DATA MANAGEMENT PROJECT

World Billionaires Analysis – Operational guidelines

Alessandro Belotti, Alessandro Castagna, Ilaria Salvatori

This document has the aim of explain in a clear way the structure of the project folder and how to run the code.

1. Structure

The submitted folder has the following structure:

→ ***DataMan_belotti_castagna_salvatori*** (the main folder)

→ ***original_data*** (raw datasets)

→ ***exported_data*** (scraped datasets and others pre cleaned datasets)

→ ***final_data*** (datasets used in MongoDB and the MongoDB output with all the aggregated data)

→ ***project_notebook.ipynb*** (python notebook of the project)

→ ***report.pdf*** (paper related to our study)

→ ***presentation.pdf*** (short description of the project)

→ ***code_guidelines.pdf*** (operation steps for running the project)

2. Python Notebook

The python file has been developed by using Jupyter Notebook. For this reason it's not required any virtual environment installation or setting.
All the required libraries can be installed by running the first chunk of the notebook.
The python files contains all the pipeline phases:

- Data Acquisition
- Data Cleaning and Enrichment
- Schema Matching (here the data preparation for mongo is developed)
- Data Quality
- Research Question

Some chuncks (like the scraping ones) require a lot of time for the computation, in order to speed up the billionaires analysis we have inserted below each part a section for the uploading of the pre computed files (it's highlighted in specific chunks).

3. Other contents

The submitted folder also contains a document which describes all the analysis and the phases of the project (*report.pdf*) and also a presentation which underlies the main concepts of the project (*presentation.pdf*).