Pasinato Alessio (887000)

DATA MANAGEMENT PROJECT

ABSTRACT

The purpose of this work is to acquire users' public data from Lichess and Chess.com using web scraping techniques, APIs and downloads, find which of them play on both websites, enrich these data with information about the players that are not available on these platforms and then store all the data on the document-based DBMS MongoDB for data exploration and insights.

DATA ACQUISITION: WEB SCRAPING

To obtain public data about players registered on Lichess and Chess.com, both websites's APIs required the username of the specific person whom public data were requested, but the APIs didn't have any endpoint which provided them. In order to overcome this problem web scraping was performed on the two websites, which allowed to obtain:

- 28.480 Lichess' usernames from «2021 Winter Marathon» Tournament[1]
- 2.600 usernames from the Lichess' leaderboards; only the top 200 players of each of the 13 game variants have been scraped because that was the maximum number of players that the website made visible
- 32.819 Chess.com usernames from «2022 Chess.com Daily Chess Championship»
- 30.000 usernames from the leaderboards of bullet, blitz and rapid game variants (10.000 usernames per variant)

It was possible though, that one player was present on more than one page where the scraping was performed (e.g. a top player who is in one leaderboard and who also played one of the tournaments); after the removal of the duplicates, the total number of the obtained Lichess usernames was 30.466 and the ones from Chess.com were 55.143.

[1] https://lichess.org/tournament/winter21

DATA ACQUISITION: DOWNLOAD

The results of "2021 Spring Marathon" Lichess tournament, from which the usernames of the 31.325 participants were extracted, were obtained by downloading them directly from Lichess[2]. The total number of Lichess' usernames became 56.383, after the removal of the duplicates.

Also, a dataset about FIDE rated players whose rating was above 1000, composed by 175.358 rows, was downloaded by Kaggle[3]. Each row represents certain characteristics of a player such as: his name, his title, his country and his rating in that country, his FIDE (Fédération Internationale des Échecs) rating, his age and his k-factor**.**

Down below the first five rows of this dataset are shown.

	Country Rank	Name	Title	Country	FIDE	Age	K-factor
0	1	Mirzaad, S.wahabuddin	FM	Afghanistan	1999.0	35.0	20.0
1	2	Rahmani, Asef	unranked/unrated	Afghanistan	1871.0	49.0	20.0
2	3	Sarwari, Hamidullah	unranked/unrated	Afghanistan	1866.0	33.0	20.0
3	4	Sakhawaty, Sepehr	unranked/unrated	Afghanistan	1846.0	19.0	20.0
4	5	Jamshedy, Mohammad Ismail	unranked/unrated	Afghanistan	1790.0	78.0	20.0

After the exploration of this dataset was discovered that data about Spanish players was missing: web scraping on the website[4] where this dataset was originally taken from was performed to obtain data about these players.

The complete dataset has 191.971 rows.

DATA ACQUISITION: API

After the collection of all the usernames, API queries were performed to collect all the users' public data. The total number of queries was the same of the number of usernames, that is 56.383, which lead to a file with a total weight of 66 MB.

Lichess' API[5] provided JSONs with the following shape:

[2] https://lichess.org/tournament/spring21
[3] https://www.kaggle.com/datasets/deepcontractor/international-chess-statistics-2022
[4] https://chess-rankings.com/
[5] https://lichess.org/api#operation/apiUser

{
"id": "georges",
"username": "Georges",
"online": true,
"createdAt": 1290415680000,
"disabled": false,
"tosViolation": false,
"seenAt": 1522636452014,
"patron": true,
"verified": true,
"title": "NM",
"url": "https://lichess.org/@/georges",
"playing":
"https://lichess.org/yqfLYJ5E/black",
"completionRate": 97,
"perfs": {…},
"profile": {…},
"playTime": {…},
"count": {…}
}

- id: player's ID
- username: player's username
- online: boolean variable indicating if the player at the time of the query was online
- createdAt: timestamp of the time when the player joined the platform
- disabled: boolean variable indicating if the player's account is banned
- tosViolation: boolean variable indicating if the player has violated the terms of service
- seenAt: timestamp related to the last time the player entered on the platform
- patron: boolean variable indicating if the player has ever made a donation to the platform
- verified: variable indicating if the player has a verified profile
- title: player's title
- url: URL redirecting to the player's Lichess profile
- playing: URL redirecting to the game the player was playing at the time of the query
- completionRate: percentage of the player's completed games

The keys "perfs" is composed in the following way:

"perfs": {
"chess960": {},
"atomic": {},
"racingKings": {},
"ultraBullet": {},
"blitz": {},
"kingOfTheHill": {},
"bullet": {},
"correspondence": {},
"horde": {},
"puzzle": {},
"classical": {},
"rapid": {},
"storm": {} }

Each of these sub-keys is a game variation.

