**INDIAN SCHOOL OF BUSINESS**

**Data Collection Group Assignment**

**Instructor: Dr. Manish Gupta**

**Honor Code Scheme**: 2N-b **Weightage:** 40%

Collecting Cricket World Cup Data

**Goal**: To understand basic web page scraping.

**Expected time**: 10 hours per person.

**Group**: Team of 2-3 people.

**Instructions:**

1. This is a group assignment consisting of **three** parts.
2. You need to have a machine with Internet connectivity.
3. You should write code in **Python**. **You are not allowed to use automated scraping tools like Octoparse.**
4. Any late submission will attract a **penalty** as mentioned in the course outline.
5. Zip file submissions consisting of all the deliverables will not be accepted.
6. Only one person in the group is supposed to do the submission in LMS.
7. Please mention PGIDs and names in the .py files and the readme file.

**Deliverables:** There will be **two** deliverables for each part of the assignment

1. **A python file with .py suffix – Please do not submit Jupyter Notebooks. Submit Python files which have a .py suffix which can be run using python filename.py.** The code should be **well commented**. Mention the **PGIDs and the names** of the group members on the top of the script.
2. The **csv file** containing all the scraped material, the details of which is mentioned below for each part. Wherever there are multiple things involved in the same column, use a **comma delimiter**.

Apart from this a combined **readme file** for all the 3 parts is to be submitted which contains any specific details that you want to convey to the TA, such as missing values in the csv or any instructions for executing the .py files.

The readme file is to be named as **[your group no]\_readme.txt**. Mention the **PGIDs and the names** of the group members on top of the file.

**So, in total there will be 7 deliverables for each group:**

1. **Three .py files – one for each part,**
2. **Three csv files – one for each part and**
3. **1 readme file.**

The task is to collect data from <https://www.espncricinfo.com/> for [ICC Cricket World Cup 2019](https://www.espncricinfo.com/series/icc-cricket-world-cup-2019-1144415/match-results). The assignment has 3 parts:

(1) collect match results info

(2) collect match details info

(3) collect player details info.

**Out of 40 marks, 10 marks are for part 1, 15 marks for part 2 and 15 marks for part 3.**

**Part 1:**

From this page: <https://www.espncricinfo.com/series/icc-cricket-world-cup-2019-1144415/match-results>, scrap the following information

1. Match number

2. Location

3. Date

4. Winning country

5. Other country

6. Match result

7. Score by winning country

8. Score by other country

9. Link to match report

10. Link to match summary

11. Link to match scorecard

**Save results in a file [your group no]\_matchResults.csv.**

**Name the .py file as [your group no]\_matchResults.py.**

**Part 2:**

For each match, go to the scorecard link like <https://www.espncricinfo.com/series/icc-cricket-world-cup-2019-1144415/india-vs-new-zealand-1st-semi-final-1144528/full-scorecard> and extract the following:

1. Player of the match with the picture. Save the url to the picture in the csv.

2. Country that the player of the match belongs to.

3. Runs scored by every batsman.

4. Balls played by every batsman.

5. Strike rate for every batsman.

6. Wickets taken by every bowler.

7. Economy rate for every bowler.

8. which country won the toss.

9. who were the umpires?

10. who was the match referee

**Save results in a file [your group no]\_matchDetails.csv.**

**Name the .py file as [your group no]\_matchDetails.py.**

**Part 3:**

For each player across all matches (using the player page like <https://www.espncricinfo.com/newzealand/content/player/506612.html>) extract the following:

1. Full name of player.

2. Date and place of birth.

3. Current age.

4. Major teams.

5. Playing role.

5. Batting style.

6. Bowling style.

7. Highest ODI batting score.

8. ODI debut information.

9. Profile information.

10. Pic of the player. Save the url of the image in the csv.

11. Country of the player.

**Save results in a file [your group no]\_playerDetails.csv.**

**Name the .py file as [your group no]\_playerDetails.py.**

**Other notes:**

1. Selector Gadget tool will not get you everything. You will have to try out various things out of what you learned in the class.
2. You can use beautifulsoup or scrapy to scrap content from webpages.
3. You will need to download relevant pages.