



INTERNAL

SAP Data Intelligence hands-on exercises

This document will guide you step-by-step through the process of training and implementing a text analysis and developing a cluster machine learning model using Python and SAP DI Pipelines.

THE BEST RUN



www.sap.com/contactsap

© 2018 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies. See www.sap.com/copyright for additional trademark information and notices.

THE BEST RUN



Table of Contents

DISCLAIMER	4
OBJECTIVE	4
SCENARIO.....	4
ENVIRONMENT ACCESS	5
STEP 1 – USE A JUPYTER NOTEBOOK.....	6
STEP 2 – BUILD MODEL PIPELINES	9
STEP 3 – USE YOUR CLUSTER MODEL	23
APPENDIX 1 – UNDERSTANDING PYTHON CODE TO TRAIN THE CLUSTER MODEL	ERROR!
BOOKMARK NOT DEFINED.	
APPENDIX 2 – UNDERSTANDING PYTHON CODE TO APPLY THE CLUSTER MODEL	ERROR!
BOOKMARK NOT DEFINED.	

DISCLAIMER

The information shared in this document is confidential and proprietary to SAP and may not be disclosed without the permission of SAP. All functionality presented here is subject to change and may be changed by SAP at any time for any reason without notice.

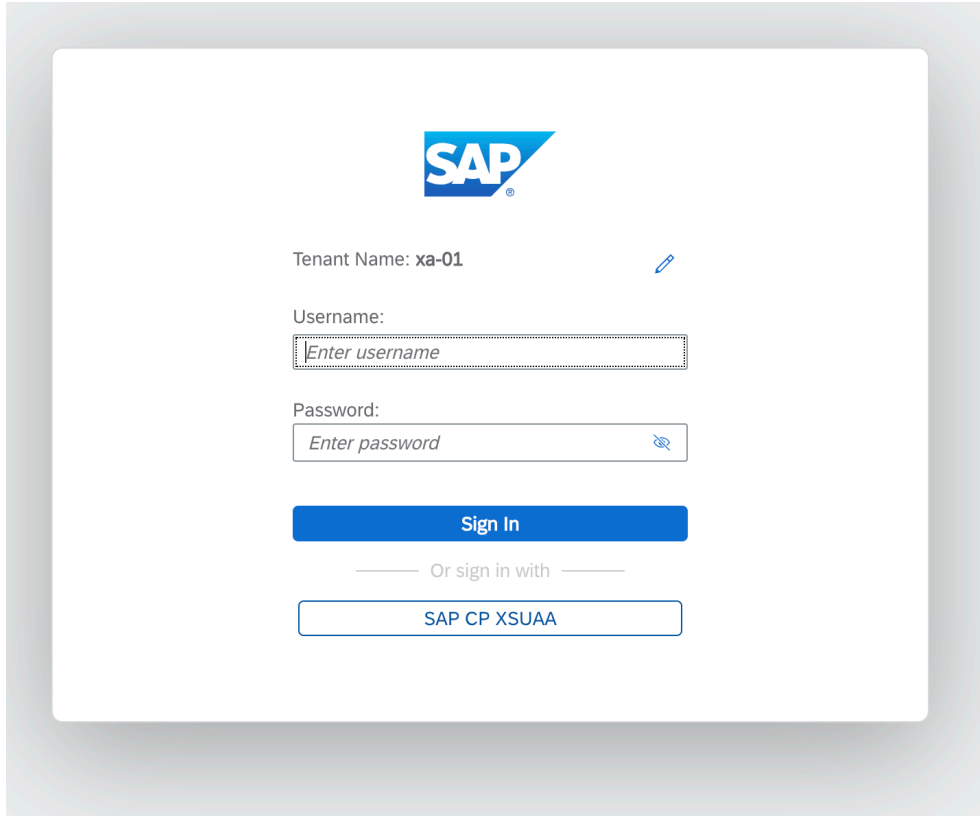
OBJECTIVE

The objective of this exercise is to give you an overview of how you can use the machine learning capabilities in SAP Data Intelligence.

SCENARIO

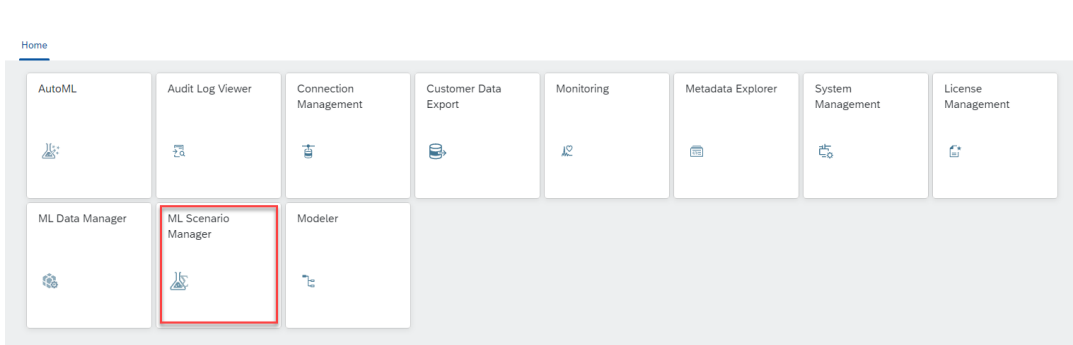
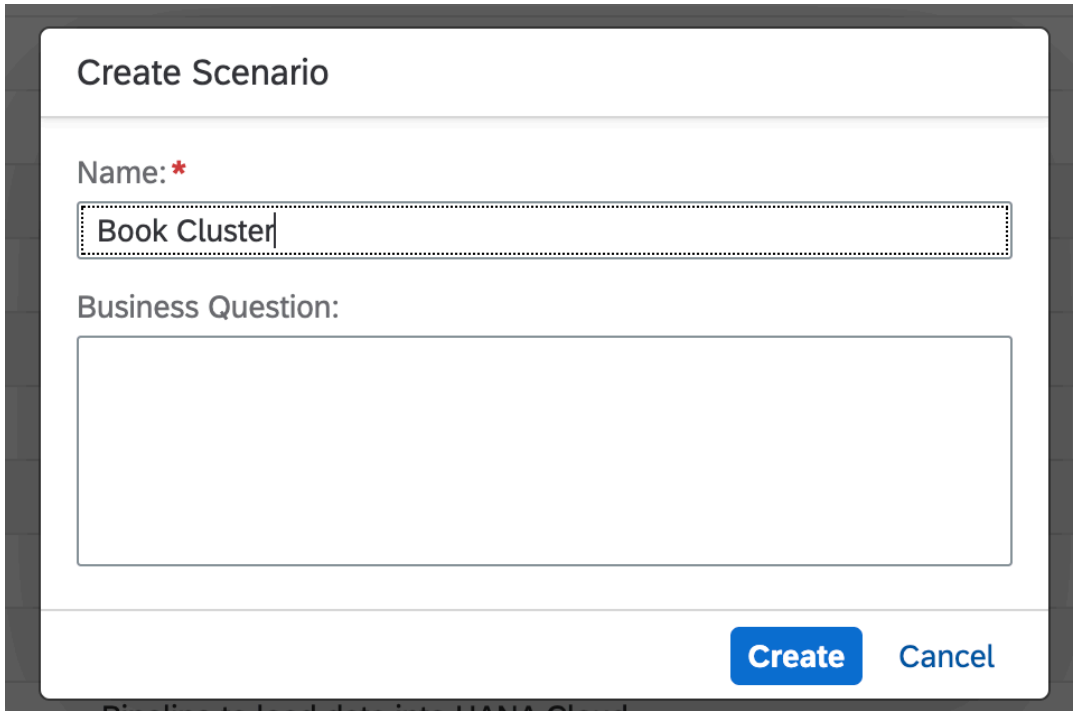
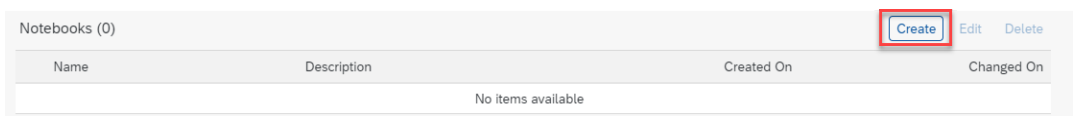
Books are grouped together in a bookshop based on their similarity, so that customers browsing for a book will find lots of similar books on the same shelf. This exercise analyzes the book description data using Python text mining algorithms and then uses this information to assign each book to a cluster. The books within a cluster are as similar as possible (based on the book description), so they are as homogeneous as possible, and there is as wide a difference as possible between clusters, so the different clusters are heterogeneous.