The "storm" variation has the following structure:

"storm": {
"runs": 44,
"score": 61
}

- runs: number of "puzzle storm" played
- score: best score

All the other keys are composed in this way:

“game variant name”: {
"games": 2945,
"rating": 1609,
"rd": 60,
"prog": -22,
"prov": true
}

- games: number of total played games of that variant
- rating: actual player’s rating in that variant
- rd: rating deviation, which is the amount of confidence the system has in your rating
- prog: progress in rating in the last 12 games
- prov: variable indicating if the rating is provisional

The key “profile” is composed in this way:

“profile”: {
"country": "EC",
"location": "string",
"bio": "Free bugs!",
"firstName": "Thibault",
"lastName": "Duplessis",
"fideRating": 1500,
"uscfRating": 1500,
"ecfRating": 1500,
"links":
"github.com/ornicar\r\ntwitter.com/ornicar"
}

- country: Alpha-2 country code indicating the player’s country
- location: string representing players’ location
- bio: player Lichess’ biography
- firstName: player’s first name
- lastName: player’s last name
- fideRating: player’s FIDE rating
- uscfRating: player’s USCF rating
- ecfRating: player’s ECF rating
- links: URLs that the player included in his profile

The “playtime” key is composed in this way:

"playTime": {
"total": 3296897,
"tv": 12134
}

- total: seconds that the player’s spent playing
- tv: seconds that the player’s spent on Lichess TV

The “count” key is composed in this way:

"count": {
"all": 9265,
"rated": 7157,
"ai": 531,
"draw": 340,
"drawH": 331,
"loss": 4480,
"lossH": 4207,
"win": 4440,
"winH": 4378,
"bookmark": 71,
"playing": 6,
"import": 66,
“me”: 0
}

- total: total games played
- rated: total number of rated games

- ai: number of games played against Lichess' bots
- draw: number of drawn games, including the games against the bots
- drawnH: number of drawn games, excluding the games against the bots
- loss: number of lost games, including the games against the bots
- lossH: number of lost games, excluding the games against the bots
- win: number of won games, including the games against the bots
- winH: number of won games, excluding the games against the bots
- bookmark: number of bookmarked games
- playing: number of matches currently in progress
- import: number of PGN matches imported
- me: number of games played against me, the person who made this project.

Chess.com[6] didn't have a unique endpoint that provided a JSON containing both public data about a user and its statistics, but instead these data were split into two.

The statistics JSON's size was 55 MB and has this shape:

{

- "fide": integer,
- "chess_daily": {…},
- "chess960_daily": {…},
- "chess_rapid": {…},
- "chess_bullet": {…},
- "chess_blitz": {…},
- "tactics": {…},
- "lessons": {…},
- "puzzle_rush": {…}

}

"chess_daily" and "chess960_daily" keys are composed in this way:

```
{
 "last": {
  "date": timestamp,
  "rating": integer,
  "rd": integer
 },
 "best": {
  "date": timstamp,
  "rating": integer,
  "game": URL,
 },
 "record": {
  "win": integer,
  "loss": integer,
  "draw": integer,
  "time_per_move": integer,
  "timeout_percent": integer
 },
 "tournament": {
  "count": integer,
  "withdraw": integer,
  "points": integer,
  "highest_finish": integer
 } }
```

- date: timestamp of the last or best game, depending if the main key is "last" or "best"
- rating: rating of the last or best game, depending if the main key is "last" or "best"
- rd: rating deviation, which is the amount of confidence the system has in your rating
- win: number of games won
- loss: number of games loss

[6] https://www.chess.com/news/view/published-data-api#pubapi-general-response-codes

- draw: number of games drawn
- time_per_move: integer number of seconds per average move
- timeout_percent: timeout percentage in the last 90 days
- count: number of tournaments joined
- withdrawn: number of tournaments withdrawn
- points: total number of points earned in tournaments
- highest_finish: best tournament place

“chess_blitz”, “chess_bullet” and “chess_rapid” have the same structure described above, except for the lack of “time_per_move' and 'timeout_percent' in the “record” key.

