استخراج اطلاعات مورد نیاز از سایت imdb

با توجه به بخش توضیحات کلی تمرین برای انجام مرحله اول و استخراج اطلاعات از این لینک استفاده کنید. این کار را با استفاده از روشهای مختلف کراولینگ این اطلاعات انجام دهید. هر یک از این 250 فیلم خود شامل لینکی هستند که شما را به صفحه آن فیلم هدایت میکند.

یس از ورود به صفحه هر فیلم شما باید اطلاعات زیر را استخراج کنید:

- ۱. نام یا عنوان فیلم (title)
- ۲. سال انتشار فیلم (year)
- ۳. درجه سنی فیلم (parental_guide)
- ۴. مدت زمان فیلم بر حسب دقیقه (runtime)
 - ۵. ژانر یا ژانرهای فیلم (genre)
- ۶. نام کامل کارگردان یا کارگردانان (directore)
 - ۷. نام کامل نویسنده یا نویسندگان (writer)
 - ۸. نام کامل بازیگر یا بازیگران (star)
- ۹. توضیحات فیلم یا خلاصه فیلم (storyline) (اختیاری)
- ۱۰. میزان فروش فیلم در منطقه آمریکا و کانادا (gross_us_canada)

▼ راهنمایی مربوط به عنوان فیلم

دقت کنید برای اینکه اطلاعات شما برای فیلمهایی که به زبان انگلیسی نیستند مطابق قسمت مشخص شده در سوال باشد باید در header درخواست خود accept-language را انگلیسی تنظیم کنید.

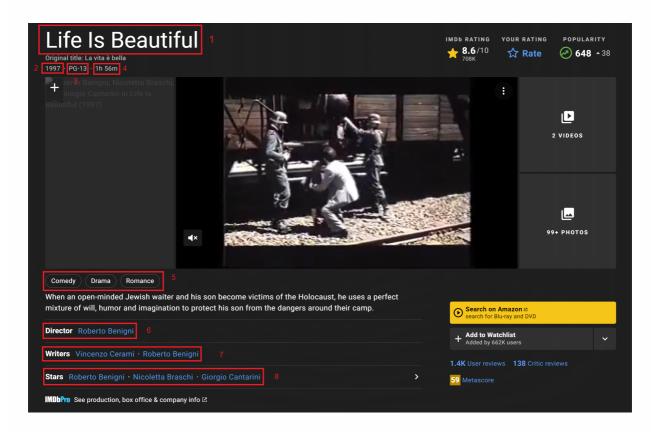
▼ راهنمایی مربوط به توضیحات فیلم

دقت کنید که توضیحات فیلم یا storyline در سایت به صورت پویا یا میباشد و الزاما میباشد و الزاما با روشهایی که اطلاعات دیگر را جمع اوری کردهاید ممکن است بدست نیاید. به عنوان مثال میتوانید از stoyline استفاده کنید، همچنین فقط برای stoyline میشود. همچنین برای راحتی شما امتیاز توضیحات فیلم به صورت اختیاری اعمال میشود.

▼ توجه

با توجه به اینکه اطلاعات مورد نیاز در بخشهای مختلفی از صفحه فیلم ممکن است تکرار شود اکیدا توصیه میشود که اطلاعات را از قسمتهای مشخص شده جمع اوری کنید.

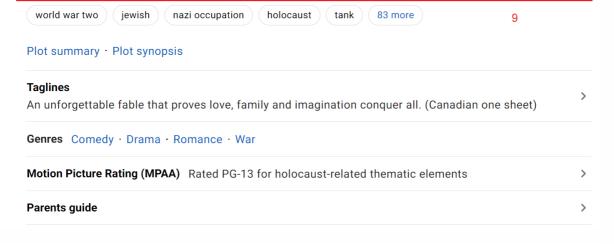
برای تشخیص هر یک از موارد بالا و جایگاه آنها در صفحه فیلم میتوانید از عکسهای زیر استفاده کنید.







In 1930s Italy, a carefree Jewish waiter named Guido starts a fairy tale life by courting and marrying a lovely woman from a nearby city. Guido and his wife have a son and live happily together until the occupation of Italy by German forces. In an attempt to hold his family together and help his son survive the horrors of a Jewish Concentration Camp, Guido imagines that the Holocaust is a game and that the grand prize for winning is a tank.—Anthony Hughes <a href="https://www.nushcombo.nushc



Box office

Budget

\$20,000,000 (estimated)

Opening weekend US & Canada

\$118,920 · Oct 25, 1998

IMDbPro See detailed box office info on IMDbPro ☑

Gross US & Canada

\$57,563,264

Gross worldwide

\$230,099,013

همچنین برای ذخیره اطلاعات در پایگاه داده نیاز به شناسه (id) فیلمها و افراد موجود در سایت است. برای بدست آوردن شناسه فیلمها و افراد باید به لینک هر یک توجه کرد و قسمت عددی بعد از title و name را بدون در نظر گرفتن قسمت رشتهای آنها (tt و nm) به عنوان شناسه لحاظ کرد. برای درک بهتر شناسه به مثالهای زیر توجه کنید:

https://www.imdb.com/title/tt0111161/

شناسه فيلم: 0111161

10

https://www.imdb.com/name/nm0000209/

شناسه فرد: 0000209

تذکر: شما برای استخراج اطلاعات مجاز به استفاده از api نیستید و در صورت استفاده از api سایت imdb نمرهای به بخش اول تعلق نخواهد گرفت.

ارسال پاسخ

پاسخ شما باید یک فایل پایتونی با نام crawl.py باشد که با اجرای آن مطلوبات مساله به صورت کامل انجام شود.