Lab Assignment Week 1 - Getting familiar with Data types and Visualization

Part 1: Tabular Data Processing

Step 1: **Download** the dataset files belong to the following data formats from internet. The files may belong to any dataset available online. If these data formats are not available you can convert any of the existing data files to specific file formats.

Step 2: **Read** these files inside the python code. Some of the file formats cannot be read using default python packages. In this case, explore the python packages suitable for reading the files.

Step 3: **Print** the properties of the data files such as size, shape, dimensions, number of rows and columns etc.

Step 4: Perform matrix related operations such as transpose, inversion, dot product (with transpose of same matrix),

Step 5: **Visualize** each of these data files using graphs, diagrams, etc. Table data visualization: line graph, bar graph, histogram chart, pie chart, scatter plot (Choose applicable columns), Save the plot to disk

Step 6: **Save** the processed matrices as PKL – Pickle format, HDF5, Zip, SQL, MAT, NPY, NPZ. Reload the saved files back to the python environment.

Tabular, Spreadsheet and Interchange Data Formats

- "Table" — generic tabular data (.dat), "CSV" — comma-separated values (.csv), "TSV" — tab-separated values (.tsv), “ARFF” - Attribute-Relation File Format (.arff) – Read and visualize the data
- "XLS" — Excel spreadsheet (.xls), "XLSX" — Excel 2007 format (.xlsx), "ODS" — OpenDocument spreadsheet (.ods), "SXC" — OpenOffice 1.0 spreadsheet file (.sxc), "DIF" — VisiCalc data interchange format (.dif) – Read and visualize the data

Part 2: Audio Visual Textual Dataset Formats

1. Image Data Formats

- Download images belong to each of these image formats JPG, PNG, BMP, TIFF. Read and display the image
- Image reading can be done using Matplotlib/PIL/Opencv packages. Use all three for a single image reading and compare their executing time.
- Print properties of images such as height, width, number of channels.
- Perform operations such as conversion to binary, grayscale image formats, cropping the image, rescaling the image etc
- Download, read and visualize 3D medical Images belong to the formats DICOM, MHA

2. Video Data Formats

- Download any video you like in the formats MP4, AVI, MPEG. Read and play the video in your ipython notebook.
- Print properties of the video such as frame rate, frame height & width, total number of frames etc.
- Save frames of video in a separate folder. Read them again and convert back into original video (audio may me missing in the new video).

3. Audio Data Formats

- Download audio files in the formats such as MP3, MIDI, WAV. Read and play the audio
- Audio visualization: audio player, spectrogram
- Print properties of the audio files such as sampling rate, length in seconds, bps etc.
- Extract and visualize audio features such as MFCC, STFT etc.
- Librosa package can be utilized for the above tasks.

4. Text Data Formats

- Download text files in the formats such as TXT, PDF, DOC. Read and parse the data.
- Download files in the formats such as "JSON" — JavaScript Object Notation (.json), "UBJSON" — Universal Binary JSON (.ubj), “HTML” – Hypertext Markup Language (.html), “XML” - eXtensible Markup Language (.xml) - Read and Parse the data
- Explore different python packages for the parsing of different structured and unstructured text files
- Text visualization: Word cloud, bubble cloud (some more in http://vallandingham.me/textvis-talk/)

Submission: Submit your files in **Single ipython Notebook zipped with files (Zip file size lesser than 5mb)** in LMS before deadline**.**