Home assignment – Erez Wasserman

(Note: Run in python3.6)

Your task is to write a python script that receives the URL as a command-line argument and:

- downloads the file

- arranges the files according to the DICOM hierarchy in an appropriate directory structure (patient/study/series). Note that patient names were replaced with IDs to protect their privacy.

>> **import** DICOM\_processor

>> DICOM\_processor.get\_dcm\_from\_url(**"http://www.example.com/dicom\_files/dicom\_file.dcm"**)

The "get\_dcm\_from\_url" function downloads the dcm file and saves it to the CWD and then calls the "hierarchical\_saving" function, which moves the file to the appropriate directory based on the structure "patient/study/series". If any branch is missing in the folders hierarchy then it is formed before moving the file to its destination.

The "get\_dcm\_from\_url" function can be wrapped in a loop and called repeatedly to download files given by a list. Since I didn't have a remote folder with dcm files, I downloaded the tar file you linked in the email and wrote another function, "get\_metadata\_list", which go over a list of dcm files in a given folder (./DICOM\_files in this case), creates a txt file with the hierarchy information (row for each dcm file, describing its PatientName, StudyInstanceUID, SeriesInstanceUID, filename).

Like the original "get\_dcm\_from\_url" function, the "get\_metadata\_list" function calls the "hierarchical\_saving" function which organize the files in the required hierarchical structure. The "hierarchical\_saving" function can be called with either 'copy' or 'move' argument, depending on whether you want the file/s in the source folder to be copied or moved.

1) generate a list of patients, their age and sex

>> DICOM\_processor.generate\_patients\_details()

The "generate\_patients\_details" function iterate over the structured folders, extracts the requested data from each dcm file, and saves it in a space-delimited txt file under the name "patients\_details.txt" in the CWD.

2) how many different hospitals do the data come from?

>> DICOM\_processor.recursively\_extract\_hospitals()

The "recursively\_extract\_hospitals" function interate over the structured folders, extract the "InstitutionName" from each dcm file, sorts & unique the list, and returns an output to the screen with the number of hospitals and the ordered list of their names.

3) explore the following DICOM tags, and try to explain what they mean, and the differences and relationships between them. Feel free to use appropriate visualizations as necessary.

- 0x0008,0x0013

- 0x0008,0x0032

- 0x0020,0x0012

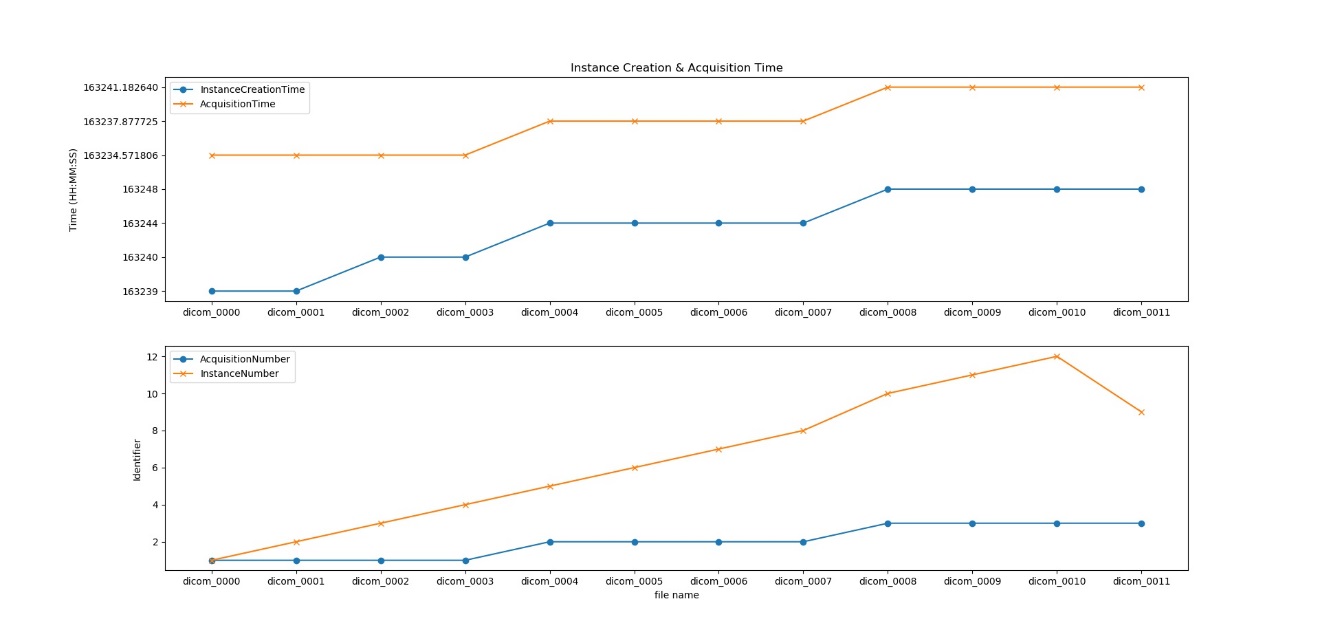
- 0x0020,0x0013

4) How long does a typical CT scan take?