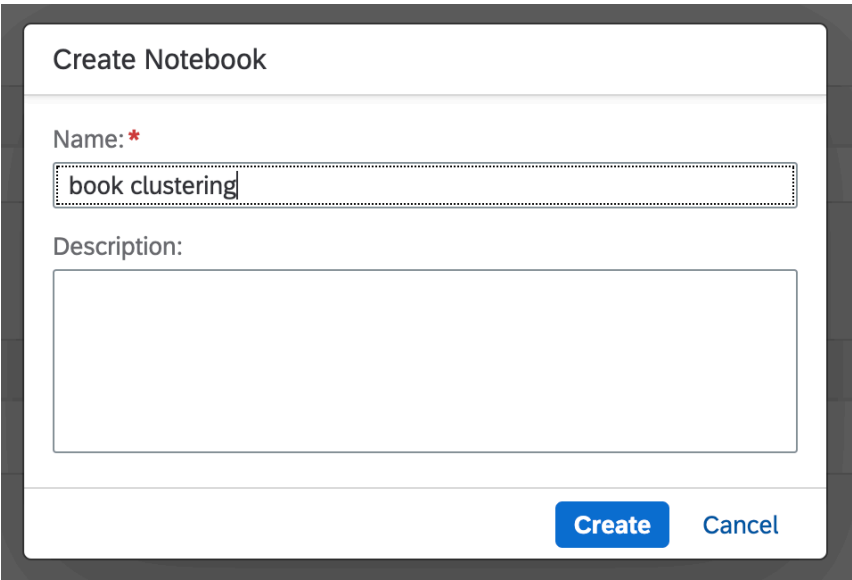
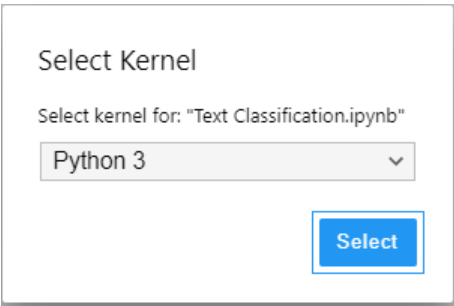
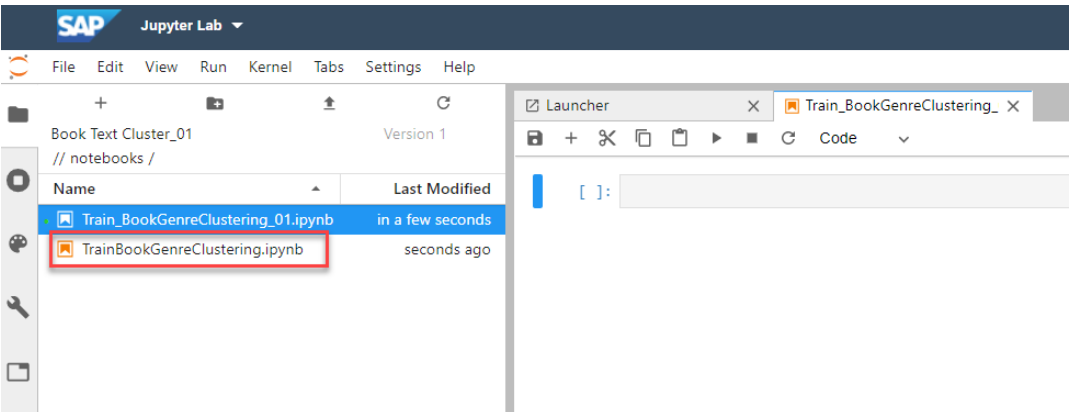
ENVIRONMENT ACCESS

Explanation	Screenshot
<p>In order to open the SAP Data Intelligence application, follow the URL:</p> <p>https://di-eu10.sapexperienceacademy.com/login/?redirectUrl=%2Fapp%2Fdatahub-app-launchpad%2F&tenant=xa-01</p> <p>Login with the username and password provided by your instructor.</p>	

STEP 1 – USE A JUPYTER NOTEBOOK

A Jupyter Notebook environment is used to explore the data, and to run predictive model tests to compare the accuracy of different algorithms and the best settings for the hyper-parameters for the algorithms.

Explanation	Screenshot
Click to open ML Scenario Manager	 The screenshot shows the Databricks home dashboard. It features a grid of tiles for various tools: AutoML, Audit Log Viewer, Connection Management, Customer Data Export, Monitoring, Metadata Explorer, System Management, License Management, ML Data Manager, ML Scenario Manager (highlighted with a red border), and Modeler.
Click the Create button. Create a new scenario. Name the scenario “Book Cluster” You see the empty scenario. First, you will use the Notebooks to explore the data and to script the text analysis and cluster model in Python. Next, pipelines bring the code into production. Executions of these pipelines will create Machine Learning models, which are then deployed as REST-API for inference.	 The screenshot shows the 'Create Scenario' form. It has a 'Name' field with a red asterisk, containing the text 'Book Cluster'. Below it is a 'Business Question' text area. At the bottom right, there are 'Create' and 'Cancel' buttons. The 'Create' button is highlighted with a red border.
In the Notebooks section, click Create to create a new notebook.	 The screenshot shows the 'Notebooks (0)' section. It has a table with columns: Name, Description, Created On, and Changed On. Below the table, it says 'No items available'. At the top right, there are 'Create', 'Edit', and 'Delete' buttons. The 'Create' button is highlighted with a red border.

Explanation	Screenshot
<p>Name the notebook "book clustering"</p>	
<p>Select the Python 3 Kernel option.</p>	
<p>Use the Upload Files function to upload the Python code into the console. Upload file BookGenreClustering.ipynb and text_clustering.png (you can find them in the bookcamp github repository) Double click on the notebook that is uploaded.</p>	 <p>Select file BookGenreClustering.ipynb</p>

Explanation

Double click on the notebook that is uploaded.

Step through the code to check it works correctly. Highlight each cell in sequence and click the arrow button to run the selected cell and advance. Note that line 9 loads the word embedding data and line 19 runs a tsne model. Both of these steps will take a few minutes to complete. Analyze the results of your exploratory data analysis and understand the text analysis and cluster model results.

Screenshot

Book Genre Clustering

In this notebook, we will walk through some **text clustering** with open source Python libraries. Let's think about the following scenario: Mr. Cricket, the owner of the best children bookshop in Walldorf, would like to put some order in his book inventory. He would like to classify the books into different categories based on topic similarity. This would allow him to improve customer experience both in his physical store and his web store, by grouping similar items into homogeneous shelves. Pretty nice idea, but how to do that? 🤖

Mr. Cricket asks for help to his SAP trusted partner. Their consultants, after taking a look at the bookshop book inventory, come up with a plan. Each book in the inventory comes with a very concise description field. They are going to implement some text analysis on this field and group books with a similar content using an unsupervised clustering strategy. Their project will be then based on the following steps:

- 1- Text Preprocessing
- 2- Word Embedding
- 3- Text Clustering

Workflow Diagram:

```

    graph LR
      A["Shadow  
Alghan War, 2001--Juvenile fiction : Dogs--War use : Human-animal relationships--  
Stronger than magic  
Seventeen coffins  
Edinburgh (Scotland)--History--19th century--Juvenile fiction; Historical fiction"] -- "Text Preprocessing (tokenization)" --> B["Tokens  
alghan war 2001 juvenile fiction dogs use human animal relationships  
unicorn juvenile fiction ponies fantasy  
edinburgh scotland history 19th century juvenile fiction historical"]
      B -- "Embedding (word2vec)" --> C["Embedding Table  
t_1 t_2 ... t_100  
0.129 0.009 ... 0.129  
0.590 0.928 ... 0.678  
0.001 0.787 ... 0.556"]
      C -- "Clustering (KMeans)" --> D["Clustering Result  
t-SNE plot showing clusters for 'Stronger than magic' and 'Seventeen coffins'"]
  
```

Let's put this into practice!

