

Operational Guide Data Management Project

Eleonora Brambatti (858098), Matteo Pasotti (901810), Marta Privitera (898017)

18th September 2023

Our folder contains different sub-folders in which we can find all the codes and JSON files created. A brief explanation of the contents:

1. The WebScraping codes can be found in the “Data Acquisition” folder, in particular we have to make reference to 3 sub-folders:
 - “Seasons”, which contains the four python codes about each year;
 - “Qualifying_Max.Speed”, which contains four folders with the pdf we downloaded for each year and then the python codes of conversion from pdf to JSON for each year;
 - “Weather”, which contains two folders, one for the Qualifying session and the other for the Race where we can find the python code for each year.
2. The output of these WebScraping can be found in the “RawDataCollected” folder which contains the “WebScraping” folder in which the JSON output relative to Season, Max Speed and Weather Conditions are collected (divided by sub-folders).
3. In the “Data Acquisition” folder we can also find the codes relative to the merge between Season JSON, Maximum Speed JSON and Weather Conditions JSON. In particular they are contained in the “Merge” folder. The outputs of the merge process are contained in the “Merged_Data” folder in which there are the integration between Season JSONs and Maximum Speed JSONs. The JSONs with the last integration plus Weather Condition are stored in the “Final_Data” folder;
4. The “Data quality” folder contains several codes and JSON files:
 - The ones to check data quality on season JSON (*rawdata_seasons.py* which has as output the *laps2013.JSON*, *laps2014.JSON*, *laps2021.JSON* and *laps2022.JSON*), Max Speed JSON (*rawdata_speed.py*) and Weather Condition JSON (*rawdata_weather.py*);
 - The four python codes for doing WebScraping (e.g *laps2013fia.py*) on the FIA website with respective outputs (e.g *laps2013fia.JSON*) and the code that checks correspondences between FIA website and our JSON files (*lapscheckfia.py*);
 - The two python codes about the data quality check on merged data (*merge_speed.py* and *merge_weather.py*);
 - The python code for data quality on stored data (*finaldata.py*)
5. Finally, the “Storage” folder contains:
 - The python codes that make connection and insertion of the final JSON files to MongoDB;
 - The 3 query codes we used for our analysis and another code to plot useful graphs we used and inserted in the Report (*DataManagementReport.pdf*).

All the explication of these codes can be found in the Report.

To access the final database on MongoDB the owner needs to send an invitation to an email associated with a Mongoddb account. The owner’s email address is m.pasotti5@campus.unimib.it.