02-rzepinskip-book_ids_types_analysis

January 19, 2019

1 Important

`make data data/raw/books_xml` has to be run before any cell in this notebook

2 Imports

```
In [1]: import os
        import pandas

In [2]: books_xml_dir = "../data/raw/books_xml"
```

2.1 IDs

- `work_id` - globally unique id of a book(abstract, disregarding edition or language)
- `goodreads_book_id`, `isbn`, `isbn13` - id of a specific edition of the book
- `best_book_id` - id of most popular edition of the book

`book_id` is used through data files as a new abstract identifier for a book: - in range 1-10000 - semantically identical to `work_id`

It is used in ratings.csv and to_read.csv, which were aggregated by `work_id`, so they contain data for all editions of a book.

2.2 book.csv

```
In [3]: book_df = pandas.read_csv("../data/raw/book.csv")

In [4]: book_df.head(1)
```

```
Out[4]:    book_id  goodreads_book_id  best_book_id  work_id  books_count       isbn  \
        0        1            2767052       2767052  2792775          272  439023483

                isbn13          authors  original_publication_year    original_title  \
        0  9.780439e+12  Suzanne Collins                     2008.0  The Hunger Games

                                  ...                         ratings_count  \
        0                         ...                               4780653

          work_ratings_count  work_text_reviews_count  ratings_1  ratings_2  \
        0            4942365                   155254      66715     127936
```

```
   ratings_3  ratings_4  ratings_5  \
0     560092    1481305    2706317

                                                image_url  \
0  https://images.gr-assets.com/books/1447303603m...

                                           small_image_url
0  https://images.gr-assets.com/books/1447303603s...

[1 rows x 23 columns]
```

To check whether the dataset contains multiple editions of the same book, we should look for duplicates in columns `work_id` or `best_book_id`.

```
In [5]: len(book_df[book_df.duplicated(['work_id'])])

Out[5]: 0

In [6]: len(book_df[book_df.duplicated(['best_book_id'])])

Out[6]: 0
```

In the dataset there is no duplicate `work_id`, so `book_id` has the same meaning as `work_id`

```
In [7]: import booksuggest.data.clean_book

In [8]: book_extra_info_rows =
        booksuggest.data.clean_book.extract_book_extra_info(books_xml_dir)
        book_extra_info_df =
        booksuggest.data.clean_book.process_book_extra_info(book_extra_info_rows)

In [9]: book_extra_info_df.head(1)

Out[9]:    work_id         isbn13                                        description
        0   517040  9780517222850   The Shell Seekers is a novel of connection: o...

In [10]: set(book_df.work_id.unique()) ^ set(book_extra_info_df.work_id.unique())

Out[10]: set()
```

2.3 similar_books.csv

```
In [11]: import booksuggest.data.prepare_similar_books

In [12]: similar_books_rows =
         booksuggest.data.prepare_similar_books.extract_similar_books(books_xml_dir)
         similar_books_raw_df =
         booksuggest.data.prepare_similar_books.process_similar_books(similar_books_rows)

In [13]: similar_books_raw_df.head(1)

Out[13]:    work_id  similar_book_work_id
         0   517040               1004459
```

Here, data rows are identified by `work_id`. To maintain consistency we should change ids to `book_id`.

```
In [14]: len(set(book_df.work_id.unique()) &
         set(similar_books_raw_df.similar_book_work_id.unique()))
```

```
Out[14]: 6025
```

```
In [15]: len(set(similar_books_raw_df.similar_book_work_id.unique()) -
         set(book_df.work_id.unique()))
```

```
Out[15]: 50644
```

Section `similar_books` contains 6025 books from the dataset. Aditionally, more than 40k books are out of the dataset and provide no value to the analysis, so they should be omitted.

2.4 book_tags.csv

```
In [16]: book_tags_df = pandas.read_csv("../data/raw/book_tags.csv")
```

```
In [17]: book_tags_df.head(1)
```

```
Out[17]:    goodreads_book_id  tag_id   count
         0                  1   30574  167697
```

Here, data rows are identified by `goodreads_book_id`. To maintain consistency we should change ids to `book_id`.

```
In [18]: set(book_tags_df.goodreads_book_id.unique()) ^ set(book_df.goodreads_book_id.unique())
```

```
Out[18]: set()
```

There 14 books that were not marked as `to_read` by any user.

2.5 ratings.csv

```
In [19]: ratings_df = pandas.read_csv("../data/raw/ratings.csv")
```

```
In [20]: ratings_df.head(1)
```

```
Out[20]:    user_id  book_id  rating
         0        1      258       5
```

```
In [21]: set(ratings_df.book_id.unique()) ^ set(book_df.book_id.unique())
```

```
Out[21]: set()
```

2.6 to_read.csv

```
In [22]: to_read_df = pandas.read_csv("../data/raw/to_read.csv")
```

```
In [23]: to_read_df.head(1)
```

```
Out[23]:    user_id  book_id
         0        9        8
```

```
In [24]: set(to_read_df.book_id.unique()) ^ set(book_df.book_id.unique())
```

```
Out[24]: {3151,
          3539,
          3996,
          4206,
          4439,
          5130,
          5898,
          6262,
          7330,
          7803,
          8055,
          9120,
          9161,
          9426}
```

```
In [25]: (set(book_df.book_id.unique()) - set(to_read_df.book_id.unique())) ==
         set(to_read_df.book_id.unique()) ^ set(book_df.book_id.unique())
```

```
Out[25]: True
```