First, we will make sure the required libraries are installed. We will use a set of very common python libraries, for dataframe handling and visualization (pandas, numpy, matplotlib, seaborn), regex and nltk for text cleaning and preprocessing, gensim for the word embedding and sklearn for the clustering. hana_ml will be used in this notebook only to access the book inventory data, that are stored into Mr. Cricket HANA Cloud database.

```
[1]: # basic Python
      pip install pandas
      pip install numpy
      pip install matplotlib
      pip install seaborn

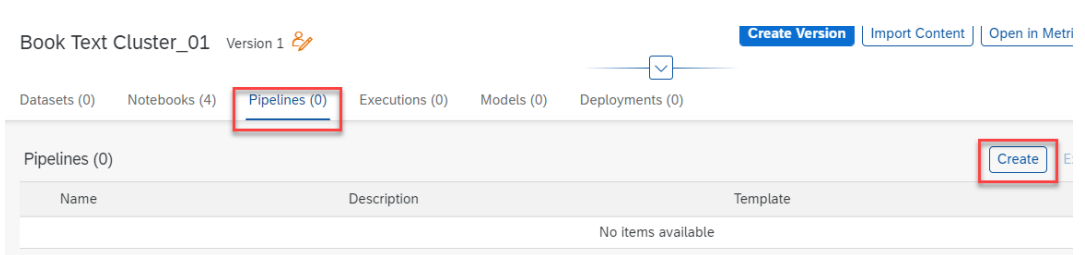
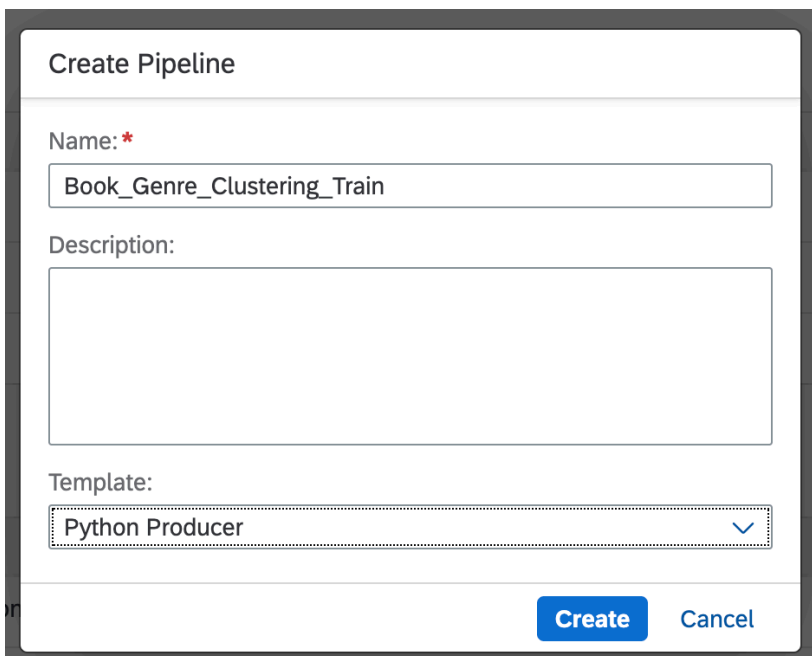
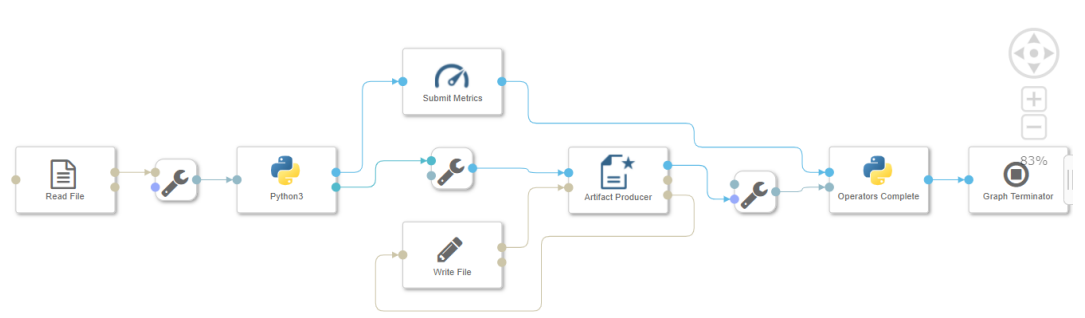
      # text preprocessing
      pip install regex
      pip install nltk
```

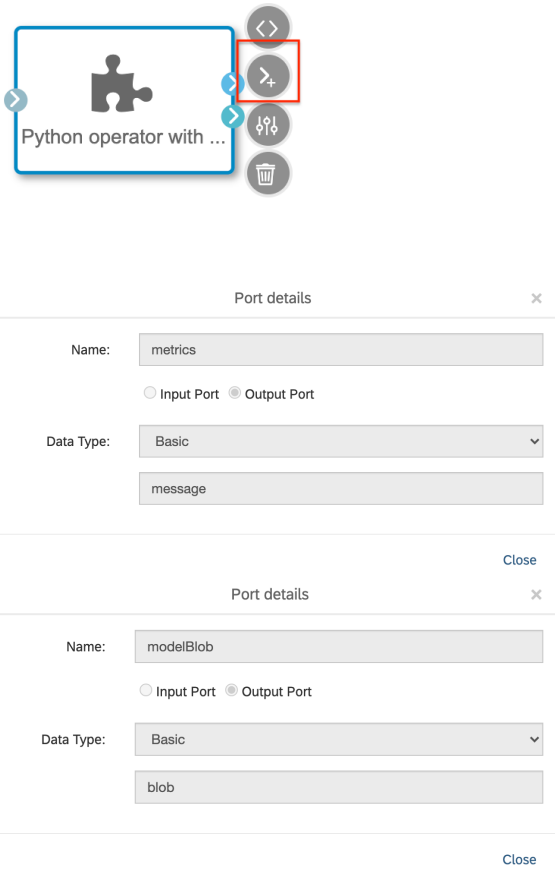
Once you have stepped through to the end of the notebook, and there are no errors, save the notebook.

