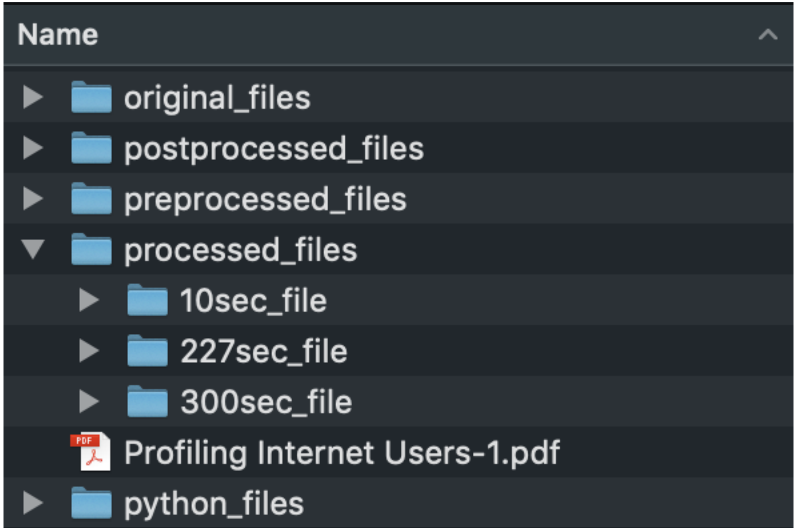
Sarvender Dahiya

Workflow

The project is done using python. Please install 'Pandas' and 'Scipy' libraries as they are required for the code to work. Please also install 'xlrd' (*pip install xlrd* **).**

❖ Project Structure:

The project folder has the following contents

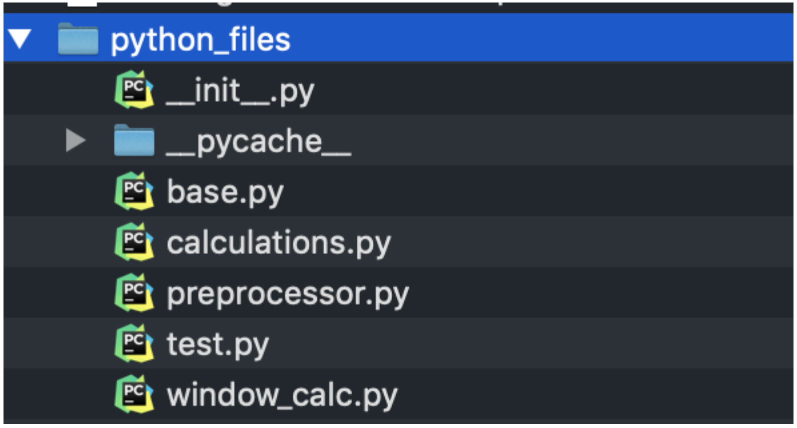


```
Name
▶ original_files
▶ postprocessed_files
▶ preprocessed_files
▼ processed_files
    ▶ 10sec_file
    ▶ 227sec_file
    ▶ 300sec_file
  Profiling Internet Users-1.pdf
▶ python_files
```

- **Original files: Should** Contain the Source files provided as input**.**
- **python_files** contains all the python files with the required code
- Please create all the other folders shown before proceeding as they are required. And add the source Files to the Original Files folder.

❖ Python Files:

NOTE: *The files are well documented with comments, and effort has been made to describe every action in the program.*



```
▼ python_files
    __init__.py
  ▶ __pycache__
    base.py
    calculations.py
    preprocessor.py
    test.py
    window_calc.py
```

You only need to run base.py in order to proceed with the project.

- **Base.py**: The main function defined in base.py does the following:

    - It asks for choice to perform the initial cleaning of the files, if chosen Y (yes) then it calls the preprocess function of the preprocessor.py, (see below) . Otherwise the program moves on.

    - Next choice is asked to process the clean files, which calls window_ generator function of the window_calc.py (see below).

    - Next it calls perform_calc function of calculations.py (see below)

- **Preprocessor.py:**

  # only read [Octets, Real First Packet, Duration] columns from the source file
  # only keep rows with Duration != 0
  # convert epochs from milisecs to secs
  # only keep rows in between Monday Feb 4 8 am and Friday Feb 15 5 pm
  # create new column 'doctets/Duration'
  # drop columns doctets and Duration as they are no longer needed
  # path to save the processed file and change format from .xlsx to .csv
  # Saves output to 'preprocessed' folder
  # results in size reduction from 566 MB of the original files to 56 MB of preprocessed files.

- **Window_calc.py:**

  # This function calculates a list containing windows,

  # Then calls the function window_value_calculator() for each file which returns two lists containing doctets/duration and and week for corresponding window

  # for each window size (and saves that into new data file in the folder for that respective window size)

  # Saves output to 'processed' folder [which contains subfolder for respective window sizes]

NOTE : *This file has been heavily documented with comments. Inside the files, effort has been made to document every task that the program logic is performing. Please see the file (window_calc.py) for detailed explainations.*

- **Calculations.py**:

The perform_calc function is used to # pass files to ‘calculate’ function

The calculate function creates a dataframe which contains the spearman coefficients (it passes every user from week 1, and calculates **spearman’s correlation coefficient** against every other user in week 2. Saves that to a list and appends the corresponding user's lists to the dataframe.)

# After this is done, the function proceeds to calculate Z and P values