Albert Amurgo i Àlex Peláez

PRÀCTICA 1:
Trail Running Races 2020-2029

Context

Les curses de muntanya cada cop tenen més adeptes i l'ITRA [1] és una organització que pretén fomentar la pràctica d'aquest esport. Una de les seves missions és calcular un índex per cada corredor (ITRA Score) per tal de comparar corredors tot i que aquests no competeixin en les mateixos esdeveniments.
Per tant, l'ITRA publica el llistat de curses homologades per a tal de calcular l'índex. El càlcul de l'índex no és públic i només que un corredor participi en una cursa oficial ja té el valor calculat per a poder comparar tot i que és a partir de 5 esdeveniments en 3 anys que l'índex té un resultat més acurat.
Com a aficionats a les curses de muntanya ens sembla interesant poder tenir un data set de tots els esdeveniments oficials per a poder planificar els calendaris de curses. Partint d'aquestes dades també es podria plantejar l'elaboració d'algun tipus d'indicador sobre alguna característica de les curses (per exemple la seva sostenibilitat). Addicionalment, altres organitzacions podrien fer servir aquest data set per a oferir serveis relacionats (agències de viatges, entrenadors personals, etc.).

Títol del dataset

Definir un títol pel dataset. Triar un títol que sigui descriptiu.

- Curses ITRA 2020-2029

Descripció

Tal i com indica el títol, la base de dades mostra les curses homologades per l'ITRA en el període 2020-2029. Per a cada cursa es mostren els seus atributs principals. La descripció dels camps es descriu una mica més avall. El format de la base de dades és un fitxer CSV que facilita la seva visualització i tractament.

Representació gràfica

L'ITRA es divideix en 5 regions geogràfiques amb centenars o milers de curses anuals.

[1] *ITRA* [en línia] [data de consulta: 30 d'Abril de 2021]. Disponible a: https://itra.run/

Albert Amurgo i Àlex Peláez

Figura 1: Distribució geogràfica de les curses del dataset

Contingut

Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

El data set conté les dades de les curses en el període 2020-2029 publicades. És una base de dades en constant canvi degut a nous esdeveniments que s'hi afegeixen i eliminen. És especialment dinàmica durant els últims mesos degut a la situació de pandèmia actual on moltes curses no es poden confirmar fins a poques setmanes abans de la data d'inici.

A la "Figura 2" es mostra un esquema de l'estructura seguida durant el desenvolupament:

Albert Amurgo i Àlex Peláez

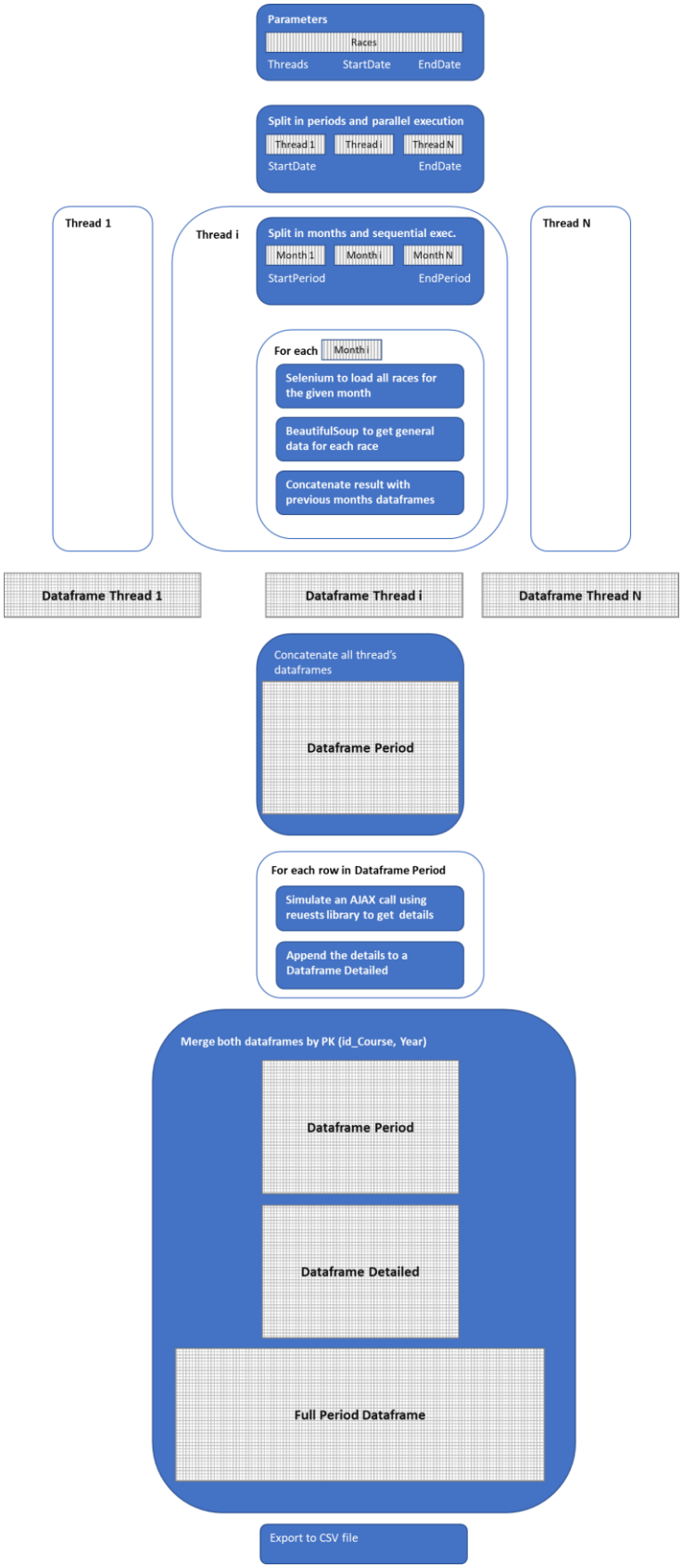

Figura 2: Esquema de la solució

Albert Amurgo i Àlex Peláez

Podem veure com, inicialment, es parteix el període en tants subperíodes com fils d'execució desitgem i executem en paral·lel cada fil.
Cada fil torna a dividir el subperíode en subperíodes de 30 dies (mesos) ja que hem detectat errors quan l'explorador manegat per Selenium carregava molts continguts. Per a cada període mensual obrim una nova instància del navegador i:

- Carreguem totes les curses del període;
- Amb BeutifulSoup obtenim les dades principals;
- Concatenem totes les dades dels mesos anteriors.

Un cop han acabat els fils executats es concatenen tots els dataframes i generem la clau primària (ID, Any)
Per a cada cursa fem una crida AJAX per a obtenir els detalls. Donat que és una execució molt ràpida 6 requests/segon no hem cregut oportú fer una execució paral·lela. Amb els detalls anem construint una dataframe amb la mateixa clau primària (id_course, annee) i finalment fem un merge amb l'inicial. Per acabar, exportem el dataframe a un fitxer CSV.
El data set conté els següents camps:

Camp	Descripció	Exemple
Name	Nom prova	Tnf 100 Malaysia 2020 - 100km Ultra Challenge 100.40 K
Link	Enllaç	https://itra.run/race/24223
Distance	Distància	100.40 km
Date	Data	05/12/2020
id_course	Identificador de la cursa	24223
nom	Nom cursa	100KM Ultra Challenge
annee	Any	2020
nb_pts_mont	Punts muntanya	7
pays	País	MY
type_partic	Tipus de participació	solo
pct_route	% de carretara	20
pct_piste	% de pista	50
pct_sentier	% de senders	30
challenge	Repte	Asia Trail Master
championnat	Campionat	TNF Ultra Challenge
inscr_nb	Número d'inscrits	700
inscr_dt_min	Data inici inscripcions	43862
inscr_dt_max	Data fi inscripcions	44027
inscr_sur_place	Inscripcions presencials (Sí/No)	1
inscr_tarif	Preu inscripció	0
inscr_tarif_devise	Divisa inscripció	0
nb_etp	Nombre d'etapes	1