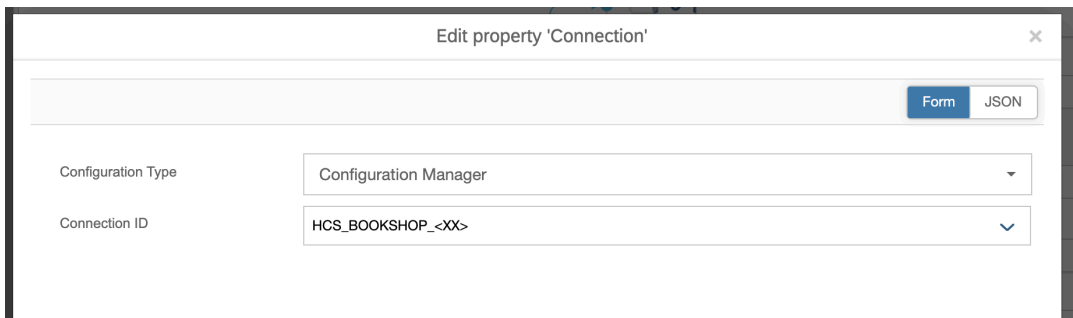
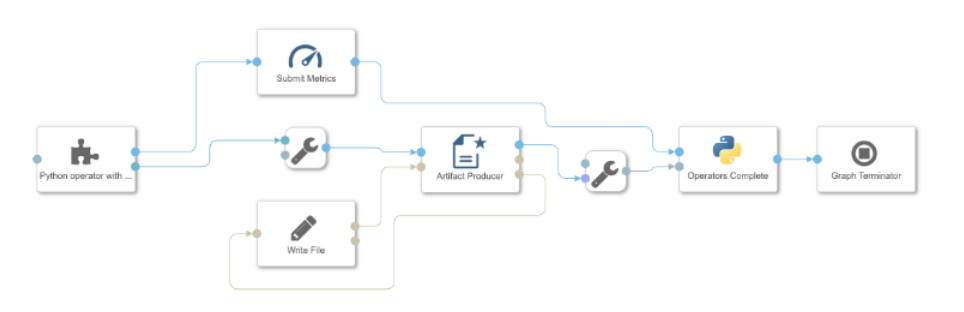
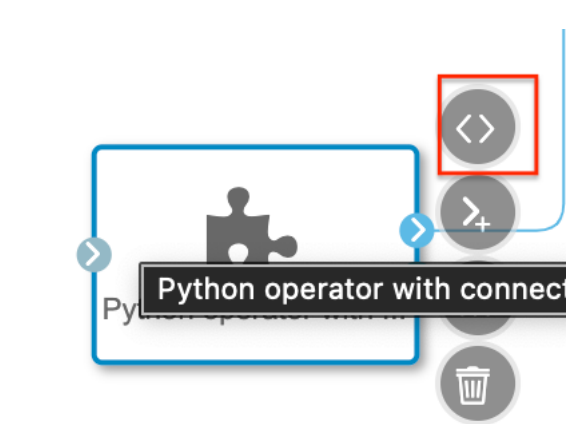
Book Genre Clustering

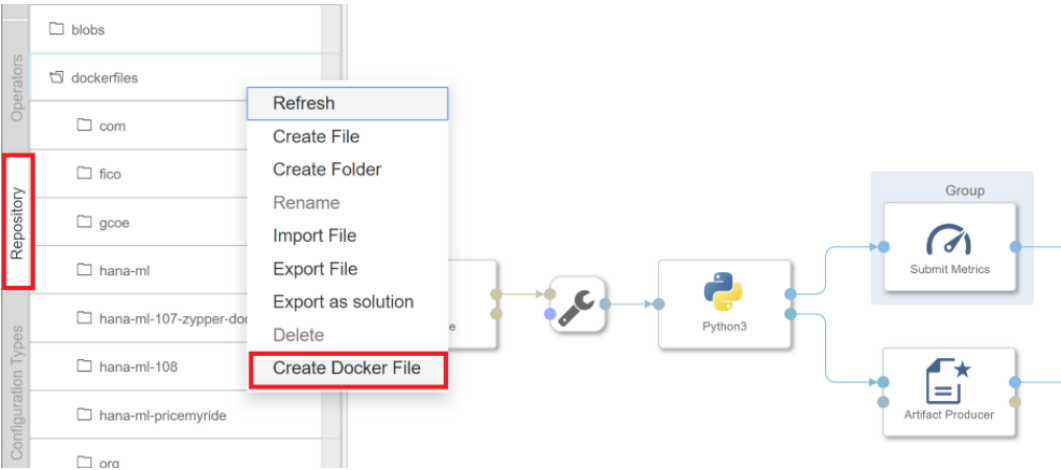
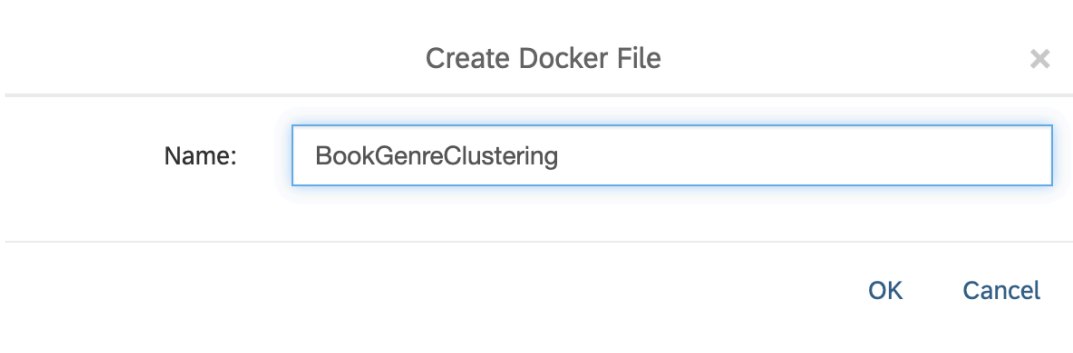
STEP 2 – BUILD MODEL PIPELINES

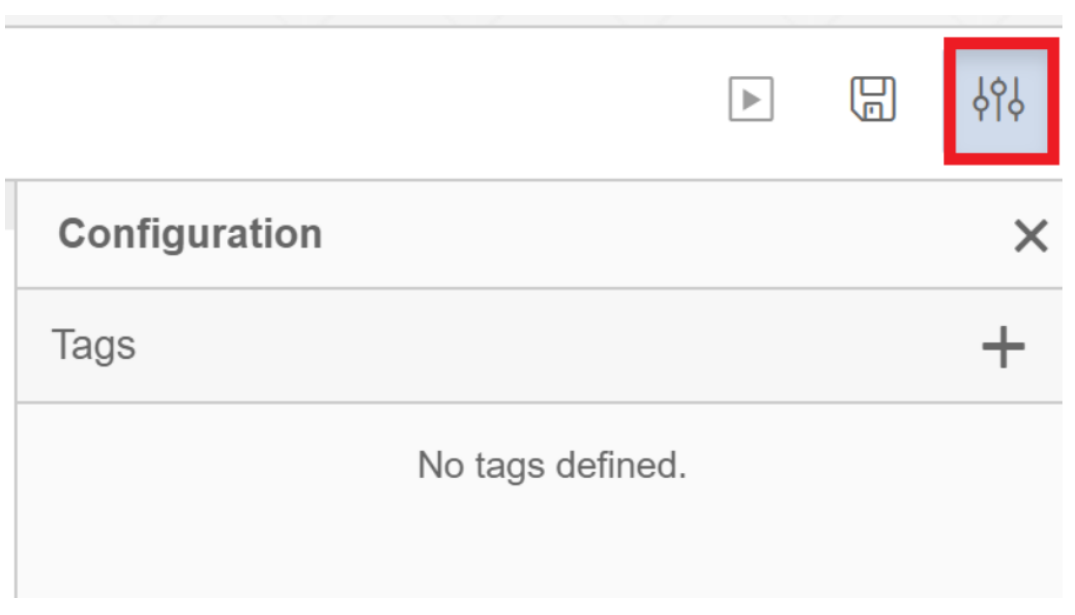

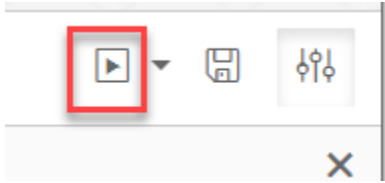
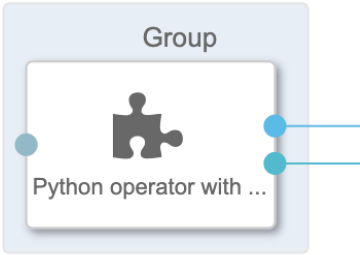
Model pipelines are used to operationalize the cluster model. Now that you have prepared the data, identify the best algorithm to use and which hyper-parameters work best, you want to take the model to production. You will build two pipeline. The first one is used to automate the model training, while the second one is used to inference the model on new data.