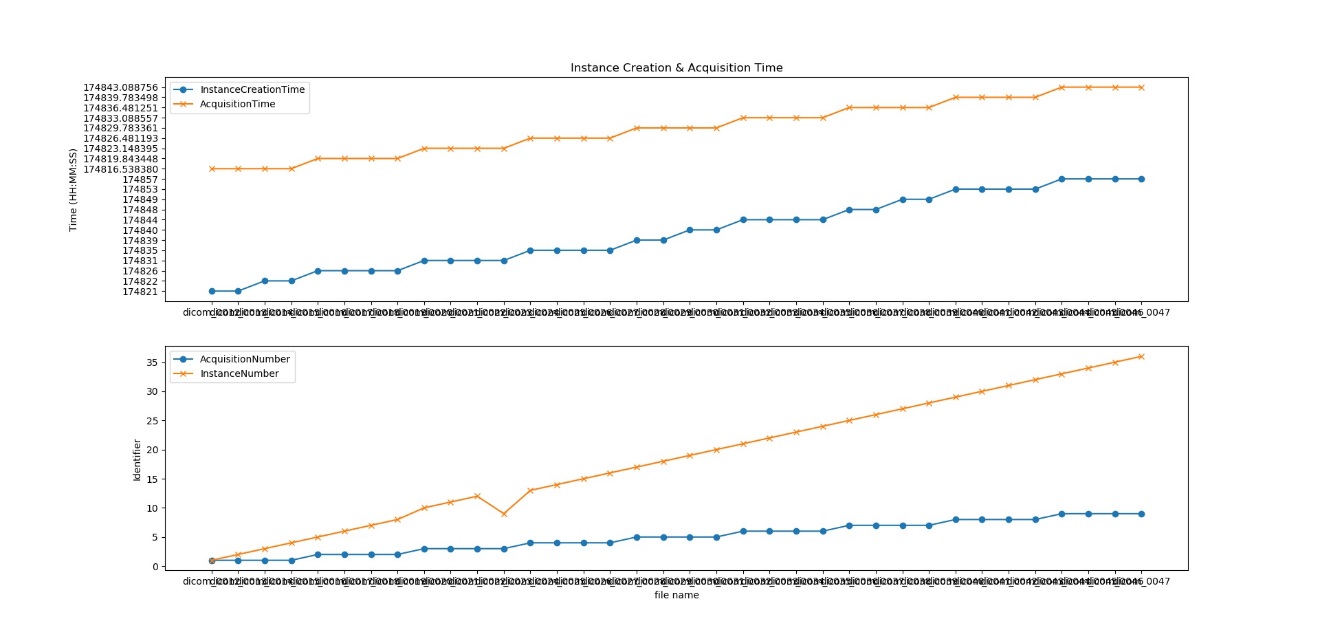
>> DICOM\_processor.explore\_DICOM\_tags(patientName)

The function goes over the requested list of tag numbers which are extracted from the dcm files stored under the folder of a given patientName. It then builds a filename vector for the x axis, and a vector of values for each of the tag numbers which are plotted as y values over each filename point. To plot the tag values for each point in the filename vector the 'plot\_subject\_DICOM\_tags' function is called with the 'x' & 4 'y' lists as arguments. Here's the result for two patientNames:

Subject **'1.2.840.113619.2.337.3.2831186181.442.1421722000.421'**:



Subject **'1.2.840.113619.2.337.3.2831186181.704.1420253349.16'**:



Based on visualizing these 4 tags in 2 patients, it seems that scans are taken in groups of 4, and that AcquisitionTime is locked to the beginning of each "quadruplets" and doesn't change within this period. InstanceCreationTime represents the time of each individual scan, but since the scan rate is high, in many cases this parameter is similar within a quadruplet, unless a second pass in between and then they don't have the exact InstanceCreationTime value.

For each scan, the AcquisitionTime is earlier than the InstanceCreationTime, and the gap between these 2 tags ranged between 5 to 14 seconds in the examined graphs.

The AcquisitionNumber strictly matches the AcquisitionTime pattern and in fact gives consecutive numeric value for each quadruplet, based on their chronological order (it numbers the quadruplets).

The InstanceNumber simply numbers the scans within a series according to the order of their occurrence.

The 'jumps' which occur sometimes on the InstanceNumber graph reveals that the numbering in the file names do not reliably represent the chronological order of the scans, but only roughly, and sometimes a later scan can be file-named before an earlier one.

This 'jump' in the InstanceNumber was not reflected in the InstanceCreationTime and the AcquisitionTime graphs, since it occurs within a quadruplet. As such, the AcquisitionTime tag is set to the most earlier scan in the quadruplet and will not change as a result from a 'switch' between the four scans. The InstanceCreationTime shows the time rounded to the second, without decimal points values, and therefore in most cases this period is not sensitive enough to be visualized in this graph as the scans within each quadruplet occurs much faster.

There's a period of 3 to 5 seconds between the starting point of each quadruplet.

The whole instance depends on the number of scans which are taken. For example, 40 scans mean 10 quadruplets, which means 10 x ~4 seconds (between each quadruplets), which gives ~40 seconds (a single scan takes less than a second).

Next, use a DICOM viewer suitable for your OS to view the scans. anything seems particularly interesting? You're obviously not expected to make a medical diagnosis, just spend 5-10 minutes browsing through the scans and share your general impressions.

I used 'MicroDicom viewer' to browse through the scans. In all patients the brain is presented in the transverse plane, and the images progress from ventral to dorsal. There are different types of CT – "CerebrumSeq" (with 4.8mm gap between scans), "5/5mm plain", and "thin plain", which scans in 0.63mm gaps.

- one of the patients has more than one series. Use a DICOM viewer suitable to your OS to explore the series and try to explain the differences.

Patient xxx…1414550448.623 has something that looks like a tumor (clearest in scan of file dicom\_0065.dcm) in the mid-frontal part of the brain. Perhaps this is the reason why she had another series, in addition to the "common" 5/5mm plain. The second series appears as "thin plain", and is much more dense - 0.63mm between each slice. Under this type of scanning the suspected tumor starts to appear in file 0191 up to file 0239, which means that on the ventro-dorsal axis its diameter is ~30.2mm.