A Note on the Files in the 01code Folder

The folder contains the python code used to filter and analyse the raw database, as well as the published R Markdown document containing the code used to produce the plots.

A Note on the Files in the 02Processed Folder

This folder contains the initial dataset as produced by the python code, and the dataset in long format, which is the format used for Plot 1. These were uploaded separately to the R Markdown document to save time converting the data from the initial format to the long format. After some tests this seemed to save 30 seconds or so in loading the document.

Converting the Caissabase file to utf-8

The Caissabase dataset seems to require being converted from ISO8859-1 (AKA Latin1) encoding to utf-8 to work with the python code that filters and analyses it.

This can be done in a linux terminal with the following code:

1.

```
file -i file.name
```

[checks the encoding - returns “8859_1”]

2.

```
iconv -f 8859_1 -t UTF-8//TRANSLIT file.name -o file.name
```

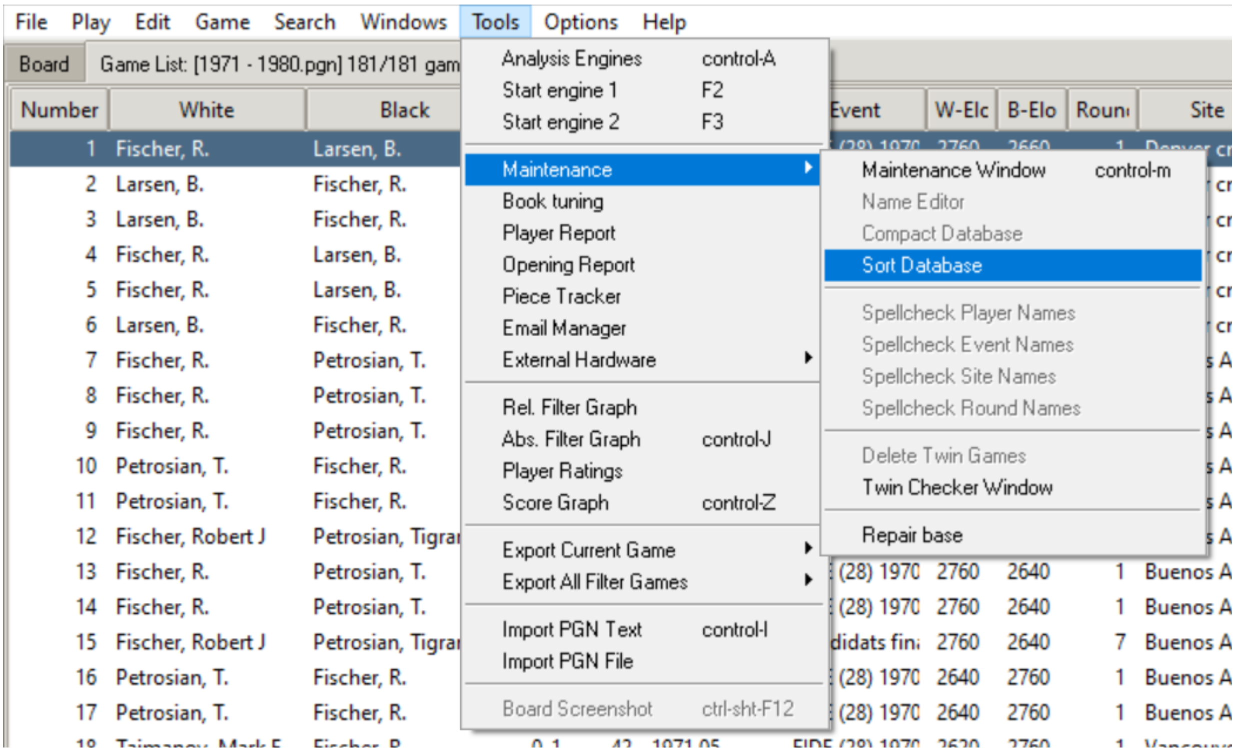
This can also be done inside python.

Alternative Sorting Methods

Another method to sort the full Caissabase data, which depending on the purpose may be faster than using the included python code, is to download the Scid, a programme for interfacing with pgn files: http://scidvspc.sourceforge.net/

Upload the .pgn file, and go to Game -> List All Games.

The games can then be filtered according to a range of information using the ‘Sort Database’ window. Found in Tools -> Maintenance -> Sort Database


File Play Edit Game Search Windows Tools Options Help
Board
Game List: [1971 - 1980.pgn] 181/181 gam
Number
White
Black
Event
W-Elo
B-Elo
Roun
Site
1 Fischer, R. Larsen, B.
2 Larsen, B. Fischer, R.
3 Larsen, B. Fischer, R.
4 Fischer, R. Larsen, B.
5 Fischer, R. Larsen, B.
6 Larsen, B. Fischer, R.
7 Fischer, R. Petrosian, T.
8 Fischer, R. Petrosian, T.
9 Fischer, R. Petrosian, T.
10 Petrosian, T. Fischer, R.
11 Petrosian, T. Fischer, R.
12 Fischer, Robert J Petrosian, Tigra
13 Fischer, R. Petrosian, T. (28) 1970 2760 2640 1 Buenos A
14 Fischer, R. Petrosian, T. (28) 1970 2760 2640 1 Buenos A
15 Fischer, Robert J Petrosian, Tigra didats fin 2760 2640 7 Buenos A
16 Petrosian, T. Fischer, R. (28) 1970 2640 2760 1 Buenos A
17 Petrosian, T. Fischer, R. (28) 1970 2640 2760 1 Buenos A
Analysis Engines control-A
Start engine 1 F2
Start engine 2 F3
Maintenance
Book tuning
Player Report
Opening Report
Piece Tracker
Email Manager
External Hardware
Rel. Filter Graph
Abs. Filter Graph control-J
Player Ratings
Score Graph control-Z
Export Current Game
Export All Filter Games
Import PGN Text control-I
Import PGN File
Board Screenshot ctrl-sht-F12
Maintenance Window control-m
Name Editor
Compact Database
Sort Database
Spellcheck Player Names
Spellcheck Event Names
Spellcheck Site Names
Spellcheck Round Names
Delete Twin Games
Twin Checker Window
Repair base

Scid allows you to remove multiple games at once, resort according to different criteria, and remove multiple games again. This makes it a great way to manually but quickly narrow down a large dataset according to multiple conditions.