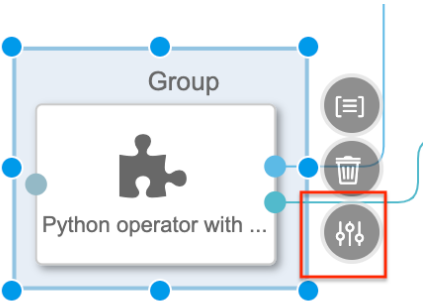
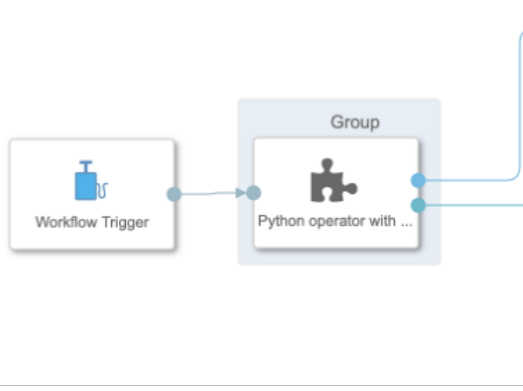
Explanation	Screenshot
To create the graphical pipeline to retrain the model, go to your ML Scenario's main page, select the "Pipelines" tab and click Create.	 The screenshot shows the 'Book Text Cluster_01' page with 'Version 1'. The 'Pipelines (0)' tab is highlighted with a red box. Below it, a table shows no items available. A 'Create' button is highlighted with a red box in the top right corner of the Pipelines section.
Name the pipeline "Text Clustering Train" and select the "Python Producer" template. Click Create.	 The 'Create Pipeline' dialog box is shown. The 'Name' field contains 'Book_Genre_Clustering_Train'. The 'Description' field is empty. The 'Template' dropdown menu is set to 'Python Producer'. At the bottom, there are 'Create' and 'Cancel' buttons.
You need to adjust the pipeline template. The pipeline loads data with the "Read File" operator. The data is passed to a Python operator, where the ML model is trained. The same Python-operator stores the model in the ML	 The screenshot shows a graphical pipeline diagram. It starts with a 'Read File' operator, followed by a 'Python' operator. The 'Python' operator is connected to an 'Artifact Producer' operator. There are also 'Submit Metrics' and 'Write File' operators. The pipeline ends with a 'Graph Terminator' operator showing a progress bar at 83%.

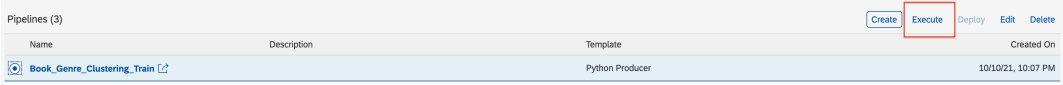
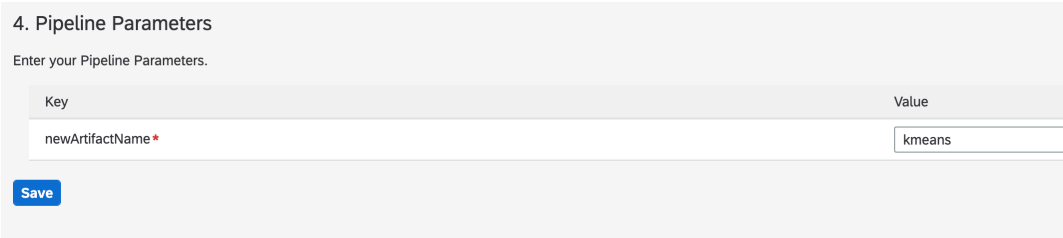
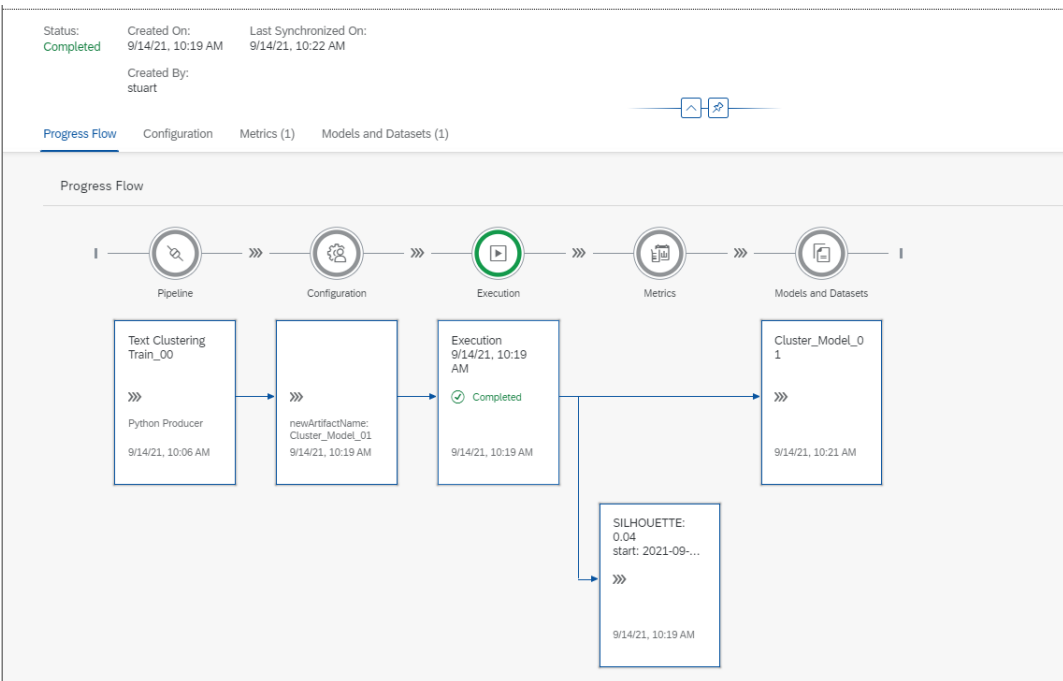
Explanation	Screenshot
Scenario through the “Artifact Producer”. The Python-operator’s second output passes a model quality metric to the ML Scenario. Once both model and metric are saved, the pipeline’s execution is ended with the “Graph Terminator”.	
<p>Since for us the input data are not stored in a file, but in HANA Cloud, we don’t need the Read File operator. We will adapt the custom python operator we built in the previous exercise.</p> <p>Insert a Python with HANA ML connection operator in the canvas. Click on the Add port symbol to add two output ports.</p> <p>Configure the ports with the parameter shown in the picture</p>	

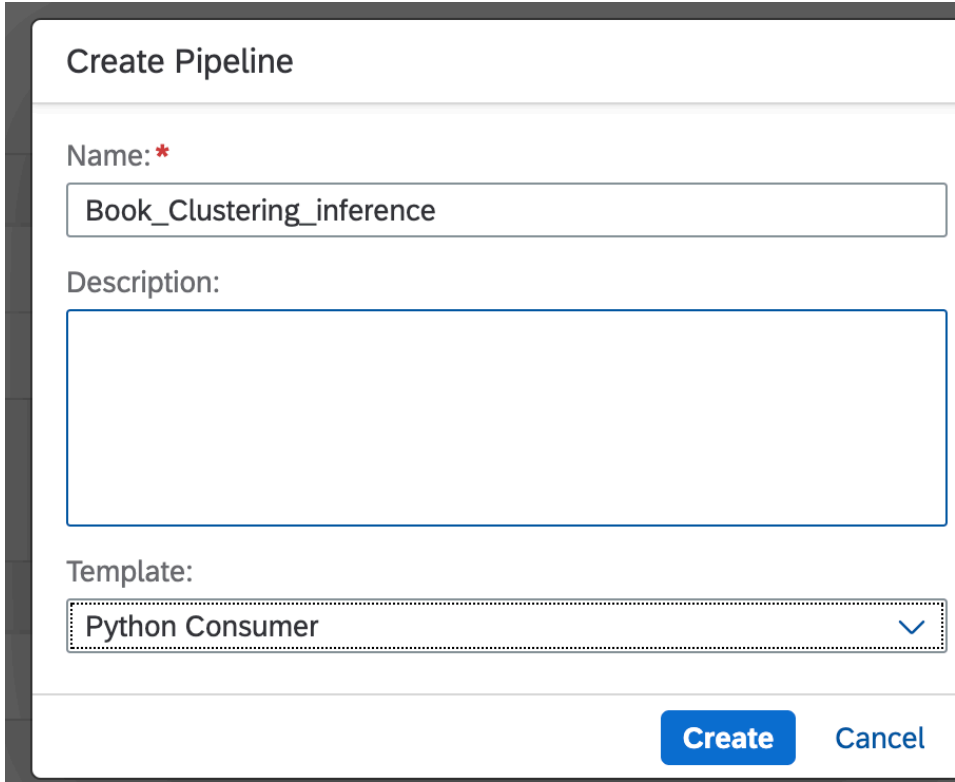
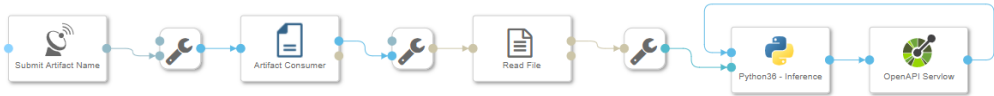
Explanation	Screenshot
<p>In the Configuration panel, open the “Connection” configuration, set “Configuration type” to “Connection Management” and select from the drop-down menu the connection ID related to your Hana Cloud instance</p>	
<p>Delete the Read File operator, and the Conversion operator. Replace the Python Operator with the custom operator that you have just configured your pipeline should look like in the picture.</p>	
<p>Next, adjust the Python code that analyzes the text data and trains the cluster model. Select the custom operator and click the Script option.</p>	
<p>The template code opens up. It shows how to pass the text analysis, model and metrics into the ML Scenario.</p>	<pre> import hana_ml from hana_ml import dataframe import pandas as pd import numpy as np def on_input(data): conn = hana_ml.dataframe.ConnectionContext(api.config.hanaConnection['connectionProperties']['host'], api.config.hanaConnection['connectionProperties']['port'], api.config.hanaConnection['connectionProperties']['user'], api.config.hanaConnection['connectionProperties']['password'], encrypt='true', </pre>

