

Natural Language Understanding

Requirements:

1. Bluemix account; [sign up for a trial account](#)
2. Knowledge studio account; [sign up for a trial account](#)
3. Python 2.7 or 3
4. [Get the Watson Developer Cloud Python SDK](#)

Step-by-step instructions:

Let's go through all the steps to build our project!

Create Project

It is assumed that you have an IBM Bluemix account. Sign in to your account and select **Catalog** and search **Natural Language Understanding** and create new instance of service.

View Credentials

You will need to click on the **Service Credentials** and **View Credentials** link to get the details that we need to populate the Jupyter notebook.

Configure Machine Learning Annotator in IBM Knowledge Studio

As defined in the Requirements section, it is assumed that you have a Knowledge Studio account. Sign in to your account and launch Knowledge Studio.

Add Entities Type

The first step is to click **Assets > Entity types**. We defined ten entity types.

IBM Watson Knowledge Studio

Entity Types

Entity Types 12 | Mention Classes | Mention Types

Buttons: Add Entity Type, Upload, Download Types

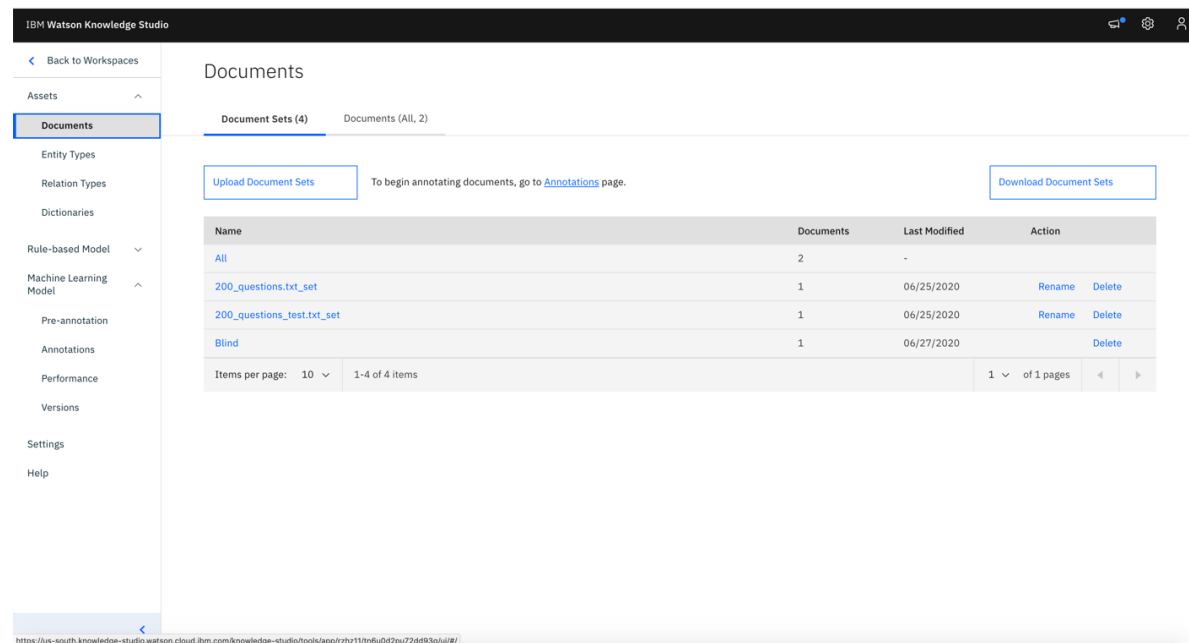
Search: Enter text to filter

<input type="checkbox"/>	Entity Type Name	Roles	Subtypes	Action
<input type="checkbox"/>	language	language		Edit Delete
<input type="checkbox"/>	theorem	theorem		Edit Delete
<input type="checkbox"/>	action	action		Edit Delete
<input type="checkbox"/>	thing	thing		Edit Delete
<input type="checkbox"/>	subject	subject		Edit Delete
<input type="checkbox"/>	profession	profession		Edit Delete
<input type="checkbox"/>	achievement	achievement		Edit Delete
<input type="checkbox"/>	city	city		Edit Delete
<input type="checkbox"/>	place	place		Edit Delete
<input type="checkbox"/>	insect	insect		Edit Delete

Items per page: 10 | 1-10 of 12 items | 1 of 2 pages

Upload Documents for Annotation

The next step is to click **Documents** > **Upload Document Set**. These documents contain questions data.