Albert Amurgo i Àlex Peláez

Camp	Descripció	Exemple
latitude	Latitud	4.85468
longitude	Longitud	100.744
pays_arrivee	País d'arribada	0
ville_depart	Localitat de sortida	Taiping, Perak
ville_arrivee	Localitat d'arribada	Taiping, Perak
pays_depart	País de sortida	0
inscr_site	Link	0
heure_depart	Hora de sortida	44170.08
dist_tot	Distància	100.4
deniv_tot	Desnivell positiu	5680
deniv_neg_tot	Desnivell negatiu	5680
nb_pts	Punts ITRA	5
temps_max	Temps màxim	115200

Agraïments

El propietari de les dades és l'ITRA i s'ha fet servir la seva web: https://itra.run/races. El fitxer robots.txt no indica cap restricció:
User-agent: *
Disallow:

Hem pogut trobar un projecte similar d'extracció de les dades de les curses de l'ITRA i els corredors que hi han participat fins al 2018.[2] Aquest projecte ja no és vàlid ja que la web ha canviat i el codi que es feia servir per extreure les dades ha deixat de funcionar. Les dades extretes en aquell projecte es van fer servir per escriure un article científic.[3] D'alguna manera el nostre projecte pren el relleu i permet donar continuïtat a la possibilitat d'extreure i analitzar les dades de l'ITRA.

Inspiració

Tal i com s'ha indicat, l'anterior projecte d'extracció de les dades de l'ITRA ja no és compatible amb la nova versió de la pàgina web. Aquest nou projecte d'extracció es pretén respondre les següents preguntes o situacions:

- Organització del calendari de curses de muntanya dels aficionats;

[2] El codi i el data set d'aquest projecte es troba disponible a github al següent enllaç: ricfog/ScrapITRA

[3] Veure *Fogliato, Riccardo, Natalia L. Oliveira, and Ronald Yurko. "TRAP: a predictive framework for the Assessment of Performance in Trail Running." Journal of Quantitative Analysis in Sports 1.ahead-of-print (2020).*

- Creuar dades amb d'altres data sets (per exemple, amb vols, hotels, situació de la pandèmia, etc.)
- Analitzar o classificar les curses en funció del seu desnivell o del tipus de sòl pel que passen (senders, camins, carreteres);
- Creuar les dades de les curses amb la dels països d'origen dels corredors que hi participen i determinar la petjada de carboni que suposen els desplaçaments relatius a aquests esdeveniments;
- Analitzar l'impacte de la pandèmia en la cancel·lació de curses;
- Calcular els punts que atorguen les curses per metre de desnivell acumulat (algunes curses potser són comparativament més assequibles).

Llicència

Hem escollit publicar el data set sota la llicència **CC BY-NC-SA 4.0 License** [4] per no permetre l'explotació comercial de les dades propietat de l'ITRA.
Aquesta llicència permet **compartir** (copiar i redistribuir) el material en qualsevol mitjà i format. També permet **adaptar** el material, per exemple, transformant-lo o afegint-hi dades addicionals.
Els termes de la llicència especifiquen les següents condicions:

- Reconeixement: cal donar crèdit a la font original, proporcionar un enllaç a la llicència i indicar si es van fer canvis.
- No comercial: no es pot utilitzar el material amb finalitats comercials.
- ShareAlike (Compartir igual): si es barregen, transformen o es fan servir les dades com a la base per una altra base de dades, s'ha de fer sota la mateixa llicència que l'original.
- No hi ha restriccions addicionals: no es poden aplicar termes legals ni mesures tecnològiques que restringeixin legalment altres persones a fer allò que la llicència permet.

Codi

El codi està publicat al següent projecte:
https://github.com/amurgo/TrailRunningRaces

Dataset

El data set ha estat publicat a Zenodo en el següent enllaç:
Trail Running Races 2020-2029 | Zenodo
DOI 10.5281/zenodo.4671018

[4] CreativeCommons - *Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)* [en línia] [data de consulta: 30 d'Abril de 2021]. Disponible a: https://creativecommons.org/licenses/by-nc-sa/4.0/

Albert Amurgo i Àlex Peláez

Contribucions

Contribucions	Signa
Recerca prèvia	Albert Amurgo i Àlex Peláez
Redacció de les respostes	Albert Amurgo i Àlex Peláez
Desenvolupament codi	Albert Amurgo i Àlex Peláez