Machine Learning Pipelines Assignment
Instructions

Overview

In this assignment, you will be building a machine learning pipeline using a Jupyter Notebook. Your work on the assignment will include the following:

- Downloading data files from a publicly available data source.
- Downloading starter files from our Weekly Schedule.
- Creating a SageMaker notebook instance using the AWS Academy Learning Lab.
- Uploading starter files to the notebook instance.
- Uploading data files to the notebook instance.
- Creating code in the Jupyter Notebook provided and testing it to create a working pipeline.
- Taking a screenshot image to document having completed your work on AWS.
- Downloading files from your SageMaker notebook instance to your computer.
- Assembling a submissions directory on your computer to hold the files that you will be submitting for the assignment.
- Zipping up your submissions directory and submitting the .ZIP file to the Canvas assignment activity.

Tools

Tools used to complete this assignment should include:

- The data files downloaded from a publicly available data source.
- The starter files.
- AWS Learner Lab environment.
- Jupyter Notebook.
- A utility program that will take and save a screen shot on your computer.

Starter Files

The following starter file is available for download in the Weekly Schedule:

- surname_givenname_machine_learning_pipelines_assignment.zip

Assignment Details

Follow the process below to complete your assignment:

1. Download the data file.
   This assignment will use data from NSF research awards abstracts between 1990 and 2003. The complete dataset contains (a) 129,000 abstracts describing NSF awards for basic research, (b) bag-of-word data files extracted from the abstracts, and (c) a list of words used for indexing the bag-of-word data.

   For this assignment, we will use only the first zipped file (*Abstracts_Part1.zip*), which consists of 51,979 abstracts. You can find the data set and information about the data at http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html .

   For more information regarding these data, please consult the description of the data link provided on the page from which you downloaded the data.

2. Download the starter file.
   Download the starter file and unzip it into a directory that normally holds your assignment submissions. Because the files in the starter zip are named and organized in a manner consistent with the zip file that you will be submitting for this assignment, you should consider the directory that you create to be both the input to your assignment and eventually the output from your assignment.

3. Create a SageMaker notebook instance.
   Using your AWS Academy Learner Lab access, create a SageMaker notebook instance with the following characteristics:

   - Name: machine-learning-pipelines-assignment
   - Notebook Instance Type: ml.t3.xlarge

4. Start your SageMaker notebook instance.

5. Upload files to your SageMaker notebook instance.

   - Upload the *Abstracts_Part1.zip* file to the main directory of the notebook image.
   - Upload the *surname_givenname_machine_learning_pipelines_assignment.ipynb* file to the main directory of the notebook image.

6. Code and test the Jupyter Notebook.
   Follow the instructions contained in the Jupyter Notebook regarding coding and testing in the notebook. As a reference, you may use the notebook that was used for the tutorial video for this assignment. Remember to rename the notebook file to include your name.

7. Download the revised Jupyter Notebook file.
   Before downloading, shut down the kernel for the notebook file. Then, download the file to the directory from which it came on your own computer.

8. Follow the instructions below regarding preparing your zip file and submitting it to Canvas.

Deliverables

You are expected to upload a single .ZIP file to the appropriate submission activity on the Canvas site for our course. This is on the Illinois Canvas system – Not the Canvas system used by AWS Academy. See below for details regarding .ZIP file contents and naming.

Creating and Submitting the Submission File

1. Use the directory that was created when you downloaded the starter files. Rename that directory to include your surname and given name. The resulting directory name should follow the scheme below:

   surname_givenname_machine_learning_pipelines_assignment

   If this were my own submission, I would name my directory as follows:

   trainor_kevin_machine_learning_pipelines_assignment

2. Make sure that you downloaded the Jupyter Notebook file that you populated and renamed while using your SageMaker notebook instance into the directory.

3. Make sure that you include the screenshot file that you created during your notebook session that documents how you created your work using the intended AWS resources.

4. Make sure to include the *Abstracts_Part1.zip* file that contains the data. This will make it easier for us to test and grade your work.

5. Please delete the original version of the Jupyter Notebook file. It was named *surname_givenname_machine_learning_pipelines_assignment.ipynb* .

6. Use a zip utility to create one zip file that contains the directory. The zip file should be named according to the following scheme:

   surname_givenname_machine_learning_pipelines_assignment.zip

   If this were my own submission, I would name the zip file as follows:

   trainor_kevin_machine_learning_pipelines_assignment.zip

7. Submit the .ZIP file to the appropriate submission activity on the Illinois Canvas site for our course.

Due By

Please submit this assignment by the date and time shown in the Weekly Schedule.