The “lessons” and “tactics” keys are composed in this way:

```
"highest": {
   "rating": "integer",
   "date": "timestamp" },
"lowest": {
   "rating": "integer",
   "date": "timestamp" }
```

- rating: highest or lowest rating, depending if the main key is “highest” or “lowest”
- date: date when the rating happened to be the highest or lowest, depending if the main key is “highest” or “lowest”

The public data JSON’s size is 2 MB, which is significantly smaller than the statistics file because only the players who included their FIDE rating have been kept; this parameter will be important later for matching records from the different datasets. The JSON has this shape:

```
{"@id": URL,
"url": URL,
"username": string,
"player_id": integer,
"title": string,
"status": string,
"name": string,
"avatar": URL,
"location": string,
"country": string,
"joined": timestamp,
"last_online": timestamp,
"followers": integer,
"is_streamer": Boolean,
"twitch_URL": URL }
```

- @id: the location of the player’s profile (always self-referencing)
- url: the chess.com user's profile page username: the player’s username
- player_id: the non-changing Chess.com player’s ID
- title: player’s title
- status: variable that indicates whether the player has a premium account, a basic one, if he’s been banned or if he’s a staff member
- name: the player’s first and last name
- avatar: URL of a 200x200 image
- location: player’s city or location
- country: API location of this player's country's profile
- joined: timestamp of registration on Chess.com
- last_online: timestamp of the most recent login
- followers: number of players tracking this player's activity
- is_streamer: boolean variable indicating if the player’s a streamer
- twitch_URL: Twitch.tv URL

NAME MATCHING

To figure out whether a user played on both Lichess and Chess.com it was not possible to directly look for matching names between the two platforms; the entries are not verified, hence the data may not me semantically accurate. To be more accurate, FIDE's players dataset had been used as a reference.
In order to understand if a player from one of the two platform was also present in the FIDE's players dataset, three keys have been utilized as a proxy: the name of the player, his FIDE rating and his title. These keys have been used since they're the most representative information about a player that we have in our data. The title and the rating weren't always both present for each player, but it was sufficient that only one was, then, jointly with the name, it was possible to look for a match.

After the data exploration process it was found that 3.885 Lichess players had in their profile the information about their FIDE rating, and 426 carried a title. 3882 of the rated players also had the information about their name, but, by using the "alphabet_detector" library[7], it was discovered that 363 of them were written in Cyrillic; with the help of another library, "transliterate"[8], those names have been transliterated.
It was also found that Lichess had a title created only for its platform, which is LM (Lichess Master); it was removed because of its unofficial nature from the only player who held it, and his new title became "unranked/unrated", which is the same nomenclature present in the FIDE dataset.

As regards Chess.com players, 5.394 of them had a FIDE rating, 4231 people of this group also provided their name, but 201 of them were written in Cyrillic. The same procedure applied to Lichess players had been repeated here.

The total number of names to be searched was the sum of the players that decided to include their real name on their profile and their FIDE rating or title, which was 8113; each of these names needed to be compared with every existing name in the FIDE dataset (around 190.000). This was a very computational demanding operation, which needed to be optimised.

The optimisation consisted in the creation, for each of the 8113 comparisons, of a sample of the original FIDE dataset, which was composed only by players who had:

- the same title as the queried user. For the untitled ones their title was "unraked/untitled"
- the same FIDE rating of the queried user, with a variation interval of ± 5%; this percentage was chosen after a test that showed that the rating of the first 50 matched users was always inside this variation range.

After the creation of the sample the user's name was compared with all the names in the new subset.

[7] https://github.com/EliFinkelshteyn/alphabet-detector
[8] https://pypi.org/project/transliterate/

The library used for matching strings was "FuzzyWuzzy"[9], especially its "token_sort_ratio()" function which was not influenced by the order of the strings and which returned a level of similarity between them in the 0-100 interval. The threshold was set to 90, so that when the name of a user had a higher value of similarity with a name from the other dataset, the names were considered the same. If more than one string happen to have more than 90% of similarity with another one, the indexes of the rows of the FIDE dataset containing those names and the user's name were saved for a manual check. The amount of user's names that matched with more than one name in the dataset was 11; those names had been manually processed and for 7 of them was possible to find a match.

The total number of matches found was 891 for Lichess, over a total of 3885 users, corresponding to 23% of total; the matches were higher for Chess.com, at 32%, which corresponds to 1351 correspondences over 4228 users.

DATA CLEANING AND INTEGRATION

This step of data integration consists of integrate information from the FIDE dataset with the Lichess and Chess.com's JSONs and also between JSONs themselves when a player had an account on both platforms.