Explanation	Screenshot
<p>You need to create a Docker image for the custom python operator. This gives the flexibility to leverage virtually any Python library. The Docker file installs the necessary libraries. You find the docker files by clicking into the “Repository” tab on the left, then right-click the “dockerfiles” folder and select “Create Docker File”.</p>	
<p>Name the file BookGenreClustering</p>	
<p>Enter this code into the Docker File window. This code leverages a base image that comes with SAP Data Intelligence and installs the necessary libraries on it.</p>	<pre>FROM \$com.sap.sles.base RUN pip3.6 install --user numpy==1.16.4 RUN pip3.6 install --user pandas==0.24.0 RUN pip3.6 install --user sklearn RUN pip3.6 install --user hana_ml RUN pip3.6 install --user regex RUN pip3.6 install --user nltk RUN pip3.6 install --user gensim</pre>

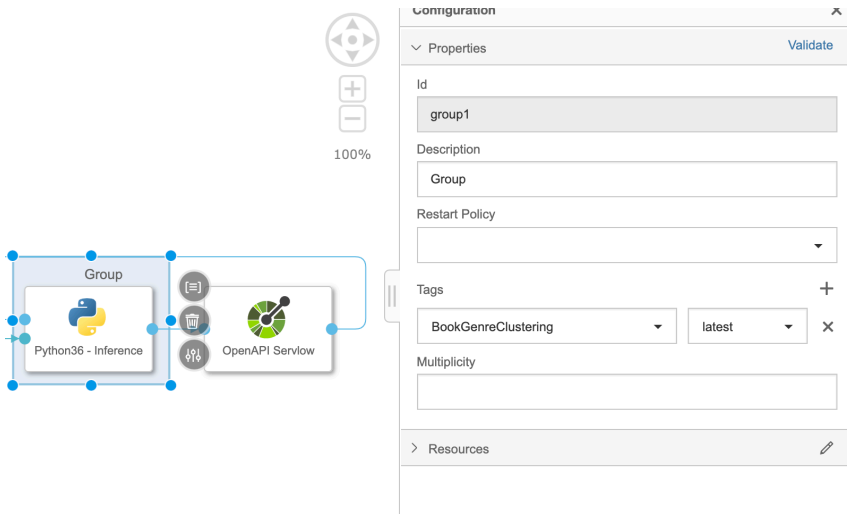
Explanation	Screenshot
<p>Open the Configuration panel for the Docker File with the icon on the top-right hand corner.</p>	
<p>Assign a tag to the docker image</p>	
<p>Now save the Docker file and click the “Build” icon to start building the Docker image. Wait a few minutes and you should receive a confirmation that the build completed successfully.</p>	
<p>Now you need to configure the custom operator, which trains the model, to use this Docker image. Go back to the graphical pipeline and right-click the custom operator and select “Group”.</p>	

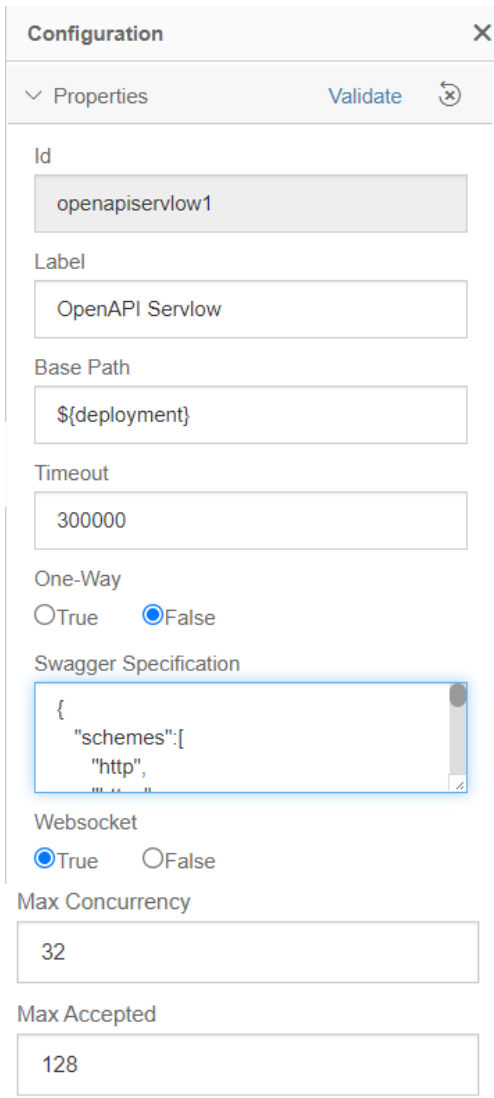
Explanation	Screenshot
<p>You specify which Docker image should be used. Select the group which surrounds the “Python 3” Operator. In the group’s Configuration select the docker image you have built.</p>	 <p>The screenshot shows a configuration window for a 'Group'. Inside the group, there is a 'Python operator with ...'. A red box highlights the configuration options for the operator, including a dropdown menu for 'Name: metricsResponse' and 'Data Type: message'. Below this, the 'Tags' section is highlighted with a red box, showing 'BookGenreClustering' and 'latest' as selected tags. The 'Multiplicity' field is also visible.</p> <p>Configuration ×</p> <p>▼ Properties Validate</p> <p>Id group1</p> <p>Description Group</p> <p>Restart Policy Name: metricsResponse Data Type: message</p> <p>Tags + BookGenreClustering latest ×</p> <p>Multiplicity</p>
	 <p>The screenshot shows a workflow diagram. A 'Workflow Trigger' is connected to a 'Python operator with ...' which is part of a 'Group'. The group is highlighted with a blue border.</p>

