01-rzepinskip-szymanskir-xml-file-content-analysis

January 19, 2019

1 XML parsing

1.1 Data sections

1.1.1 Ids

- `book.id`, `book.isbn`, `book.isbn13` - id of a specific edition of the book
- `book.work.id` - globally unique id of a book(abstract, disregarding edition or language)
- `book.work.best_book_id` - id of most popular edition of the book
- https://www.goodreads.com/book/show/book.id
- https://www.goodreads.com/work/editions/book.work.id

```xml
<GoodreadsResponse>
   <book>
      <id>1</id>
      <title><![CDATA[Harry Potter and the Half-Blood Prince (Harry Potter, #6)]]></title>
      <isbn><![CDATA[0439785960]]></isbn>
      <isbn13><![CDATA[9780439785969]]></isbn13>
      <work>
         <id type="integer">41335427</id>
         <best_book_id type="integer">1</best_book_id>
      </work>
   </book>
</GoodreadsResponse>
```

1.1.2 Authors

Not parsed - already in csv

1.1.3 Ratings

Not parsed - already in csv

1.1.4 Shelves

Not parsed - already in csv

1.1.5 Simple columns

Some simple columns(simple type value) are not present in `books.csv` or are wrongly formatted. Thus, we want to extract ths additional data from XML files.

- `description`
- `isbn13`

```
<?xml version="1.0"?>
<GoodreadsResponse>
   <book>
      <isbn13><![CDATA[9780439785969]]></isbn13>
      <description><![CDATA[The war against Voldemort is not going well: even Muggle governmen
   </book>
</GoodreadsResponse>
```

1.1.6 Similar books

Similar books as proposed by Goodreads. The methodology of theirs recommendation is not yet known.

Similar items for specific book should be saved as pairs `(book_id, similar_book_id)` to maintain consistency with other data files.

```
<?xml version="1.0"?>
<GoodreadsResponse>
   <book>
      <similar_books>
         <book>
            <id>153795</id>
            <title><![CDATA[Squire (Protector of the Small, #3)]]></title>
            <title_without_series>Squire</title_without_series>
            <link><![CDATA[https://www.goodreads.com/book/show/153795.Squire]]></link>
            <small_image_url><![CDATA[https://s.gr-assets.com/assets/nophoto/book/50x75-a91bf2
            <image_url><![CDATA[https://s.gr-assets.com/assets/nophoto/book/111x148-bcc042a9c9
            <num_pages>400</num_pages>
            <work>
               <id>1142883</id>
            </work>
            <isbn />
            <isbn13 />
            <average_rating>4.27</average_rating>
            <ratings_count>39948</ratings_count>
            <publication_year>2007</publication_year>
            <publication_month>12</publication_month>
            <publication_day>18</publication_day>
            <authors>
               <author>
                  <id>8596</id>
                  <name>Tamora Pierce</name>
                  <link><![CDATA[https://www.goodreads.com/author/show/8596.Tamora_Pierce]]></
               </author>
            </authors>
         </book>
         <!-- More data of the same type -->
      </similar_books>
   </book>
</GoodreadsResponse>
```