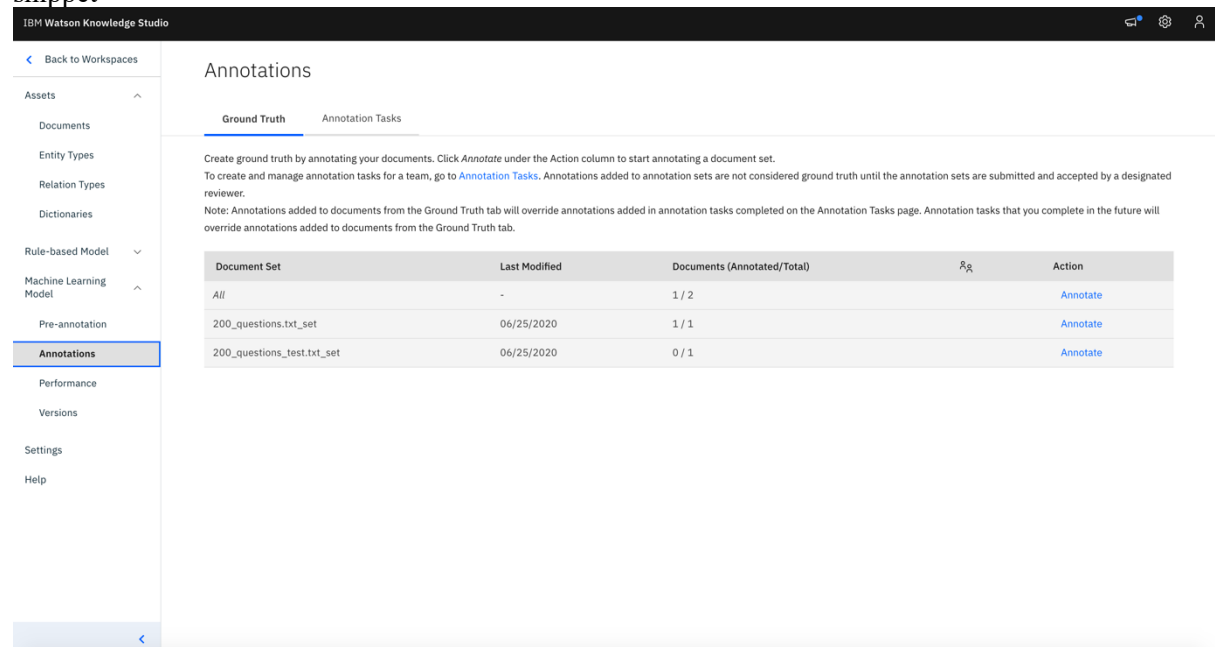


The screenshot shows the IBM Watson Knowledge Studio interface. The left sidebar contains a navigation menu with options: Back to Workspaces, Assets, Documents (selected), Entity Types, Relation Types, Dictionaries, Rule-based Model, Machine Learning Model, Pre-annotation, Annotations, Performance, Versions, Settings, and Help. The main content area is titled 'Documents' and has two tabs: 'Document Sets (4)' and 'Documents (All, 2)'. The 'Document Sets (4)' tab is active, showing a table with columns: Name, Documents, Last Modified, and Action. The table lists four document sets: 'All' (2 documents), '200_questions.txt_set' (1 document, last modified 06/25/2020), '200_questions_test.txt_set' (1 document, last modified 06/25/2020), and 'Blind' (1 document, last modified 06/27/2020). Each row has a 'Rename' and 'Delete' link. Below the table, there is a pagination bar showing 'Items per page: 10' and '1-4 of 4 items'. At the bottom of the page, there is a URL: 'https://us-south.knowledge-studio.watson.cloud.ibm.com/knowledge-studio/tools/app/rzh11/tm6ud2pu72dd93q/ul/#/'.

Name	Documents	Last Modified	Action
All	2	-	
200_questions.txt_set	1	06/25/2020	Rename Delete
200_questions_test.txt_set	1	06/25/2020	Rename Delete
Blind	1	06/27/2020	Delete

Create Annotation Set

The system needs a set of human annotators for identifying all entities for each document containing questions. Under **Machine Learning Model** > **Annotations** you can see two documents in below snippet



The screenshot shows the IBM Watson Knowledge Studio interface. The left sidebar contains a navigation menu with options: Back to Workspaces, Assets, Documents, Entity Types, Relation Types, Dictionaries, Rule-based Model, Machine Learning Model, Pre-annotation, Annotations (selected), Performance, Versions, Settings, and Help. The main content area is titled 'Annotations' and has two tabs: 'Ground Truth' and 'Annotation Tasks'. The 'Ground Truth' tab is active, showing a table with columns: Document Set, Last Modified, Documents (Annotated/Total), $F_{0.5}$, and Action. The table lists three document sets: 'All' (1 / 2), '200_questions.txt_set' (06/25/2020, 1 / 1), and '200_questions_test.txt_set' (06/25/2020, 0 / 1). Each row has an 'Annotate' link. Above the table, there is a text box explaining how to create ground truth by annotating documents and how annotations added to annotation sets are not considered ground truth until accepted by a designated reviewer. Below the table, there is a note stating that annotations added to documents from the Ground Truth tab will override annotations added in annotation tasks completed on the Annotation Tasks page.