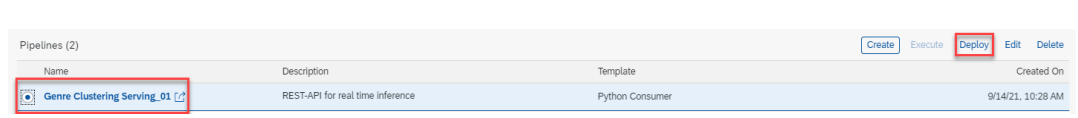
Explanation	Screenshot								
<p>The pipeline is now complete and you can run it.</p> <p>Go back to the ML Scenario. Select the pipeline in the ML Scenario and click the “Execute” button on the right.</p>	 <p>Pipelines (3)</p> <table><thead><tr><th>Name</th><th>Description</th><th>Template</th><th>Created On</th></tr></thead><tbody><tr><td>Book_Genre_Clustering_Train</td><td></td><td>Python Producer</td><td>10/10/21, 10:07 PM</td></tr></tbody></table>	Name	Description	Template	Created On	Book_Genre_Clustering_Train		Python Producer	10/10/21, 10:07 PM
Name	Description	Template	Created On						
Book_Genre_Clustering_Train		Python Producer	10/10/21, 10:07 PM						
<p>Click “Step 3” and then “Step 4” to skip the optional steps until you get to the “Enter your Pipeline Parameters”.</p> <p>Set “newArtifactName” Value to “kmeans”. Click Save.</p> <p>The trained model will be saved under this name.</p>	 <p>4. Pipeline Parameters</p> <p>Enter your Pipeline Parameters.</p> <table><thead><tr><th>Key</th><th>Value</th></tr></thead><tbody><tr><td>newArtifactName*</td><td>kmeans</td></tr></tbody></table> <p>Save</p>	Key	Value	newArtifactName*	kmeans				
Key	Value								
newArtifactName*	kmeans								
<p>Wait a few minutes until the pipeline executes and completes.</p> <p>The metrics section shows the trained model’s silhouette metric.</p> <p>The model itself was saved successfully under the name “kmeans”.</p> <p>The model has a Technical Identifier.</p>	 <p>Status: Completed Created On: 9/14/21, 10:19 AM Last Synchronized On: 9/14/21, 10:22 AM Created By: stuart</p> <p>Progress Flow Configuration Metrics (1) Models and Datasets (1)</p> <p>Progress Flow</p> <pre>graph LR; Pipeline[Pipeline] --> Configuration[Configuration] --> Execution[Execution] --> Metrics[Metrics] --> Models[Models and Datasets]</pre> <p>Execution 9/14/21, 10:19 AM Completed</p> <p>Cluster_Model_01</p> <p>SILHOUETTE: 0.04 start: 2021-09-... 9/14/21, 10:19 AM</p>								

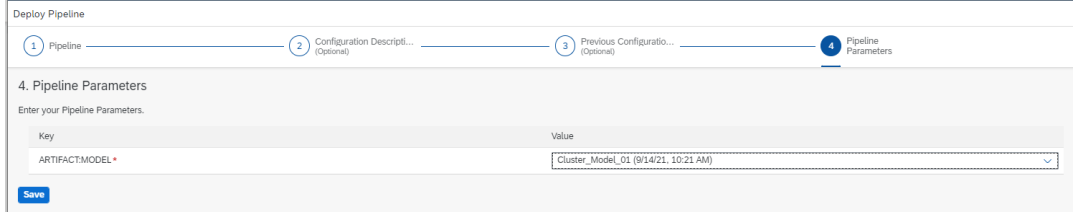
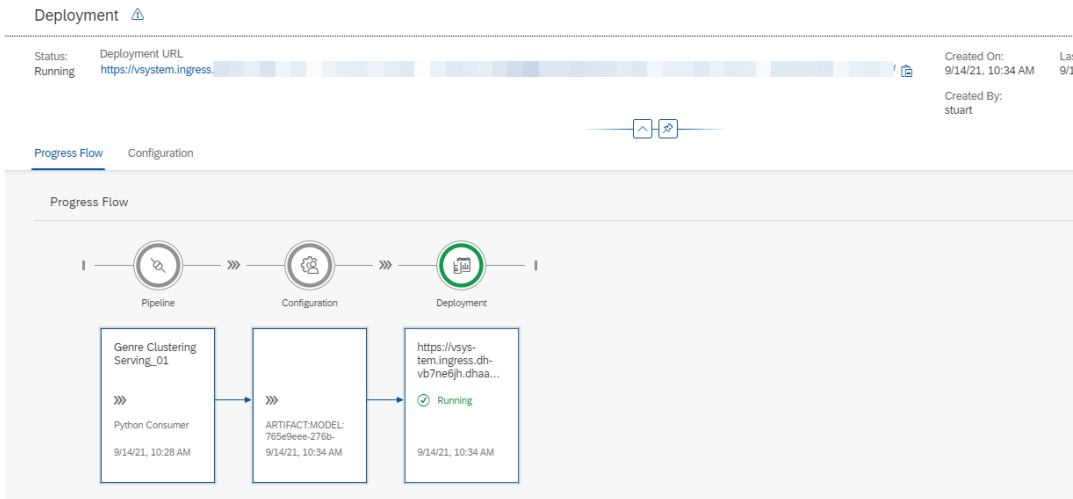
Explanation	Screenshot
<p>You will now use the model for real-time inference with REST-API.</p> <p>Go back to the main page of your ML Scenario and create a second pipeline.</p> <p>This pipeline will provide the REST-API to obtain predictions in real-time. Name the pipeline “Book Clustering Consumer”.</p> <p>Select the template “Python Consumer”.</p> <p>This template contains a pipeline that provides a REST-API.</p>	
<p>The “OpenAPI Servlow” operator provides the REST-API.</p> <p>The “Artifact Consumer” loads the trained model from your ML scenario.</p> <p>The “Python36 – Inference” operator ties the two operators together. It receives the input from the REST-API call (here the user’s text input book description) and uses the loaded model to assign the cluster, which is then returned by the “OpenAPI Servlow” to the client, which</p>	

Explanation	Screenshot
had called the REST-API.	
<p>You only need to change the “Python36 – Inference” operator. Open its “Script” window.</p> <p>Carefully copy and paste to replace the whole code with the code given here.</p> <p>Close the editor window.</p>	<pre> Import json # Global vars to keep track of model status model = None model_ready = False # Validate input data is JSON def is_json(data): try: json_object = json.loads(data) except ValueError as e: return False return True # When Model Blob reaches the input port def on_model(model_blob): global model global model_ready import pickle model = pickle.loads(model_blob) model_ready = True api.logger.info("Model Received & Ready") # Client POST request received def on_input(msg): error_message = "" success = False prediction = None try: api.logger.info("POST request received from Client - checking if model is ready") if model_ready: api.logger.info("Model Ready") api.logger.info("Received data from client - validating json input") user_data = msg.body.decode('utf-8') # Received message from client, verify json data is valid if is_json(user_data): api.logger.info("Received valid json data from client - ready to use") # apply your model # load new data books = json.loads(user_data)["book_description"] # preprocessing import nltk from nltk.corpus import stopwords nltk.download("stopwords") stopwords=set(stopwords.words("english")) import regex as re #Transform to lower case books=[x.lower() for x in books] #Remove punctuation books=[re.sub("[-.!?:\\(\\)]", '', x) for x in books] #Remove stopwords books=[' '.join([t for t in x.split() if not t in stopwords]) for x in books] # Remove short tokens books=[' '.join([t for t in x.split() if len(t) > 1]) for x in books] #Remove extra spaces books=[re.sub(' +', '', x) for x in books] # Remove duplicate tokens books=[' '.join(list(dict.fromkeys(x.split())))) for x in books] # GloVe Vectorization import gensim.downloader as gensim_api word_embedding = gensim_api.load("glove-wiki-gigaword-100") # load pre-trained word-vectors from gensim-data import numpy as np features=[] for book in books: tokens_features=[] for word in book.split(): try: tokens_features.append(word_embedding[word]) except: continue features.append(np.mean(np.array(tokens_features),axis=0)) # deploy cluster model predictions = model.predict(features) success = True else: api.logger.info("Invalid JSON received from client - cannot apply model.") error_message = "Invalid JSON provided in request: " + user_data </pre>

Explanation	Screenshot
	<pre> success = False else: api.logger.info("Model has not yet reached the input port - try again.") error_message = "Model has not yet reached the input port - try again." success = False except Exception as e: api.logger.error(e) error_message = "An error occurred: " + str(e) if success: # apply carried out successfully, send a response to the user msg.body = json.dumps({'Cluster': predictions.tolist()}) else: msg.body = json.dumps({'Error': error_message}) new_attributes = {'message.request.id': msg.attributes['message.request.id']} msg.attributes = new_attributes api.send('output', msg) api.set_port_callback("model", on_model) api.set_port_callback("input", on_input)</pre>
<p>Finally, you just need to assign the Docker image to the “Python36 – Inference” operator. As before, right-click the operator and select “Group”. Add the tag of your docker image. Save the changes.</p>	

