UCS538 – Data Science Fundamentals

Lab Assignment 8

Program 1: Develop a Python program to extract the features of a protein sequence file (download from LMS).

Input File

Sequence	Class
PGGGKVQIVYKPV	+
PGGGKVYKPV	-
PGGGKNAEVYKPV	-
PGGGKVQIVEKPV	-
QTAPVPMPDLKNVKVY	-
KPVDLSKVTSKCGSLGNIHLDF	-

Output File

SN	F1	F2	F3	F4	F5	F6	Class
1	4	0	2	3	2	1	1
2	5	3	3	2	2	2	0
3	4	5	2	3	3	4	0
4	3	4	0	2	5	5	1
….	….	…	….	…..	…..	…..	…..

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:

SN→ SN of sequence
F1→ Count the number of N in sequence
F2→ Count the number H in sequence
F3→ Count the number Q in sequence
F4→ Count the number G in sequence
F5→ Count the number D in sequence
F6→ Count the number T in sequence
Class→ Replace "+ with 1" and "– with 0"

1.2 Input/Output Files:

- Input File(s) → file1.csv
- 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-" + roll number + ".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-" + roll number + ".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

1.3 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: Write a program to implement TOPSIS (Download topsis.csv from LMS)

Input File

Model	Corr	Rseq	RMSE	Accuracy
M1	0.79	0.62	1.25	60.89
M2	0.66	0.44	2.89	63.07
M3	0.56	0.31	1.57	62.87
M4	0.82	0.67	2.68	70.19
M5	0.75	0.56	1.3	80.39

Output File

Model	Corr	Rseq	RMSE	Accuracy	Topsis Score	Rank
M1	0.79	0.62	1.25	60.89	0.55	5
M2	0.66	0.44	2.89	63.07	0.87	1
M3	0.56	0.31	1.57	62.87	0.6	4
M4	0.82	0.67	2.68	70.19	0.79	2
M5	0.75	0.56	1.3	80.39	0.66	3