**IBM Training**

Student Exercises

Lab-1: Develop Socioeconomic Annotators for COVID-19

Hands-On Lab

|  |  |
| --- | --- |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |

## 

Table of Contents

[Introduction 3](#_Toc40664714)

[Objectives 3](#_Toc40664715)

[Exercise 1: Create a Watson Knowledge Studio Instance 3](#_Toc40664716)

[Exercise 2: Create a Watson Discovery Instance 6](#_Toc40664717)

[Exercise 3: Create a Type System 7](#_Toc40664718)

[Exercise 4: Create a Dictionary 11](#_Toc40664720)

[Exercise 5: Upload a corpus of documents 13](#_Toc40664721)

[Exercise 6: Perform Manual Annotation 16](#_Toc40664722)

[Exercise 7: Train and create a machine learning (ML) annotator 22](#_Toc40664726)

[Exercise 8: Save and Deploy the ML Annotator to Discovery 26](#_Toc40664727)

# 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.**

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

1. Click **Create Resource.**

A screenshot of a social media post

Description automatically generated

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

A screenshot of a cell phone screen with text

Description automatically generated

1. Click on **Knowledge Studio**.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

# Exercise 2: Create a Watson Discovery Instance

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

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a social media post

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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.

1. Although we can manually enter the entity types for our type system, we will instead upload the type system CSV file downloaded from the Box folder or the GitHub repository.

On the Entity Types screen, click **Upload**.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

A screenshot of a computer

Description automatically generated

# Click on Upload.

A screenshot of a cell phone

Description automatically generated

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.

A screenshot of a social media post

Description automatically generated

# 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**.

A screenshot of a cell phone

Description automatically generated

1. 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**.

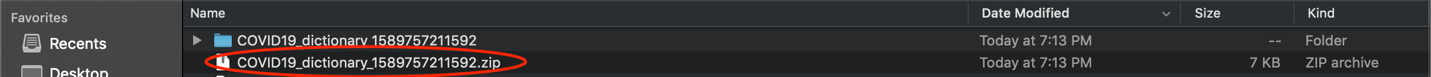
A screenshot of a cell phone

Description automatically generated

1. Click on the **Upload icon** and select the **COVID19\_dictionary\_1589757211592.zip** file.

A screenshot of a cell phone

Description automatically generated



1. Click on **Upload**.

A screenshot of a cell phone

Description automatically generated

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.

A screenshot of a computer

Description automatically generated

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.

A screenshot of a computer

Description automatically generated

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

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

A screenshot of a cell phone

Description automatically generated

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**.

A screenshot of a cell phone

Description automatically generated

1. Click **Upload Document Sets**.

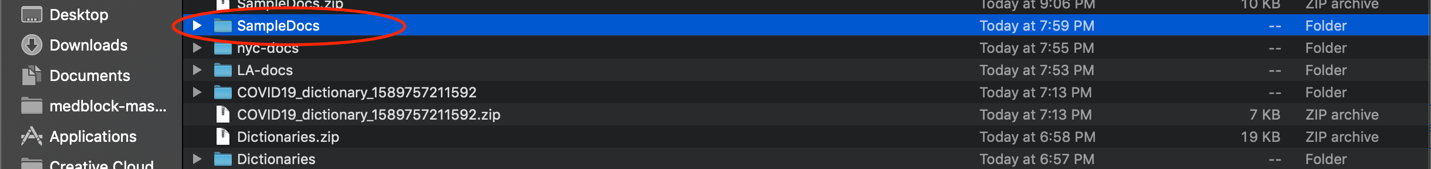
A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated



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

A screenshot of a computer screen

Description automatically generated

1. Click **Upload**.

A screenshot of a cell phone

Description automatically generated

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.

A screenshot of a cell phone

Description automatically generated

# 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**.

A screenshot of a cell phone

Description automatically generated

1. Click **Run Pre-annotators.**

A screenshot of a cell phone

Description automatically generated

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.

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

A close up of a logo

Description automatically generated

1. Under Select document sets, click the **checkbox** next to dc-doc-43.txt\_set and click **Run.**

A screenshot of a social media post

Description automatically generated

1. Click **OK**.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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

A close up of a logo

Description automatically generated

# 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.

1. Click the second **Annotate** link on the dc-doc-43.txt\_set row.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

1. 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**.”

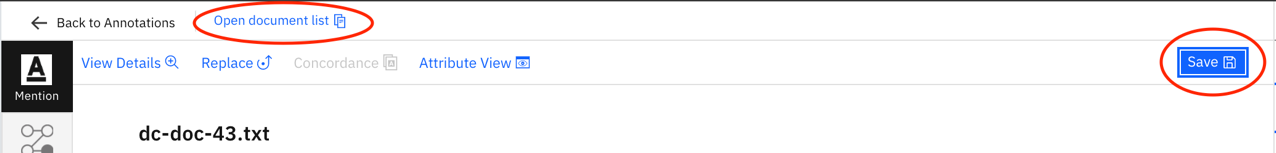
To highlight a sentence with an entity type, click on the first word of the sentence and then the last word before clicking on the corresponding entity type on the right, which in this post is only University\_Student.

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

A screenshot of a cell phone

Description automatically generated

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



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

A screenshot of a cell phone

Description automatically generated

1. 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.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

1. 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.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

1. 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.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

1. 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.

A screenshot of a cell phone

Description automatically generated

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.

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

A screenshot of a cell phone

Description automatically generated

# 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.

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

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

A close up of a logo

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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.

A screenshot of a social media post

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

1. 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**.

A screenshot of a cell phone

Description automatically generated

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:

A screenshot of a cell phone

Description automatically generated

# 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**.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

1. 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**.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

1. Select **Discovery** and click **Next**.

A screenshot of a cell phone

Description automatically generated

1. 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**.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

1. 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.

A screenshot of a cell phone

Description automatically generated

**You have completed Lab 1!**