Documentation: data scraping and structuring (scraping_stl.py)

1. Data sources

This script extracts data from rankings and individual runner pages from the LiveTrail.net website, for the 2013 to 2024 editions of the SaintéLyon ultra-trail.

The runners' individual pages, which contain their lap times, have a URL that depends on their bib number. Many bib numbers correspond to other SaintéLyon race formats, or are not assigned, so rather than loop through all possible bib numbers, two types of pages are used:

- **General ranking pages**: used to retrieve bib numbers specifically for the SaintéLyon.
- **Individual runner pages**: used to obtain lap times, rankings and personal information, as well as checkpoint information.

2. Code principle - scraping

1) Retrieve the source code of the general ranking page in XML,
2) Extraction of bib numbers using a regular expression, and checkpoint information
3) Loop over all bib numbers:
4) Retrieve source code for each individual runner page in XML,
5) Extraction of all information, via specific identified tags.
6) Loop for each year.
7) Loop for the special case of 2017

Note on maintainability: the LiveTrail site is currently being updated, and the scraping used here is no longer functional on the beta of the new version.

3. Data cleaning

Certain choices have been made in order to obtain the following results in a form as clean as possible :

- All editions have a total of 7 points of interest with recorded times (including Saint-Etienne with time 00:00), but the two last editions had 8. The added points of interests, “Animation 500m” in 2024 and “KM BV SPORT” in 2023, located at 600m from the finish line and 1.4km from the start respectively, were ignored in the scraping for simplification : that way, every edition’s point number 6 (0-indexed) is the finish line. Besides, the time split data from these points probably doesn’t bring much value.
- All past achievements have been stored in a dictionary, to be treated (or not) later.
- A 2nd csv file was created to store information about each point of interest : name, distance from start, height, elevation from start (all differ every year).
- Some other data fields were ignored, deemed irrelevant to the study or repetitive, but they could be scraped just like the rest.

4. Limitations and constraints

Missing data :

- 2020 edition cancelled (COVID)
- Individual pages from 2017 are not accessible. Global results had to be retrieved on the global ranking page, with a modified version of the script. These results unfortunately contain less information than the individual pages.
- Some lap times are missing (chip problems?)
- Some anonymized bib numbers may not contain all the information.

Differences between editions :

- Some data fields differ slightly in the source code between editions.

5. Data output

Results are stored in two CSV files (Saintelyon_Results.csv, Saintelyon_checkpoints.csv), saved for each loop over one year to avoid data loss in the event of program shutdown. The files contain :

- Runner identity (surname, first name, club, nationality).
- Rankings (general, category, gender).
- Checkpoint times in dictionary form. As checkpoints are not necessarily the same from one edition to the next, the appropriate formatting of these is left to the data analysis section.
- Performance history for other editions.
- Information about the checkpoints for each year.