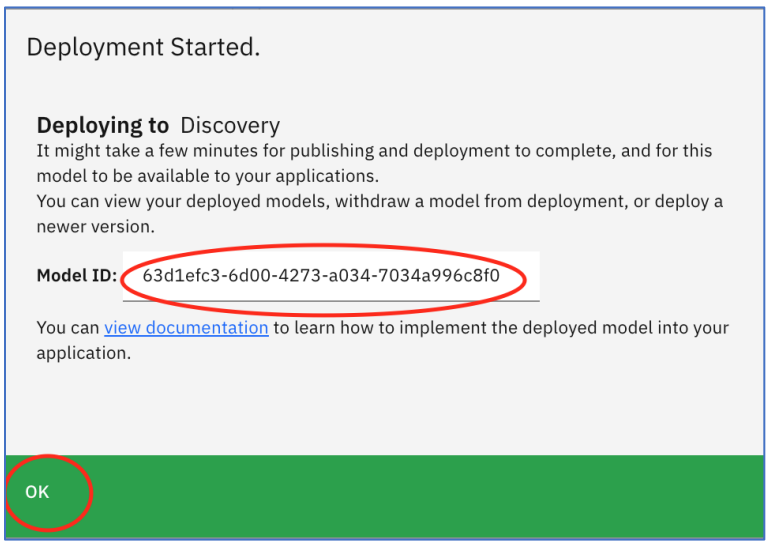


6. In order to deploy this model to your Discovery instance, you will need to select the resource group containing your instance as well as the Service name of the instance that you created. If this is your first time working with the Watson APIs on the IBM Cloud, you should only have one instance of Discovery currently provisioned.

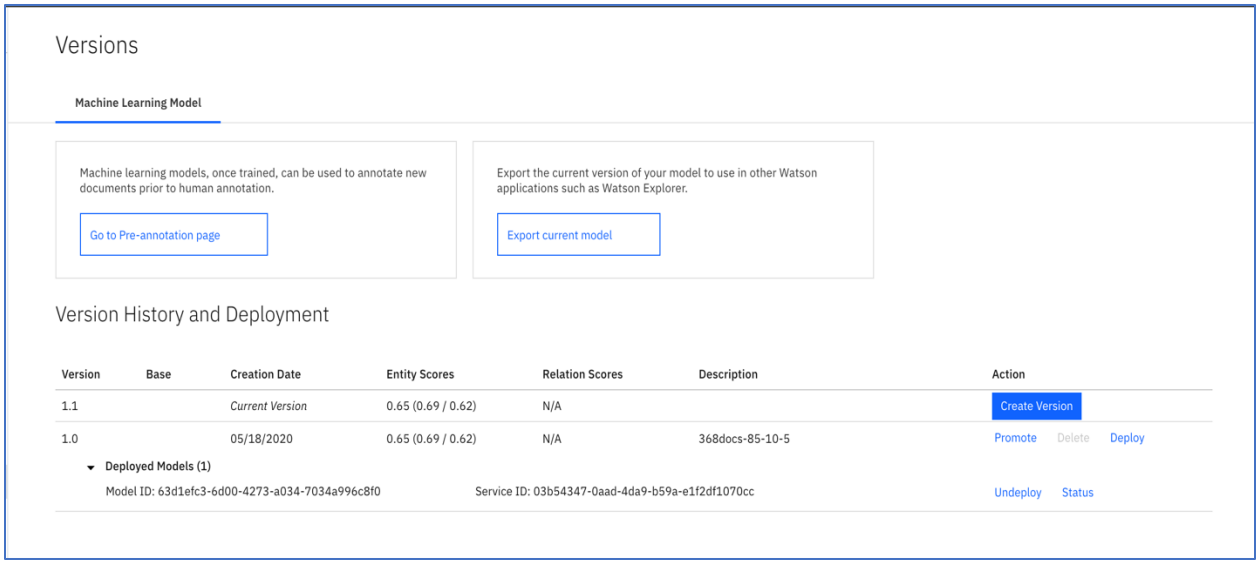
   Select **default** from the drop-down menu under **Resource group** and the **name of the Discovery instance** under **Service name**.



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



8. 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.



You have completed Lab 1!