Before integrating the documents from the two platforms though, it was necessary to:

- Rename the keys of the two different platform's JSONs in order for them to match
- Unify for each key the set of possible values that they can hold
- Create a unique JSON's structure: it was decided to adapt the Chess.com JSON's structure to the Lichess one

Regarding the first step, the Lichess' keys that have been renamed are: "createdAt", "seenAt" and "patron" who became respectively "joined", "last_online" and "status". This nomenclature is the same utilised by Chess.com.
The keys from Chess.com that have been modified are "chess_rapid", "chess_bullet" and "chess_blitz" who simply became "rapid", "bullet" and "blitz" like on Lichess, and "fide" who became "fideRating".

Subsequently, all the values of the keys that contained timestamps had been converted into datetimes, and the possible values of "status" became "patron" and "not patron", differently from before, when the key was named "patron" its values were the booleans "True" and "False".

[9] https://pypi.org/project/fuzzywuzzy/

In the matter of the adaptation of the Chess.com JSON's structure to the Lichess' one, a few steps had to be made. The first one was the creation of the "profile" key, since in the public data JSON it's not present, differently from Lichess. The information included in this key were: "name", "location", "country" and "fideRating", like on the other platform, but also "avatar", "followers", "is_streamer" and "twitch_URL" had been added, with the consideration that these were also personal information that should be put into the profile key.

Regarding the statistics' JSON, the "blitz", "rapid" and "bullet" keys, which are the same variants played on both websites, had this structure:

```
name of the variant: {
        "last": {"rating" : integer,
                 "date" : timestamp,
                 "rd" : integer},
        "best": {"rating" : integer,
                 "date: timestamp,
                 "game" : URL }
        "record": {"win" : integer,
                  "loss" : integer,
                  "draw" : integer }.
```

To unnest the structure and imitate the Lichess' one "last", "best" and "record" have been removed. The information contained in "last" became respectively "rating", "last_game_date" and "rd". The best rating simply became "best", the best date became "best_date" and the best game became "best_game". Finally, the data contained in "record" got simply unnested and kept its name.
The new structure has this shape:

```
name of the variant: {
        "rating" : integer,
        "rd" : integer,
        "last_game_date": datetime,
        "best" : integer,
        "best_date": datetime,
        "best_game": URL,
        "games" : integer,
        "win" : integer,
        "loss" : integer,
        "draw" : integer}.
```

The other game variants kept their original shape, considering that they're not played on both websites.

Furthermore, a new key “count” had been created, which contained the number of total wins, losses, draws, and the sum of these three; the corresponding names are “win”, “loss”, “draw” and “all”.

Finally, the documents relative to public data and statistics have been merged, with this result:


```
{
"profile": {…},
"@id": URL,
"url": URL,
"username": string,
"player_id": integer,
"title": string,
"status": string,
"joined": datetime,
"last_online": datetime,
"count": {
      "win": integer,
      "loss" : integer,
      "draw": integer,
      "all" : integer},
"perfs": {
         "chess_daily": {
                "rating" : integer,
                "rd" : integer,
                "last_game_date": datetime,
                "best" : integer,
                "best_date": datetime,
                "best_game": URL,
                "games" : integer,
                "win" : integer,
                "loss" : integer,
                "draw" : integer }
      "chess960_daily": {…},
      "rapid": {…},
      "bullet": {…},
      "blitz": {…},
      "fideRating": integer,
      "tactics": {…},
      "lessons": {…},
      "puzzle_rush": {…}}
```


The final step was to create a document which, given the full real name of a player as the primary key, contained the information from the FIDE dataset and the ones from one or both platforms, depending on where the user was active. The document contained on the first level the name, title, FIDE rating, federation, age and k-factor of the given player; two other keys, “lichess” and “chessCom” were present, which in turn contained the user’s JSON with all the information on that given platform. Its structure was the following:


```
{
"name": string,
"title": string,
"federation": string,
"age": integer,
"fide": integer,
"k-factor": integer,
"lichess": {…},
"chessCom": {…}}
```

DATA MODELING

The data obtained from the two platforms is not structured, they provide different public information about players and they also have different realities to represent, such as different game variations which are not playable on both websites. That's the reason why MongoDB, a document-based database, was chosen as the DBMS where the data was stored.

Three collections named "Lichess", "ChessCom" and "Corrispondenze" were created on MongoDB, which respectively contained all the players whose data was integrated with the information in the FIDE dataset, that only played on Lichess, Chess.com, and all the players that played on both platforms.

Subsequently, this DBMS had been used for data exploration of all the data.

