



مقدمه



در این تمرین شما به تحلیل داده‌هایی که از وبگاه¹ Goodreads استخراج شده است می‌پردازید. در این پرونده² به بیان ویژگی‌های برخی از موجودیت³‌هایی که در این تمرین به آن‌ها احتیاج است پرداخته می‌شود.

مفاهیم اولیه

Goodreads یک شبکه اجتماعی کتاب است، که کاربران در آن کتاب‌هایی را که خوانده‌اند پیدا می‌کنند، به کتاب‌ها امتیاز می‌دهند و کتاب‌های محبوب خود را به دوستان خود توصیه می‌کنند.

کتاب



هر کتاب صفحه‌ای ویژه دارد که در آن مشخصات کتاب درج شده است و کاربران با ورود به آن می‌توانند اطلاعات زیادی از کتاب و نویسنده آن کسب کنند. گاهی مختصری از کتاب و یا شرحی بر کتاب در این صفحه گنجانده می‌شود و از همان صفحه می‌توان به دیگر آثار نوشته شده توسط نویسنده کتاب دسترسی داشت.

می‌توان به کتاب امتیاز داد یا گفت که برای آینده آن را خواهیم خواند. بیوگرافی نویسنده کتاب و ژانر⁴ ادبی و هنری کتاب نیز قابل مشاهده است. در نهایت می‌توان با دیگران در مورد آن کتاب بحث کرد و از همان‌جا نیز کتاب را سفارش داده و وارد وبگاه فروشنده‌ی کتاب شده و آن را به شکل نسخه‌ی کاغذی و یا نسخه الکترونیک خریداری کرد.

¹ Website

² File

³ Entity

⁴ Genre

کاربران

هر کاربر در این وبگاه قفسه⁵ شخصی کتاب دارد و می‌تواند به آن کتاب بیفزاید، نگهداری کند و به دیگران پیشنهاد کند. کاربران در هر سال در رای‌گیری شرکت می‌کنند تا کتاب منتخب سال را انتخاب کنند. در آن می‌توان دوست‌یابی کرد و دوستان شما با امتیاز دادن به کتاب‌ها و توصیه‌ی آن‌ها به شما در برگزیدن کتاب خوب کمک‌تان می‌کنند.



می‌توان بیوگرافی نویسنده‌ها را خواند و کتاب‌های آنها را بررسی کرد و فهمید دیگر خوانندگان در مورد آن کتاب چه گفته‌اند. کاربران می‌توانند در مورد کتاب‌ها **نظراتشان** را بنویسند و یا به کتاب‌ها **امتیاز** بدهند.

قفسه کتاب

در این وبگاه سه قفسه اصلی وجود دارند:

- کتاب را پیش‌تر خوانده‌اند
- در حال مطالعه‌اند
- قرار است بعداً کتاب را بخوانند



هر کاربر به‌صورت پیش‌فرض این سه قفسه کتاب را دارد و می‌تواند قفسه‌ای جدید بوجود آورد و کتاب‌های مورد نظر خود را به آن بیفزاید.

نویسندگان

در Goodreads برای نویسندگان صفحه‌ای اختصاصی تولید می‌شود که در آن اطلاعات نویسنده – همچون بیوگرافی جامع، آدرس سایت، تاریخ تولد و ... – آورده شده است تا بهتر شناسایی شوند و راه‌های ارتباطی با نویسنده به شکل کامل در اختیار بازدیدکننده قرار گیرد.



⁵ Shelf

مجموعه‌های داده⁶



مجموعه‌های داده‌ای که در این تمرین به شما داده شده است در قالب CSV⁷ هستند. CSV نام یک قالب برای پرونده‌های متنی است که در آن مقادیر با استفاده از نماد کاما (,) از یکدیگر جدا می‌شوند. این قالب یکی از روش‌های پرطرفدار برای تبادل اطلاعات است. در زیر درباره هر ویژگی و نوع داده⁸ مربوط به مجموعه‌های داده‌ای که در اختیار شما قرار داده شده است، توضیح مختصری آمده است.

کاربران



اطلاعات مربوط به کاربران عادی این شبکه اجتماعی در پرونده‌ای به نام users.csv آورده شده است.



