Name : Abhishek Guleri
Roll No. : 185509
Lab : Data Mining Lab
Assignment No. : 4
Branch - CSE DD

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

1.1 : Description

- Input file contains only two 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 -> Serial number of sequence
  - F1 -> Count(#N) in sequence
  - F2 -> Count(#H) in sequence
  - F3 -> Count(#Q) in sequence
  - F4 -> Count(#G) in sequence
  - F5 -> Count(#D) in sequence
  - F6 -> Count(#T) in sequence
  - Class -> Replace “+ with 1” and “- with 0”

```
sr_no = 1 #increments with the upcoming rows
F1 = row['Sequence'].count('N')
F2 = row['Sequence'].count('H')
F3 = row['Sequence'].count('Q')
F4 = row['Sequence'].count('G')
F5 = row['Sequence'].count('D')
F6 = row['Sequence'].count('T')
Class = '1' if row['Class'] == '+' else 0
```

1.2 : Run the program through command line as:
python extractFeatures.py <inputFile1> <inputFile2> …… n

- **inputFileX** are **proteinSequenceFileX.csv** where **X = 1, 2, 3**

```
python extractFeatures.py proteinSequenceFile1.csv proteinSequenceFile2.csv
proteinSequenceFile3.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”

- ***result = "result-" + str(time.strftime("%Y%d%m")) + ".csv"***

  Example : result-20221802.csv

- **Result file content:**

SN	F1	F2	F3	F4	F5	F6	Class

- Log file:
  - It contains three columns (inputFileName, sequence, Class) having issues with the sequence or with the class label in the input file(s).
  - Missing sequence or sequences having any numeric value
  - Name of log file -> "log-" + str(time.time()) + ".csv"
    - ***log = "log-" + str(time.strftime("%Y%d%m")) + ".csv"***

      Example : log-20221802.csv
  - **Log file content:**

inputFileName	Sequence	Class

1.4 : Check for:

- Correct number of parameters

```
if re.search(r'\d', row['Sequence']) or row['Sequence'] in (None, "") or
row['Class'] in (None, "") :
```

- Show appropriate message for wrong inputs

```
print(sys.argv[i], "either missing a value or the protein sequence is contains a
numerical value at row ", i)
```

Works parallel with the log writer

Code: extractFeature.py

```
import csv
import sys
import time
import re #regex

def main():
	n = len(sys.argv)

	result = "result-" + str(time.strftime("%Y%d%m")) + ".csv"
	log = "log-" + str(time.strftime("%Y%d%m")) + ".csv"

	with open(result, 'w', newline='') as result_csv, open(log, 'w', newline='')
as log_csv:
	result_fieldnames = ['SN', 'F1', 'F2', 'F3', 'F4', 'F5', 'F6', 'Class']
```

```
	result_writer = csv.DictWriter(result_csv, fieldnames=result_fieldnames)
	result_writer.writeheader()

	log_fieldnames = ['inputFileName', 'Sequence', 'Class']
	log_writer = csv.DictWriter(log_csv, fieldnames=log_fieldnames)
	log_writer.writeheader()

	sr_no = 1

	for i in range(1,n):
		with open(sys.argv[i]) as inputfile:
		fileReader = csv.DictReader(inputfile)

		for row in fileReader:
			if re.search(r'\d', row['Sequence']) or row['Sequence'] in
(None, "") or row['Class'] in (None, "") :
				log_writer.writerow({'inputFileName' : sys.argv[i],
'Sequence': row['Sequence'], 'Class' : row['Class']})
				sr_no -= 1
				print(sys.argv[i], "either missing a value or the protein
sequence is contains a numerical value at row ", i)

			else:
				F1 = row['Sequence'].count('N')
				F2 = row['Sequence'].count('H')
				F3 = row['Sequence'].count('Q')
				F4 = row['Sequence'].count('G')
				F5 = row['Sequence'].count('D')
				F6 = row['Sequence'].count('T')
				Class = '1' if row['Class'] == '+' else 0
				result_writer.writerow({'SN' : sr_no, 'F1' : F1, 'F2' :
F2, 'F3' : F3, 'F4' : F4, 'F5' : F5, 'F6' : F6, 'Class' : Class})
				sr_no += 1

if __name__ == '__main__':
	main()
```

Result File:

```
SN,F1,F2,F3,F4,F5,F6,Class
1,1,0,0,2,2,3,1
2,0,2,0,1,1,0,1
3,2,0,1,2,0,0,1
4,0,0,1,0,0,1,0
5,1,0,0,1,0,0,0
6,0,1,1,1,0,1,0
7,0,0,0,0,0,1,1
8,0,0,0,2,0,1,0
```

9,1,0,0,0,0,1,1
10,1,0,0,0,0,0,1
11,0,0,1,1,1,0,0
12,1,0,0,0,0,0,1
13,1,1,0,1,1,2,0
14,0,1,1,2,2,0,1
15,0,0,0,0,0,1,1
16,0,0,1,0,0,1,1
17,0,0,0,0,1,0,0
18,0,0,0,0,2,0,0
19,0,1,0,0,0,0,0
20,0,0,0,0,2,0,0
21,1,0,0,1,0,0,0
22,1,0,0,1,0,0,0
"result-20221802.csv" [dos] 1147L, 21816B

Log File:

inputFileName,Sequence,Class
proteinSequenceFile1.csv,PGGGKV3KPV,-
proteinSequenceFile1.csv,NLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVE,
proteinSequenceFile1.csv,PG33GKVQIVEKPV,-
proteinSequenceFile1.csv,KDRVQSKIGSLDNITHVPGGGN,
proteinSequenceFile1.csv,QTAPVPMPDLKNVKSKIGSTE,
proteinSequenceFile1.csv,KPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDF,
proteinSequenceFile1.csv,PGGGKN8EVYKPV,-
proteinSequenceFile1.csv,QTAPVPMPDLKNVKSKIGS67ENLKHQPGGGKVQIVY,-
proteinSequenceFile1.csv,PGGG5VQIVYKPV,+
proteinSequenceFile2.csv,QTAPVPMPDLKNVKSKIGSTE,
proteinSequenceFile2.csv,PGGGKN8EVYKPV,-
proteinSequenceFile2.csv,PGGG5VQIVYKPV,+
proteinSequenceFile2.csv,QTAPVPMPDLKNVKSKIGS67ENLKHQPGGGKVQIVY,-
proteinSequenceFile2.csv,PG33GKVQIVEKPV,-
proteinSequenceFile2.csv,NLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVE,
proteinSequenceFile2.csv,KDRVQSKIGSLDNITHVPGGGN,
proteinSequenceFile2.csv,PGGGKV3KPV,-
proteinSequenceFile2.csv,KPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDF,
proteinSequenceFile3.csv,PGGGKN8EVYKPV,-
proteinSequenceFile3.csv,QTAPVPMPDLKNVKSKIGS67ENLKHQPGGGKVQIVY,-
proteinSequenceFile3.csv,KDRVQSKIGSLDNITHVPGGGN,
proteinSequenceFile3.csv,KPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDF,
"log-20221802.csv" [dos] 27L, 1357B

Program 2: Develop a web service for Program 1.
2.1 : User should get:

- Result File
- Log File

Source Code : GitHub
Hosted on : FeEX