DATA EXPLORATION

Regarding Lichess, out of a total of 891 players, 74% of people are not ranked in contrast to the 36% on Chess.com; this difference may be caused by the source of the analysed data, since around 23.000 Chess.com usernames were drawn from the top players leaderboards.
As concerns joining the platform, 79 new people (within our analysed data) joined on March 2020. The most people who joined Lichess on another month was 22, in February 2018. This increase in rate may be due to the pandemic which begun to spread exactly in that month.
The federations with the most players are Germany (9% of the total), Russia (8%) and India (7%), followed by Sweden (3%).
76% of players are included in the 10-40 years old range and the 34% are between 10 and 20 years old.

Talking about Chess.com, across all its analysed players, the most followed are Alexandra Botez and Simon K. Williams, each one having approximately 55k followers, 30k more than the third most followed person, who is Krikor Mekhitarian.
Out of a total of 1351 players, 1080 people (80%) have a premium account, and this really differs from Lichess. The reason may be that Lichess only offers a change in the patron's logo, while Chess.com gives more benefits to the premium players, such as an AI tutor who helps understanding mistakes and in-depth analysis of past games.
As concerns joining the platform, 42 was the number of new members in April 2020, probably due to the pandemic. The most people who joined Chess.com on another month was 22, the half, in May 2016. The reason may be the same as explained in the Lichess' case.
The top 5 federations by number of players are Russia (9%), US (7%), India (6%), Poland (5%) and Germany (4%).

As concerns the users who play on both websites, the statistics about the age, the joining date, the federation and the possession of a premium account are similar as the ones described above.

DATA QUALITY: CURRENCY

As concerns to currency, the API queries have been made on different days, hence not all the information about the players are up to the same date. Down below is shown the list of the data's currency.

Lichess' players from:

- Winter Tournament: 22/02/2022
- Leaderboards: 24/02/2022
- Spring Tournament: 26/02/2022

Chess.com players from:

- 2022 Chess.com Daily Chess Championship: 22/02/2022
- Leaderboards: 24/02/2022

DATA QUALITY: COMPLETENESS

Lichess and Chess.com provide both common information about a player and also data that is not mutually available on both websites. Chess.com provides information whether a player is a streamer, his Twitch URL and the number of followers he/she has on the platform, whereas Lichess provides the total time spent playing and the percentage of games completed.

As regards the game variations Chess.com has Chess Daily, Chess 960 Daily and puzzle rush, whilst Lichess has Chess960, UltraBullet, King of the Hill, Correspondence, Horde, Storm, Antichess, CrazyHouse, Three Check, Streak and Racer.

For both platforms was searched the percentage of presence of the keys contained in "perfs" and "profile"; the other data was not looked for because it was always given by the API or it was always present in the FIDE dataset.

LICHESS

Perfs

'chess960': 75%,
'puzzle': 96%,
'atomic': 51%,
'ultraBullet': 69%,
'blitz': 100%,
'crazyhouse': 60%,
'bullet': 100%,
'correspondence': 100%,
'classical': 100%,
'rapid': 100%,
'storm': 73%,
'racer': 61%,
'kingOfTheHill': 45%,
'threeCheck': 52%,
'horde': 49%,
'antichess': 50%,
'racingKings': 37%,
'streak': 54%

Profile

'country': 98%,
'location': 77%,
'bio': 60%,
'firstName': 100%,
'lastName': 100%,
'uscfRating': 5%,
'ecfRating': 3%,
'rcfRating': 1%,
'dsbRating': 2%,
'links': 34%

Chess.com

Perfs

'chess_daily': 56%,
'chess960_daily': 21%,
'rapid': 90%,
'bullet': 94%,
'blitz': 99%,
'tactics': 100%,
'lessons': 100%,
'puzzle_rush': 100%

Profile

'avatar': 85%,
'followers': 100%,
'country': 100%,
'location': 73%,
'is_streamer': 100%,
'twitch_url': 3%

SITOGRAPHY

[1] *Lichess' 2021 Winter Marathon*: https://lichess.org/tournament/winter21
[2] *Lichess' 2021 Spring Marathon*: https://lichess.org/tournament/spring21
[3] *International chess statistics 2022 dataset*:
https://www.kaggle.com/datasets/deepcontractor/international-chess-statistics-2022
[4] *Live chess ratings and rankings*: https://chess-rankings.com/
[5] *Lichess' API*: https://lichess.org/api#operation/apiUser
[6] *Chess.com's API*: https://www.chess.com/news/view/published-data-api#pubapi-general-response-codes
[7] *Alphabet detector library*: https://github.com/EliFinkelshteyn/alphabet-detector
[8] *Transliterate library*: https://pypi.org/project/transliterate/
[9] *FuzzyWuzzy library*: https://pypi.org/project/fuzzywuzzy/