03-rzepinskip-book_tags_cleaning

January 19, 2019

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

```
In [2]: book_df = pandas.read_csv("../data/raw/book.csv")
        tags_df = pandas.read_csv("../data/raw/tags.csv")
        book_tags_df = pandas.read_csv("../data/raw/book_tags.csv")
        with open("../data/external/genres.txt") as file:
            goodreads_genres = [line.rstrip('\n') for line in file]
```

0.1 Switch to book_id

```
In [3]: book_ids_df = book_df[['book_id', 'goodreads_book_id']]
        book_tags_fixed_ids_df = book_tags_df.merge(book_ids_df,
        on='goodreads_book_id')[['tag_id', 'book_id', 'count']]
```

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

```
Out[4]:    tag_id  book_id   count
        0   30574       27  167697
```

0.2 Filter non-genres tags

```
In [5]: tags_filtered_df = tags_df[tags_df.tag_name.isin(goodreads_genres)]
```

```
In [6]: tags_filtered_df.size, tags_df.size
```

```
Out[6]: (1664, 68504)
```

```
In [7]: tags_filtered_df.head(30)
```

```
Out[7]:       tag_id       tag_name
        283      283   10th-century
        297      297   11th-century
        307      307   12th-century
        317      317   13th-century
        327      327   14th-century
        341      341   15th-century
        350      350   16th-century
        362      362   17th-century
        392      392   18th-century
        509      509   19th-century
        543      543      1st-grade
        923      923   20th-century
```

941	941	21st-century
969	969	2nd-grade
1094	1094	40k
1416	1416	abandoned
1499	1499	abuse
1510	1510	academia
1511	1511	academic
1523	1523	accounting
1540	1540	action
1559	1559	activism
1586	1586	adaptations
1619	1619	adolescence
1629	1629	adoption
1642	1642	adult
1659	1659	adult-fiction
1691	1691	adventure
1710	1710	adventurers
1746	1746	africa

0.3 Add tag names to book_tags.csv

```
In [8]: book_tags_joined_df = book_tags_fixed_ids_df.merge(tags_filtered_df,
        on='tag_id')[['book_id', 'tag_name', 'count']]
```

```
In [9]: book_tags_joined_df.head(10)
```

Out[9]:

	book_id	tag_name	count
0	27	fantasy	37174
1	21	fantasy	3441
2	2	fantasy	47478
3	18	fantasy	39330
4	24	fantasy	38378
5	3275	fantasy	104
6	3753	fantasy	253
7	54	fantasy	3428
8	337	fantasy	933
9	964	fantasy	105

```
In [10]: book_tags_joined_df.sort_values(['book_id', 'count'], ascending=[True, False])
```

Out[10]:

	book_id	tag_name	count
6233	1	young-adult	25968
13501	1	fiction	13819
166503	1	dystopia	11065
2661	1	fantasy	10836
70137	1	science-fiction	8772
47716	1	romance	3341
22584	1	adventure	3190
36050	1	teen	1776
167057	1	post-apocalyptic	1461

58901	1	action	1263
126894	1	survival	1120
30700	1	novels	904
67672	1	thriller	800
195907	1	futuristic	701
162372	1	suspense	641
62379	1	love	484
55891	1	coming-of-age	447
52647	1	contemporary	382
91395	1	speculative-fiction	364
146172	1	drama	340
2	2	fantasy	47478
4261	2	young-adult	14984
7891	2	fiction	13239
16988	2	magic	4302
18885	2	childrens	3828
20239	2	adventure	2430
23900	2	classics	1898
32350	2	middle-grade	1558
26686	2	novels	1082
33264	2	paranormal	923
...	...	...	...
126600	9999	research	8
172287	9999	society	8
175311	9999	social-justice	8
103787	9999	history	7
110906	9999	essays	7
143487	9999	politics	7
100059	9999	unfinished	6
121426	9999	social-science	6
153391	9999	sexuality	6
176837	9999	chick-lit	6
102315	10000	history	776
104445	10000	non-fiction	219
184789	10000	military-history	114
124116	10000	war	104
200968	10000	world-war-i	88
186569	10000	european-history	34
111232	10000	world-history	29
142948	10000	politics	17
106793	10000	historical	16
121053	10000	germany	10
137332	10000	american-history	10
91975	10000	20th-century	9
120792	10000	france	9
94038	10000	abandoned	8
109153	10000	reference	8
175520	10000	russia	5

```
98567     10000                unfinished      4
115960    10000                 biography      3
132524    10000       historical-fiction      3
189894    10000  international-relations      3

[207127 rows x 3 columns]
```

In [ ]: