Week 7: Unstructured data; exploration

Objective: In this lab session, you will learn basic data exploration techniques for image and text data. We will use Google Colab - Python IDE for most of our tasks.

Go to <https://colab.research.google.com/> to start a new Python notebook. You will need to log in to use the Google Colab. You can use your university email or your personal.

**Image Data Exploration**

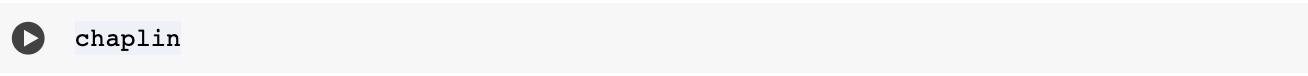
This section investigates the exploration of image data.

* We will explore basic info about image data, including format, size, mode, EXIF and histogram of an image.
* To start, please download the Image folder, unzip the folder and load images from the folder.

**Task1a: Ingest Image data and check the basic properties of the image**

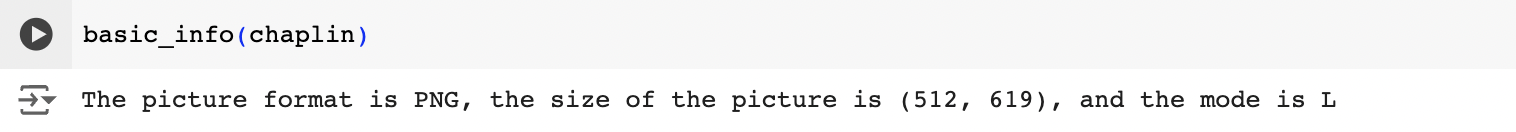
A screen shot of a computer

Description automatically generated



A computer screen shot of text

Description automatically generated



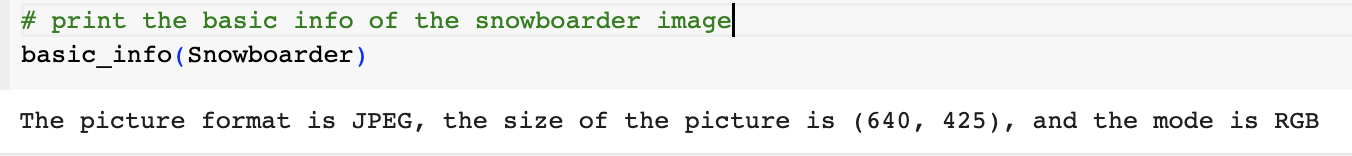
Note:

* The format attribute identifies the source of an image. If the image was not read from a file, it is set to None.
* The size attribute is a 2-tuple containing width and height (in pixels).
* The mode attribute defines the number and names of the bands in the image and the pixel type and depth. Typical modes are “L” (luminance) for grayscale images, “RGB” for actual colour images, and “CMYK” for pre-press images.

**Task1b: Ingest another Image data and check the basic properties of the image**

A screenshot of a website

Description automatically generated

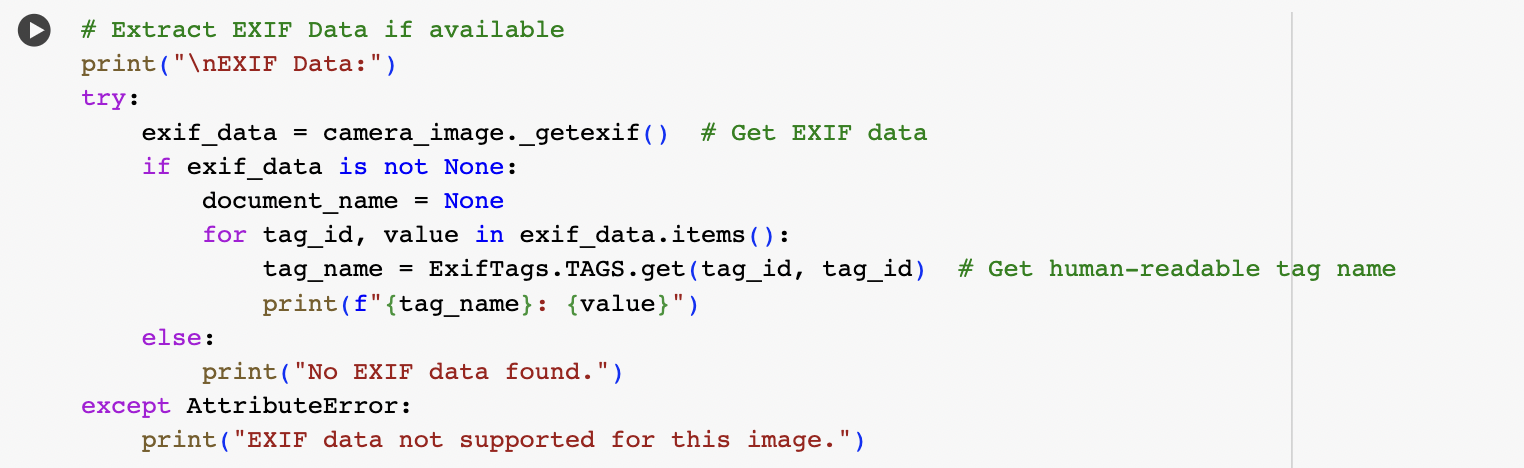


**Task2: EXIF Data for Image data**

As part of image exploration.

* We can also get the EXIF of an image, particularly images from digital devices.
* EXIF (Exchangeable Image File Format) files store essential data about photographs. Almost all digital cameras create these data files when you snap a new picture. An EXIF file holds all the information about the image itself — such as the exposure level, where you took the photo, and any settings you used





**Extra Exercise on Exif**

Try to snap a picture on your phone, save it and load it here. Try getting the Exif of the image you have just snapped and loaded here.

**Task3: Histogram of images**

The histogram plots the number of pixels in the image (vertical axis) with a particular brightness or tonal value (horizontal axis).

* Let’s try to plot the histogram for the snowboarder image.

A screenshot of a computer program

Description automatically generated

**Task4: Calculate the statistics: Mean, median and Mode of the channels.**

A screenshot of a computer program

Description automatically generated

**Extra Exercise on Image Loading & Exploration.**

* Try to load 5 images from the Folder.
* Check the image format.
* Check the image size.
* Check the Image mode.
* Check the Exif data to see if it has
* Compute and print the histogram

Ps. You can do that using our basic\_info function. To invoke the function, for example, basic\_info(image\_variable\_name)

**Text Data**

This is the second part of today's seminar, covering the exploration of textual data.

To start, please download the text data from BB and unzip the folder. There are 20 subfolders. Out of the 20, we will only work on the autos folder.

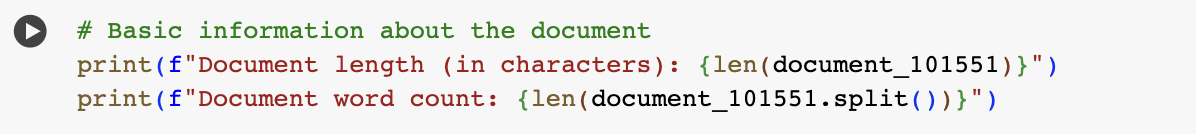
The text folder contains documents from the well-known text dataset [downloadable from here](https://www.google.com/url?q=https%3A%2F%2Fwww.cs.cmu.edu%2Fafs%2Fcs%2Fproject%2Ftheo-20%2Fwww%2Fdata%2Fnews20.html), consisting of 20,000 messages collected from 20 internet newsgroups.

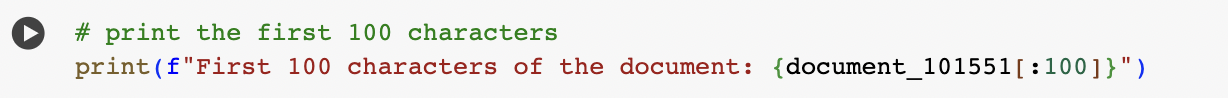
**Task5: Ingest the document file and print the file**

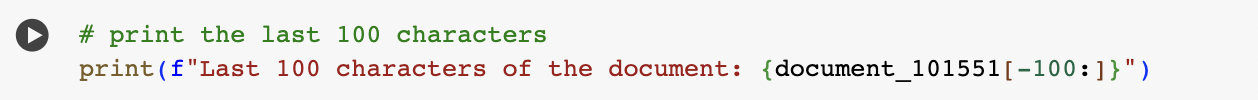
A computer code with colorful text

Description automatically generated

**Task5b: Investigate basic properties like document length, word count, first 100 and last 100 characters**







**Task6: Plot a word count for the document file**

A computer code with text

Description automatically generated with medium confidence

**Extra Exercise on the Text Data Exploration**

Ingest 2 more text documents from any folder of your choice.

* Check and print the length of the documents
* Check and print the word count of the documents
* Create a word cloud for each of the documents