نام ویژگی	توضیح	نوع داده
id	شناسه یکتا	عدد صحیح
name	نام و نام خانوادگی	رشته‌ای از کاراکترها
place_of_birth	کشوری که در آن متولد شده است	رشته‌ای از کاراکترها
member_since	تاریخ عضویت	رشته‌ای از کاراکترها
favorite_authors	شناسه نویسندگان مورد علاقه	رشته‌ای از کاراکترها
favorite_genres	ژانرهای ادبی مورد علاقه	رشته‌ای از کاراکترها
want_to_read	شناسه کتاب‌هایی که قصد خواندنشان را دارند	رشته‌ای از کاراکترها
currently_reading	شناسه کتاب‌هایی که در حال خواندنشان هستند	رشته‌ای از کاراکترها
read	شناسه کتاب‌هایی که خوانده‌اند	رشته‌ای از کاراکترها

⁶ Datasets

⁷ Comma-Separated Values

⁸ Data Type



نویسندگان



اطلاعات مربوط به نویسندگان در پرونده‌ای به نام authros.csv آورده شده است.

نام ویژگی	توضیح	نوع داده
id	شناسه یکتا	عدد صحیح
name	نام و نام خانوادگی	رشته‌ای از کاراکترها
gender	جنسیت	رشته‌ای از کاراکترها
member_since	تاریخ عضویت	رشته‌ای از کاراکترها
year_of_birth	سال تولد	عدد صحیح
place_of_birth	کشوری که در آن متولد شده است	رشته‌ای از کاراکترها
genres	ژانرهای ادبی	رشته‌ای از کاراکترها

کتاب‌ها



اطلاعات مربوط به کتاب‌ها در پرونده‌ای به نام books.csv آورده شده است.

نام ویژگی	توضیح	نوع داده
id	شناسه یکتا	عدد صحیح
title	عنوان	رشته‌ای از کاراکترها
author_id	شناسه نویسنده	عدد صحیح
genre	ژانر ادبی	رشته‌ای از کاراکترها

رابطه دنبال کردن



اطلاعات مربوط به دنبال کردن کاربران این شبکه اجتماعی در پرونده follow_edges.csv فراهم گردیده است. دقت شود که تجزیه کردن این مجموعه داده تنها برای بخش امتیازی تمرین لازم است.

نام ویژگی	توضیح	نوع داده
id	شناسه یکتا مربوط به کاربر	عدد صحیح
followings	شناسه کاربرانی که وی فعالیتشان را دنبال می‌کند	رشته‌ای از کاراکترها
followers	شناسه کاربرانی که فعالیت وی را دنبال می‌کنند	رشته‌ای از کاراکترها

نظرات کاربران



اطلاعات مربوط به نظراتی که کاربران درباره کتاب‌ها داده‌اند در پرونده‌ای به نام reviews.csv آورده شده است.



نام ویژگی	توضیح	نوع داده
id	شناسه یکتا	عدد صحیح
book_id	شناسه کتابی که بر روی آن نظر داده شده است	رشته‌ای از کاراکترها
user_id	شناسه کاربری که نظر داده است	عدد صحیح
rating	امتیازی که به کتاب داده است	عدد صحیح
date	تاریخی که نظر خود را ثبت کرده است	رشته‌ای از کاراکترها
number_of_likes	تعداد کاربرانی که این نظر مورد علاقه‌شان بوده است	عدد صحیح

نکات تکمیلی

- ستون‌هایی که تحت عنوان id مشخص شده‌اند در آن مجموعه داده یکتا هستند.
- ستون‌هایی که حاوی اطلاعات تاریخ هستند، شامل روز، ماه و سال آن رویداد هستند.
- در پرونده‌های CSV، ستون‌هایی که بیشتر از یک داده را در خود ذخیره می‌کنند، اطلاعاتشان با استفاده از جداکننده⁹ '\$' از یکدیگر جدا شده است.
- **تضمین می‌شود** که در مجموعه‌های داده‌ای که در اختیار شما قرار گرفته است، از کاراکترهای '،' و '\$' تنها برای جداسازی اجزا استفاده شده است.

منابع

<https://www.goodreads.com>

<https://en.wikipedia.org/wiki/Goodreads>

<http://essay.farsheedpress.com/goodreads>

<https://www.iconfinder.com>

⁹ Delimiter