

INTRODUCTION

The students will be appointed as new Data Scientists by a new startup looking to disrupt the space weather reporting business. Their first project is to provide better data about some phenomena and/or events than the other competitors. Therefore, the students have to get the data that has been published by the strongest competitor. Besides, they will be pointed to another messy HTML page where extra data is available to be posted by the new spiffy site of the new startup.

Of course, the students have no access to the raw datasets, therefore, as enterprising data scientists, they will scrape this information directly from each HTML page using all the great tools available to the students in Python.

In addition to the technical details, data sources and initial codes, the students will be provided with information sources to read up a bit on the target space weather phenomena and events upon releasing this project to them.

PART ONE: Data Scraping and Preparation (15 Marks)

The student will be asked to do the following tasks: -

1. Scrape the data from the strongest competitor's website. (4 Marks)
2. Clean and tidy the scraped data, deal with the missing records and drop the extra columns. (3 Marks)
3. Scrape the extra data from a messy HTML. (3 Marks)
4. Tidy the extra data and deal with the missing observations, according to detailed instructions. (5 Marks)

PART TWO: Data Analysis (15 Marks)

The student will be asked to do the following tasks: -

1. Replicate the data published by the strongest competitor using the extra data obtained from messy HTML. The student should discuss in detail the criteria she/he used to replicate the data and how she/he implemented such criteria. (3 Marks)
2. Integrate the two datasets for just the top 50 space phenomena (top 50 solar flare events as mentioned in the competitor's website), applying the best matching between the two data sets. The student should discuss the best matching between the datasets and how such matching might lead to losing

some records. Then, the student should identify the lost records (if any) and comment on the accuracy of the analysis. (4 Marks)

3. Plot attributes of the aforementioned space phenomena over time. Use graphical elements to indicate the top 50. From the plot, the student should comment on the value of some attributes with respect to the top 50 phenomena. (4 Marks)
4. Provide detailed comments on how some space phenomena do cluster in time. In this context, the student should do the necessary plot(s) to derive the comment. (4 Marks)