Document Set	Last Modified	Documents (Annotated/Total)	$F_{0.5}$	Action
All	-	1 / 2		Annotate
200_questions.txt_set	06/25/2020	1 / 1		Annotate
200_questions_test.txt_set	06/25/2020	0 / 1		Annotate

Click on **Annotate**, we have annotated different entities which found relevant according to the data

The screenshot shows the IBM Watson Knowledge Studio interface. On the left, a sidebar contains navigation links: Assets, Documents, Entity Types, Relation Types, Dictionaries, Rule-based Model, Machine Learning Model, Pre-annotation, Annotations (selected), Performance, Versions, Settings, and Help. The main area displays a document titled '200_questions.txt' with 16 numbered sentences. Each sentence contains entities highlighted in different colors (e.g., 'Alessandro Volta' in pink, 'professor' in green, 'chemistry' in blue). On the right, a legend titled 'Entity' and 'Mention' lists various entity types with corresponding colored squares: achievement, action, city, gender, human, insect, language, place, profession, subject, theorem, and thing. The 'Annotations' tab is active, showing the document's content and the entity annotations.

Create a Machine Learning Annotator Model

We trained and evaluated with separate training and evaluation files.

The screenshot shows the 'Training / Test / Blind Sets' configuration screen in IBM Watson Knowledge Studio. The top bar indicates 'Last trained on: Jun 25, 2020 5:18:30 PM' and 'Last evaluated on: Jun 25, 2020 4:57:47 PM'. Below this, there are three buttons: 'Train', 'Evaluate', and 'Train & Evaluate'. A table lists the document sets and their task status:

Document Set	Task Status
<input type="checkbox"/> All	
<input checked="" type="checkbox"/> 200_questions.txt_set	
<input type="checkbox"/> 200_questions_test.txt_set	

On the right, there are options to 'Create new sets by splitting the selected document sets'. The 'Ratio' section allows entering the percentage of documents to include in each set:

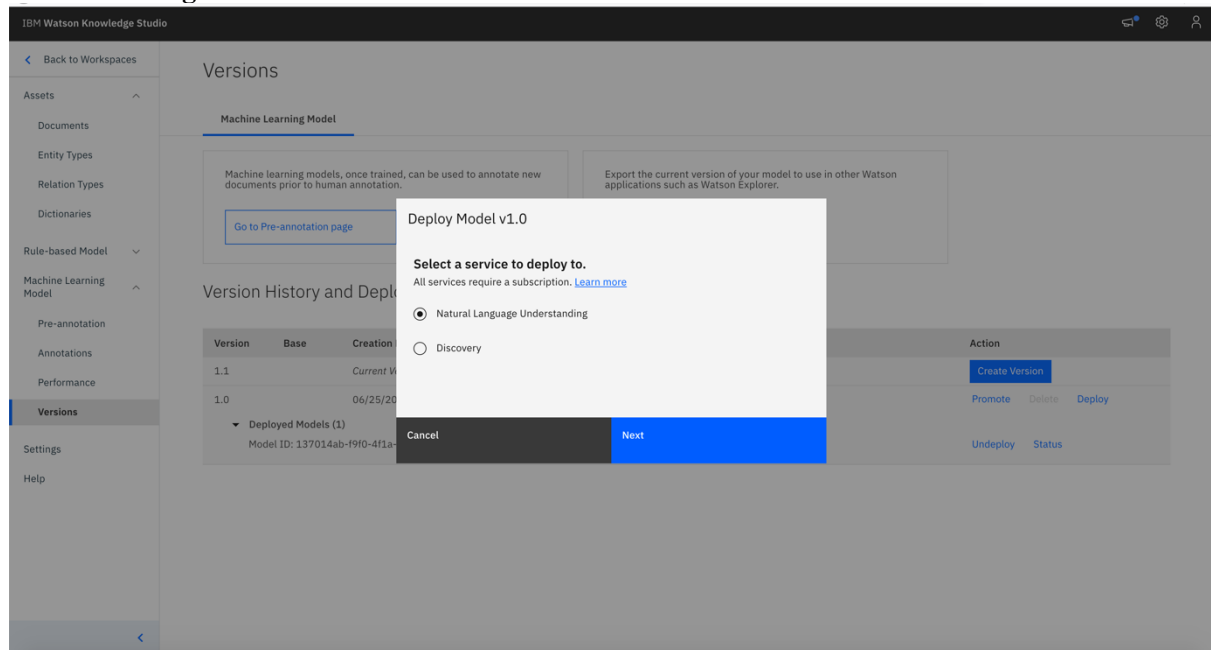
- Training Set (70% Recommended): 100
- Test Set (23% Recommended): 0
- Blind Set (7% Recommended): 0

