Codebook

Raw .pgn file Data

The raw Caissabase data comes in .png (portable game notation) format.

Files in png format can be converted to a .txt file very easily by simply editing the filename to filename.txt.

The data consists of two components: headers and game notation.

Header information

Example

```
[Event "FIDE (28) 1970-1972"]
[Site "Denver cm sf"]
[Date "1971.07.??"]
[Round "1"]
[White "Fischer, R."]
[Black "Larsen, B."]
[Result "1-0"]
[WhiteElo "2760"]
[BlackElo "2660"]
[ECO "B88"]
```

Event	The name of the tournament or event
Site	Geographical location of the game
Date	This one is fairly self explanatory
Round	Round number, if the players were playing multiple games
White	The name of the player controlling the white pieces
Black	The name of the player controlling the black pieces
Result	“1-0”: White won the game “½ -½” : Game resulted in a draw “0-1”: Black won the game
WhiteElo	The Elo points of the player on the white pieces
BlackElo	The Elo points of the player on the black pieces.

ECO	The code indicating which opening was used. Find more here: https://www.365chess.com/eco.php

Algebraic notation

Example

```
1. e4 c5 2. Nf3 d6 3. d4 cxd4 4. Nxd4 Nf6 5. Nc3 Nc6 6. Bc4 e6 7. Bb3 Be7
8. Be3 O-O 9. f4 Bd7 10. O-O a6 11. f5 Qc8 12. fxe6 Bxe6 13. Nxe6 fxe6 14.
Na4 Rb8 15. Nb6 Qe8 16. Bxe6+ Kh8 17. Bf5 Ne5 18. Qd4 Qh5 19. Nd5 Nxd5 20.
Qxd5 Qe2 21. Ba7 Rbe8 22. Rf2 Qb5 23. c3 Bh4 24. g3 Qxd5 25. exd5 Bf6 26.
Raf1 Nc4 27. Be6 Ra8 28. Bd4 Bxd4 29. cxd4 Rxf2 30. Rxf2 b5 31. Kf1 g6 32.
b3 Na3 33. Ke2 Ra7 34. Rf8+ Kg7 35. Rd8 b4 36. Rxd6 Nb5 37. Rb6 Nxd4+ 38.
Kd3 Nxe6 39. Rxe6 a5 40. Kd4 Kf7 41. Re2 1-0
```

Algebraic notation is a way to describe the moves of a chess game.

Find more about algebraic chess notation here:
https://en.wikipedia.org/wiki/Algebraic_notation_(chess)

Processed Data

completeDataset.csv

The file of processed data contains the following headings:

game	This is simply the game number as indexed in the original file of raw date - this column is re-indexed to list the games in the order they appear in the processed data file after the data is transferred to R.
year	The year the game took place
w_elo	White Elo score (0 if not recorded)
b_elo	Black Elo score (0 if not recorded)
outcome	The outcome of the game; white win, draw, or black win
moves	The list of the moves of the game, in algebraic notation
total_elo	The sum of the two players' Elo scores.
inYearNumber	This number indicates the rank of the game in the year it appeared, as sorted by total_elo. For example, if the two players with the highest Elo score played a game,

	this game would be ranked 1.
topTen	These values are all “True” - indicating that the games were selected due to being in the top ranking set of games for that year by total_elo.
n_moves	The number of moves in the game
move_list_list	This, in retrospect very poorly named column, contains a list of game scores for each move. The game scores were derived from a Stockfish analysis of each move of the game, indicating the respective state of the game, varying between -1 and +1, where -1 is a decisive advantage for black, and +1 is a decisive advantage for white.
And then a series of numbered columns	These numbered columns contain the game scores for each move which are already contained in the move_list_list column, but with 1 column per move.

LongFormat.csv

This file contains the same data as the previous file, but in long rather than wide format.