Data 603: Homework 3: Chicken Hawks

Due 3/29 by 8 PM

The goal of this homework is to find pictures of my favorite bird, the Chicken. So consider yourselves Chicken Hawks and the cluster is your farm.



### Part 1: Find Images of Chickens!

Using the Big Data Cluster, create a dataframe of images of chickens! You will need to use the following files in HDFS:

/data/google\_open\_image/metadata/class-descriptions-boxable.csv

/data/google\_open\_image/labels/\*.csv

/etl/google\_open\_image/images/**test**.avro

Read them into 3 dataframes.

Find ImageIDs containing the labels: ‘Chicken’

Find the number of images that contain a chicken. Remember that an image can have more than 1 chicken in it!

Join the data such that you have a dataframe containing image labels, images, and

Save the dataframe in **AVRO** format to your **HDFS** user directory: /user/<userid>/homework3/homework3\_part1.avro

### Part 2: Finding Friendly Chickens

In this part, you will use the Avro file generated in Part 1 to extract images of Chickens.

Friendly chickens are chickens that hang out with other chickens. Your goal is to extract profile photos of friendly chickens. To do that you’ll need to read in the Avro file you created in part 1 and the bounding box data for each image that contains more than 1 chicken.

General guidelines:

Read the bounding box data from hdfs into a dataframe:

/data/google\_open\_image/bboxes/\*.csv

Filter it for Image IDs present in your Avro file that contain more than 1 chicken.

For each Image, extract a sub-image using the bounding box information. To do this you will need to write a User-Defined Function (UDF) that takes as input, image data and 4 corner points.

Extract the jpg files into HDFS user directory of 5 RANDOM chickens Name the files <image\_id>\_[1-5].jpg

Write an AVRO file to your HDFS user directory containing the following information

ImageID, UserID, Label, XMin, XMax, YMin, YMax, Sub-Image Data

### Final Products:

At the end you should be able to turn in: Python Notebook File: **<lastname>\_<userid>\_hw3.ipynb**

5 image .jpg files with above naming convention

An AVRO file (in HDFS)

Give me access to your /user/<userid>/homework3 HDFS directory containing your data.

# Some Help!

Use PIL Image object and BytesIO in your UDF to open the image!

img = Image.open(io.BytesIO(data))

Image Coordinates in the data are floats, but PIL image coordinates need to be Integers

Use Image.crop to extract the data of interest

Xmin = Left

Ymin = Upper

Xmax = Right

Ymax = Lower

Use Image.save to a BytesIO and return the .getvalue() of the BytesIO variable.