Below this, there is a section to 'Add documents in the selected sets to:' with a table showing the number of documents for each set:

Set Name	Number of Documents
<input checked="" type="checkbox"/> Training Set	
<input type="checkbox"/> Test Set	
<input type="checkbox"/> Blind Set	1

Deploy the Machine Learning Annotator Model

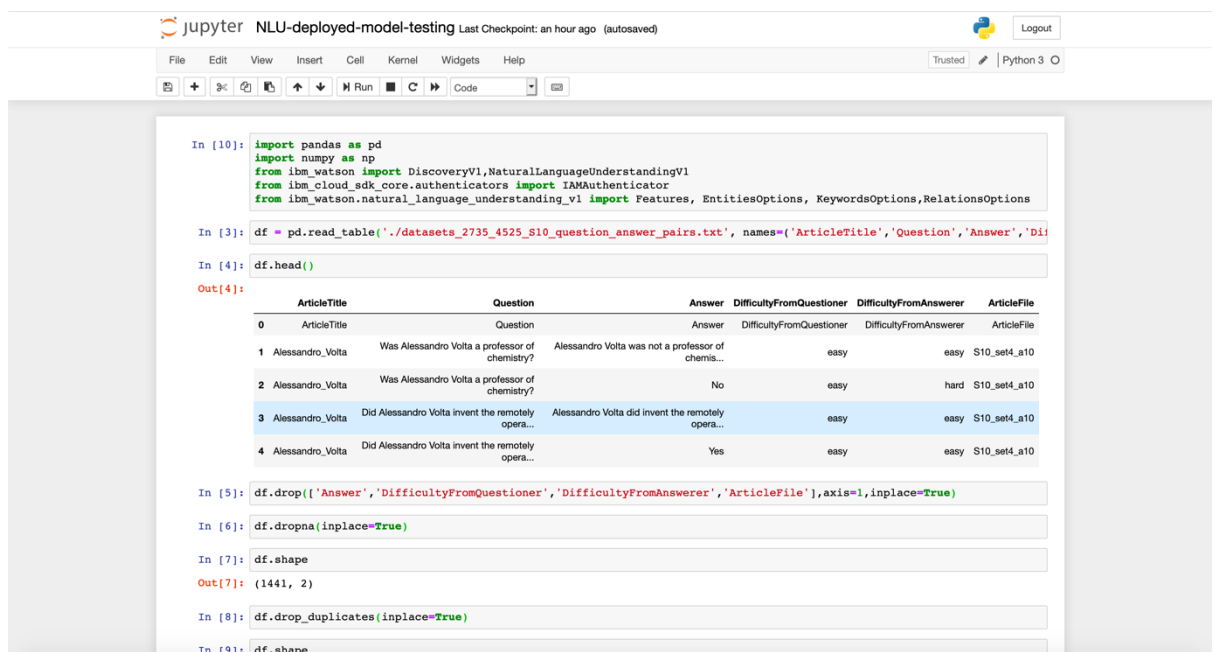
Go to **Machine Learning Model > Versions > Create Version > Deploy > Natural Language Understanding**



Copy model id of deployed model

Create Jupyter notebook Back-End on Python:

1. Load testing data



- Copy API key of NLU service and model id of deployed model. Populate IAMAuthenticator with NLU API key, and model id in model dictionary.

```

In [11]: authenticator = IAMAuthenticator("IPaKtGaZ8aVqTL4Pl1JzoO88dwQXTGs0tZqaSHMbSLI")

In [12]: url = "https://gateway.watsonplatform.net/natural-language-understanding/api"
service = NaturalLanguageUnderstandingV1(version="2018-03-16",authenticator=authenticator)

In [13]: service.set_service_url(url)

In [21]: response = service.analyze(
    text=df['Question'].iloc[1],    features={
        'relations': {
            'model': '137014ab-f9f0-4f1a-8db0-371697970218'
        },
        'entities':{ 'model': '137014ab-f9f0-4f1a-8db0-371697970218'}
    })
    ).get_result()

In [22]: response
Out[22]: {'usage': {'text_units': 1, 'text_characters': 46, 'features': 2},
  'relations': [],
  'language': 'en',
  'entities': [{ 'type': 'human',
    'text': 'Alessandro Volta',
    'disambiguation': { 'subtype': [ 'NONE' ]},
    'count': 1,
    'confidence': 0.999277},
  { 'type': 'profession',
    'text': 'professor',
    'disambiguation': { 'subtype': [ 'NONE' ]},
    'count': 1,
    'confidence': 0.998283},
  { 'type': 'subject',
    'text': 'chemistry',
    'disambiguation': { 'subtype': [ 'NONE' ]},
    'count': 1,
    'confidence': 0.982997}]}

```

In response model will return tagged entities.