1.2 Sample xml file content:

```
<?xml version="1.0"?>
<GoodreadsResponse>
   <Request>
      <authentication>true</authentication>
      <key><![CDATA[all_men_must_serve]]></key>
      <method><![CDATA[book_show]]></method>
   </Request>
   <book>
      <id>1</id>
      <title><![CDATA[Harry Potter and the Half-Blood Prince (Harry Potter, #6)]]></title>
      <isbn><![CDATA[0439785960]]></isbn>
      <isbn13><![CDATA[9780439785969]]></isbn13>
      <asin />
      <kindle_asin><![CDATA[B019PIOJZE]]></kindle_asin>
      <marketplace_id><![CDATA[A1F83G8C2ARO7P]]></marketplace_id>
      <country_code><![CDATA[GB]]></country_code>
      <image_url>https://images.gr-assets.com/books/1361039191m/1.jpg</image_url>
      <small_image_url>https://images.gr-assets.com/books/1361039191s/1.jpg</small_image_url>
      <publication_year>2006</publication_year>
      <publication_month>9</publication_month>
      <publication_day>16</publication_day>
      <publisher>Scholastic Inc.</publisher>
      <language_code>eng</language_code>
      <is_ebook>false</is_ebook>
      <description><![CDATA[The war against Voldemort is [...]]]></description>
      <work>
         <id type="integer">41335427</id>
         <books_count type="integer">275</books_count>
         <best_book_id type="integer">1</best_book_id>
         <reviews_count type="integer">2234433</reviews_count>
         <ratings_sum type="integer">8109581</ratings_sum>
         <ratings_count type="integer">1787108</ratings_count>
         <text_reviews_count type="integer">27548</text_reviews_count>
         <original_publication_year type="integer">2005</original_publication_year>
         <original_publication_month type="integer">7</original_publication_month>
         <original_publication_day type="integer">16</original_publication_day>
         <original_title>Harry Potter and the Half-Blood Prince</original_title>
         <original_language_id type="integer" nil="true" />
         <media_type>book</media_type>
         <rating_dist>5:1162549|4:459309|3:136413|2:21524|1:7313|total:1787108</rating_dist>
         <desc_user_id type="integer">5119944</desc_user_id>
         <default_chaptering_book_id type="integer">1</default_chaptering_book_id>
         <default_description_language_code nil="true" />
      </work>
      <average_rating>4.54</average_rating>
      <num_pages><![CDATA[652]]></num_pages>
      <format><![CDATA[Paperback]]></format>
      <edition_information />
      <ratings_count><![CDATA[1680064]]></ratings_count>
      <text_reviews_count><![CDATA[22133]]></text_reviews_count>
      <url><![CDATA[https://www.goodreads.com/book/show/1.Harry_Potter_and_the_Half_Blood_Princ
      <link><![CDATA[https://www.goodreads.com/book/show/1.Harry_Potter_and_the_Half_Blood_Prin
      <authors>
         <author>
            <id>1077326</id>
            <name>J.K. Rowling</name>
            <role />
            <image_url nophoto="false"><![CDATA[https://images.gr-assets.com/authors/141594517
            <small_image_url nophoto="false"><![CDATA[https://images.gr-assets.com/authors/1415
            <link><![CDATA[https://www.goodreads.com/author/show/1077326.J_K_Rowling]]></link>
            <average_rating>4.43</average_rating>
            <ratings_count>18039589</ratings_count>
            <text_reviews_count>434729</text_reviews_count>
         </author>
         <!-- More data of the same type -->
      </authors>
      <reviews_widget>
      </reviews_widget>
      <popular_shelves>
         <shelf name="to-read" count="167697" />
         <shelf name="fantasy" count="37174" />
         <!-- More data of the same type -->
      </popular_shelves>
      <book_links>
         <book_link>
            <id>8</id>
            <name>Libraries</name>
            <link>https://www.goodreads.com/book_link/follow/8</link>
         </book_link>
      </book_links>
      <buy_links>
         <buy_link>
            <id>1</id>
            <name>Amazon</name>
            <link>https://www.goodreads.com/book_link/follow/1</link>
         </buy_link>
         <!-- More data of the same type -->
      </buy_links>
      <series_works>
         <series_work>
            <id>624922</id>
            <user_position>6</user_position>
            <series>
               <id>45175</id>
               <title><![CDATA[Harry Potter]]></title>
               <description><![CDATA[Orphan Harry learns he is a wizard on his 11th birthday wh
               <note><![CDATA[Cursed Child is NOT a Primary Work. Boxsets ARE part of the serie
               <series_works_count>13</series_works_count>
               <primary_work_count>8</primary_work_count>
               <numbered>true</numbered>
            </series>
         </series_work>
         <!-- More data of the same type -->
      </series_works>
      <similar_books>
         <book>
            <id>153795</id>
            <title><![CDATA[Squire (Protector of the Small, #3)]]></title>
            <title_without_series>Squire</title_without_series>
            <link><![CDATA[https://www.goodreads.com/book/show/153795.Squire]]></link>
            <small_image_url><![CDATA[https://s.gr-assets.com/assets/nophoto/book/50x75-a91bf24
            <image_url><![CDATA[https://s.gr-assets.com/assets/nophoto/book/111x148-bcc042a9c9
            <num_pages>400</num_pages>
            <work>
               <id>1142883</id>
            </work>
            <isbn />
            <isbn13 />
            <average_rating>4.27</average_rating>
            <ratings_count>39948</ratings_count>
            <publication_year>2007</publication_year>
            <publication_month>12</publication_month>
            <publication_day>18</publication_day>
            <authors>
               <author>
                  <id>8596</id>
                  <name>Tamora Pierce</name>
                  <link><![CDATA[https://www.goodreads.com/author/show/8596.Tamora_Pierce]]></l
               </author>
            </authors>
         </book>
         <!-- More data of the same type -->
      </similar_books>
```

```
   </book>
</GoodreadsResponse>
```