Data 603 Homework 1

Using HDFS

#### **Homework: Using HDFS**

At this point you should be able to login to the UMBC Big Data Cluster and launch Jupyter Lab. The goal of this homework is to learn to manipulate files in HDFS and become familiar with the Google Open Image Dataset metadata file formats.

**Step 1: Download the CSV Files**

Using apache arrow (pyarrow) download/get the following files into your Jupyter environment.

* hdfs://data/google\_open\_image/ids/test-images-with-rotation.csv
* hdfs://data/google\_open\_image/labels/test-annotations-human-imagelabels-boxable.csv
* hdfs://data/google\_open\_image/metadata/class-descriptions-boxable.csv

**Step 2: Create a Dataframe**

Using Pandas, read the files into different dataframes. Be sure the dataframes have column names.

Use the join command to join the three dataframes into a single dataframe. Ensure you don't have repeated lines in the resulting dataframe. File 3 can be joined with file 2 based on the Label field. (File 2 calls this LabelName, File 3 calls it Label) That dataframe can be joined with File 1 based on the Image ID field.

**Step 3: Download Images**

The resulting dataframe should contain human-readable label and URLs. Use the requests package to download 3 random images and display them in your notebook.

**Step 4: Upload Images to HDFS**

Using pyarrow, create a folder called homework1\_images in your HDFS user directory. (mkdir /user/<username>/homework1\_images)

Upload/put the downloaded images into the directory you just created.

**Step 5: Email me .ipynb file**

Download a copy of your completed python notebook and email it to me. Please name it in the following format:

**<userid>\_lastname\_homework\_1.ipynb**

#### **Resources:**

[Google Open Image Dataset File Descriptions.](https://docs.google.com/document/d/1iVBCLtaC0Wd2YXXKigrF76A6ufDMQprk-2y1NUnru2s/edit?usp=sharing)

Apache Arrow Python Docs:<https://arrow.apache.org/docs/python/index.html>

Use this to connect to HDFS, download/upload the files from/to HDFS.

Pandas API:<https://pandas.pydata.org/docs/reference/index.html#api>

Read the files into Pandas DataFrames, Use Pandas to Join Dataframes.

Requests:<https://requests.readthedocs.io/en/master/>

Library for doing requests over the internet.

Numpy:<https://numpy.org/>

Pandas is backed by NumPy, you may need it for some things.

Core Python Library: io  
 Bytes IO: <https://docs.python.org/3/library/io.html#binary-i-o>