**Dharmik Patel (U34190712)**

**Infosec Project – Profiling Internet users**

**WORKFLOW DOCUMENT**

In this project I have used python programming to code and achieve the requirements that are preprocessing the excel files and saving them to .csv format and the calculations on the preprocessed files. The downloaded excel sheets are stored in the folder named **InfosecData** and the preprocessed reduced data is present in folder: **processed\_files**. Now further the splitted data according to the window size is stored in folders: **win10, win227, win300**. I have selected the 1st week starting from **Monday Feb11 8AM to Friday Feb 15 5PM.**The second week starts from **Monday Feb18 8AM to Friday Feb 22 5PM.**

The **starterfile.py** is the main file from where the whole program needs to be executed. Most of the working of code is explained using comments. The function **process\_data()** is splitting the data in the time frame of two weeks selected as above using the Real First Packet epoch Time. Now the function process data is called and the csv files are made.Further we have called the **window\_func(win\_size)** where win\_size is 10,200 and 300seconds(5min) and this window function returns a week\_list where it says whether the RFP time falls in week 1(w1) or week2(w2).

The **win\_list.py** has function **window\_func()** that is doing the task of separating week 1 and week 2 using two for loops respectively while ignoring the weekends. The loop works as it starts from Day1 8AM to 5PM that is a 9 hour window and then it skips 24 hours from 8AM day1 to get start window of day2 8AM. It calls the comp\_window() function at the end and returns week\_list. The other function defined is **comp\_window()**  that does the task of checking whether the epoch time from the processed file falls in the relative window sizes and calculates the average doctates/duration for the epoch that falls in the window slot and assigns average as very small number (0.0001) for rest epoch times. At the end creating a new csv file with attributes {real first packet,average octet/duration} and saving them according to their win\_size slots and relative path.

**Spearman.py** has the function **spearman\_files()** which does the task of making a single file of all 54 files for each window sizes such as rows indicate week and column names indicate user names and their data correspondingly . These files will be further used in calculating spearman correlation and are stored in **spearmanV** folder with names win10, win227 and win300 respectively. These 3 files will be the input to next function used which is the **calculate()** function that actually calculates the spearman correlation and stores the final output to the **final\_output** folder.The **p\_calc()** function is calling the calculate() function in a loop for all those 3 spearmanV files which is the calculation for r1a2b.

***How to run ? You have to only run the starter.py file in pycharm and there you go!***