CS594: Python Programming Lab

Take Home Assignment - 3 (2 Questions, 100 Points)
Submission Dead Line: 12-Sep-2019 23:59 Hours   Pages: 4

IIT Guwahati

29 Aug 2019 (Thu)

Question 1: (0 points)
Reading assignment from the text book

Chapter 8 Text Files

Chapter 10 Object-Oriented Programming

Regular Expressions Python programming document available `https://docs.python.org/3/howto/regex.html`

Question 2: (100 points)
Implement in Python Programming Language the following problems:

Q1. 70 Marks You are given all the ICC cricket world cup 2019 matches information. All the matches details are available at `https://www.espncricinfo.com/scores/series/8039/season/2019/icc-cricket-world-cup`. This resource includes `SCORECARD, REPORT`, and `SUMMARY` links for each match.

Your task is to:

1. Download all the 48 matches cricket text commentaries into files on your local system. For each commentary file, adhere to the file name convention as `roll-number-match-id-commentary.txt`. Locate the match ID at: `SUMMARY` link. Look for the string `Match number` under the section `MATCH DETAILS`.
2. Download all the 48 matches scorecards. For each scorecard file, adhere to the file name convention as `roll-number-match-id-scorecard-published.txt`
3. Your program should take as input the above text commentary files pertaining to one match. Read the text commentary from the given file. Your task is to *compute the scorecard* given the text commentary. Output scorecard should be *identical in format* as given at the `SCORECARD` link. Store the scorecard in an output file with the name `roll-number-match-id-scorecard-computed.txt`.
4. Scorecard should contain the following three sections:
   (a) Batsmen
   (b) Fall of wickets
   (c) Bowling
5. You have to compute scorecard for all the 48 matches.
6. You `should NOT use` the following information present in the text commentary
   - Summary lines written just after the wicket taken by a bowler such as in the URL `https://www.espncricinfo.com/series/8039/commentary/1144483/england-vs-south-africa-1st-match-icc-cricket-world-cup-2019?`

`innings=2&filter=full`
`Over 25.5` use of lines of the form
JP Duminy c Stokes b Ali 8 (12m 11b 1x4 0x6) SR: 72.72

- End of over summaries such as in the same URL above, at the end of 26 overs summary available in a grey box presented as
  END OF OVER:26 | 5 Runs 1 Wkt | SA: 142/4 (170 runs required from 24 overs, RR: 5.46, RRR: 7.08)
- Use of the above information to compute scorecard leads to penalization of 50% of marks Batting Section marks. That is you will loose 10 marks for using this information.

7. **You are permitted** to use the above information to extract number of minutes played by a batsman as this information cannot be `computed` from text commentary alone.
8. You must use regular expressions for processing each line of the text commentary.
9. You must use two classes one for constructing batting part of the scorecard; other for the bowling part of the score card.

Q2. 30 Marks Read the two files `roll-number-match-id-scorecard-published.txt` and `roll-number-match-id-scorecard-computed.txt`.

Compare the output scorecard obtained by your program with the published scorecard. Your task is to

1. Identify the differences between `computed` scorecard and the `published scorecard`
2. Print the difference in a readable format
3. Print the total number of mismatches in each match

Instructions File Naming Convention Create a directory with your roll number. Inside this directory, place all the above python programs and input files. Prefix the file name with your roll number followed by "_" followed by question number followed by ".py". Example: 194161000_q1.py.

Input File Naming Convention As described in the question.

README.txt Write a short notes on sequence of steps involved to run the your programs. Include what is the input for the program (with an example) and what will be the output from the program (with an example).

tar gzip Create (roll number).tar.gz file using the above directory. This directory must contain the following:

Q1 and Q2. The input files prepared for Q1 and Q2

Q1. Python program solution for Q1

Q2. Python program solution for Q2

README Instructions to run your program must be placed in README.txt file.

Submission Email the above tar gzip file to the CS594 TA `vaibhav18@iitg.ac.in` as per the above given dead line

Copying You should avoid indulging in copying. Every submission will be subject to software similarity using the tool `Measure of Software Similarity` available at `https://theory.stanford.edu/~aiken/moss/`. Two submissions having similarity score equal to or more than 40.0% will be declared copied. If you are found involved in copying act, your name will be referred to disciplinary committee. Therefore you are requested to place individual efforts and avoid copying.

Marking Scheme Your implementation will be evaluated as described below.

Q1 5 Marks Downloading 48 published scorecards

Q1 5 Marks Downloading 48 text commentaries

Batting Section 20 marks distributed as given below

1. 10 marks for each line of batsman data. Each line of batsman in the scorecard contains the following information. Correct population of each of these information will be awarded 1.25 marks.
   - Batsman name
   - Description of how batsman got out
   - Runs scored
   - Balls faced
   - Minutes played
   - Number of FOUR's scored
   - Number of SIX's scored
   - Strike rate (SR)
2. 2.5 Marks for Identifying not out batsman
3. 2.5 Marks for computation of extras (total runs and their details)
4. 2.5 Marks for computing total runs scored by the team
5. 2.5 Marks for identifying batsman who did not bat

Fall of wickets 10 marks

Bowling Section 20 marks distributed as given below

1. For each bowler line appearing in the scorecard the following information needs to be computed. Correct population of each of this information will be awarded 1.82 marks.
   - Bowler Name
   - Number of overs bowled
   - Number of maiden overs bowled
   - Number of runs conceded
   - Number of wickets taken
   - Economy rate (ECON)
   - Number of 0's
   - Number of FOUR's yielded
   - Number of SIX's yielded

- Number of wide balls
- Number of no balls

10 Marks Use of regular expressions for each ball of text commentary 1.67 marks for each of the following:

1. Identifying bowler
2. Identifying batsman
3. Identifying number of runs scored
4. How wicket was fallen
5. Identifying extra runs
6. Identifying run out information

Q2 10 Marks Identifying the differences

Q2 10 Marks Printing the difference in a readable format

Q2 10 Marks Printing the total number of mismatches in each match