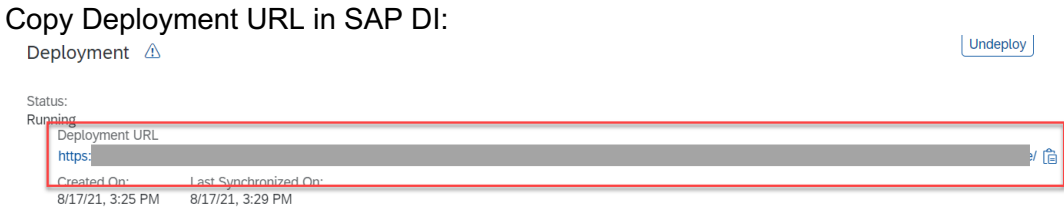
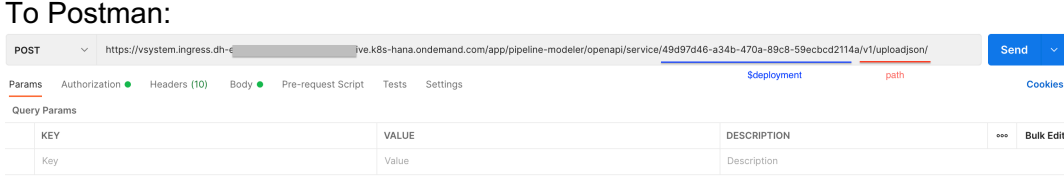
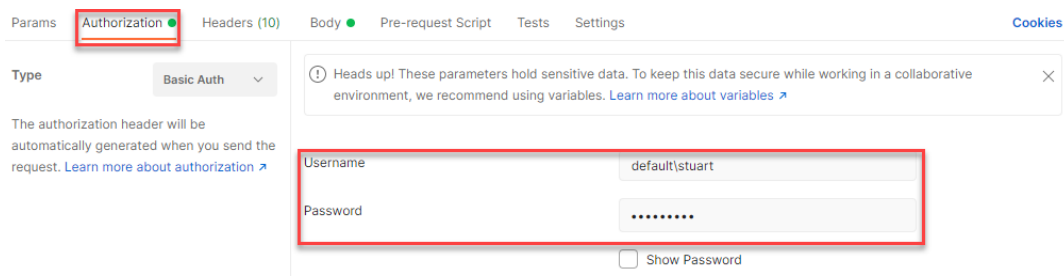
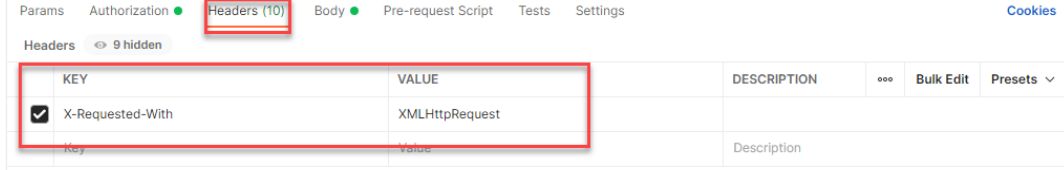
Explanation	Screenshot
<p>Click on OpenAPIServlow and have a look at the configuration</p>	
<p>Notice in particular the content of the Swagger Specification.</p>	<pre data-bbox="423 1413 1479 1902">{ "schemes":["http", "https"], "swagger":"2.0", "info":{ "description":"This is an example of using the OpenAPI Servlow to carry out inference with an existing model.", "title":"OpenAPI demo", "termsOfService":"http://www.sap.com/vora/terms/", "contact":{ }, "license":{ "name":"Apache 2.0", </pre>

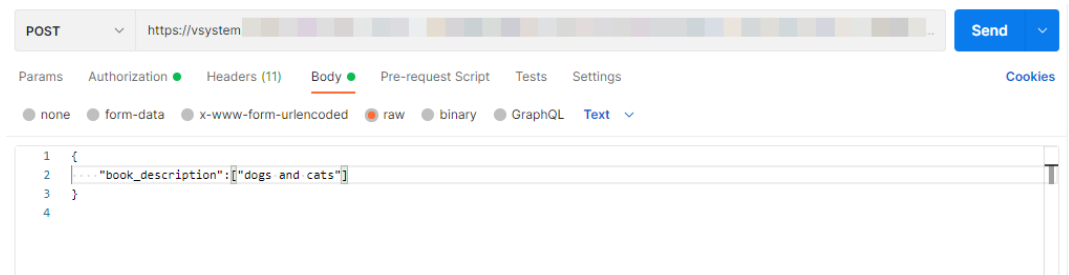

Explanation	Screenshot								
	<pre> "url":"http://www.apache.org/licenses/LICENSE-2.0.html" }, "version":"1.0.0" }, "basePath":"/\$deployment", "paths":{ "/v1/uploadjson":{ "post":{ "description":"Upload data in json format", "consumes":["application/json"], "produces":["application/json"], "summary":"Upload JSON data to be used in the Python operator's script", "operationId":"upload", "parameters":[{ "type":"object", "description":"json data", "name":"body", "in":"body", "required":true }], "responses":{ "200":{ "description":"Data uploaded" }, "500":{ "description":"Error during upload of json" } } } } }, "definitions":{ }, "securityDefinitions":{ "UserSecurity":{ "type":"basic" } } }</pre>								
Go back to the ML Scenario. Now deploy the new pipeline. Select the pipeline and click “Deploy”.	 <p>The screenshot shows a table titled "Pipelines (2)" with columns: Name, Description, Template, and Created On. There is one row with the pipeline "Genre Clustering Serving_01". Above the table, there are buttons: "Create", "Execute", "Deploy" (highlighted with a red box), "Edit", and "Delete". The "Name" column has a link icon next to the pipeline name.</p> <table><tr><th>Name</th><th>Description</th><th>Template</th><th>Created On</th></tr><tr><td>Genre Clustering Serving_01</td><td>REST-API for real time inference</td><td>Python Consumer</td><td>9/14/21, 10:28 AM</td></tr></table>	Name	Description	Template	Created On	Genre Clustering Serving_01	REST-API for real time inference	Python Consumer	9/14/21, 10:28 AM
Name	Description	Template	Created On						
Genre Clustering Serving_01	REST-API for real time inference	Python Consumer	9/14/21, 10:28 AM						

Explanation	Screenshot
Click through the screens until you can select the trained model from the drop-down. Click “Save”.	
After a few minutes the pipeline will start running.	

STEP 3 – USE YOUR CLUSTER MODEL

Now that you have deployed your model, you can use it for real-time cluster assignment. For this, you are going to use the Postman application.

Explanation	Screenshot
<p>Open Postman. Copy the deployment URL from SAP DI. Enter the Deployment URL as request URL. Extend the URL with v1/uploadjson/, the path specified in the OpenAPI servlow operator. Change the request type from “GET” to “POST”.</p>	<p>Copy Deployment URL in SAP DI:</p>  <p>To Postman:</p> 
<p>Go to the “Authorization” tab. Select “Basic Auth” and enter your username and password for SAP Data Intelligence. The username starts with your tenant’s name, followed by a backslash and your actual username.</p>	
<p>Go to the “Headers” tab and enter the key “X-Requested-With” with value “XMLHttpRequest”.</p>	

Explanation	Screenshot
Finally, pass the input data to the REST-API. Select the “Body” tab, choose “raw” and enter the syntax given here.	 <pre>{ "book_description":["dogs and cats"] }</pre>
Press “Send” and after a few minutes you will see the genre prediction that comes from SAP Data Intelligence. Try the REST-API with different text to see how the cluster allocations change.	 <pre>{ "Cluster": [9] }</pre>
You have now completed the exercise.	

