**Lab Assignment 04 - Feature Extraction**

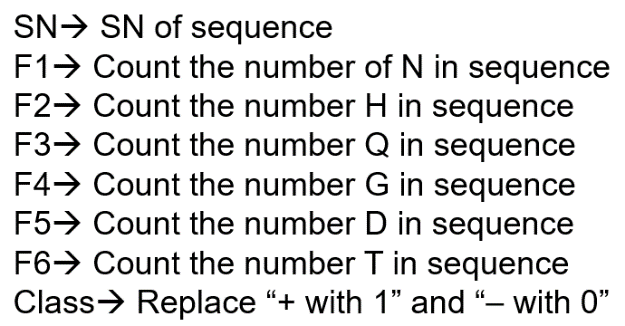
|  |
| --- |
| * Input data files are available in “**Input files for Assignment04**” folder |

**Program 1:** Develop a command line program to extract the features of a protein sequence file.

|  |  |
| --- | --- |
| **Input File** | **Output File** |
|  |  |

**1.1 Description:**

* Input file contain only two columns: first column is the protein sequence and second column is the class (either -ve or +ve).
* Extract the feature of the sequence as given below rule:



**1.2 Run the program** through command line as:

**python extractFeatures.py <inputFile1> <inputFile2> ………… n**

**Example:**

* python extractFeatures.py inputfile1.csv
* python extractFeatures.py inputfile1.csv inputfile2.csv
* python extractFeatures.py inputfile1.csv inputfile2.csv inputfile3.csv

**1.3 Input/Output Files:**

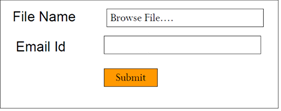
* Input File(s) 🡪 file1.csv | file1.csv file2.txt | file1.txt file2.csv file3.txt
* Output Files 🡪 One result file and one log file
  + **Result file:**
    - It contains the extracted features for every sequence present in the input file(s).
    - Name of the result file 🡪 “result-” + str(time.time()) + “.csv”
    - e.g. 🡪 “**result-20202109.csv**”
  + **Log file:**
    - * It contains three columns (inputFileName, Sequence, Class) having issues with the sequences or with the class label in the input file(s).
    - Missing sequence or sequences having any numeric value
    - Missing class label
      * Name of the log file 🡪 “log-” + str(time.time()) + “.csv”
      * e.g. 🡪 “**log-20200909.csv**”
      * Log file content

|  |  |
| --- | --- |
| FileName, Sequence, Class  file1.csv, AGERT5DCT, + | 🡨 Sequence contain numeric value |
| file2.csv, ARGVT, | 🡨 Sequence class is missing |
| file3.txt, , - | 🡨 Sequence is missing |
| file4.txt, A4ADER, | 🡨 Sequence contain numeric value  & class is missing |

**1.4 Check for:**

* Correct number of parameters
* Show appropriate message for wrong inputs.
* Handling of “File not Found” exception
* Input file(s) contain only two columns.
* Output file name will be “result-” + str(time.time()) + “.csv”
* Log file name will be “log-” + str(time.time()) + “.csv”

**Program 2:** Develop a web service for Program 1.



**2.1 User Should get:**

* Result File
* Log File