

فاز اول پروژه

**پیشبینی رده سنی بر اساس خلاصه فیلم**

پردازش زبان‌ و گفتار

**استاد درس**

دکتر مینایی

**دانشجویان**

امیرحسین احمدی – ملیکا احمدی رنجبر

بهار 1401

**فهرست مطالب**

[خلاصه فعالیت ها 3](#_Toc104322066)

[استخراج اطلاعات فیلم از IMDb 4](#_Toc104322067)

[پیش پردازش 6](#_Toc104322068)

[آمار 8](#_Toc104322069)

# خلاصه فعالیت ها

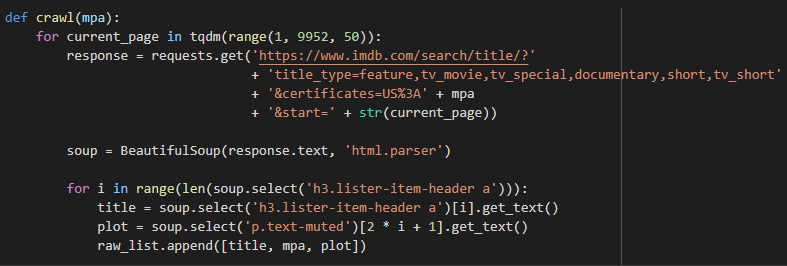
در فاز اول این پروژه با Crawling روی صفحات مربوط به رده های سنی (MPA) در سایت IMDb اطلاعات مربوط به فیلم و خلاصه فیلم را بدست آورده و ذخیره میکنیم. در ادامه داده های بدست آمده را نرمال سازی کرده و داده های تمیز شده را ذخیره میکنیم. در انتها نیز آمار خواسته شده مربوط به داده ها را بدست میاوریم.

# استخراج اطلاعات فیلم از IMDb

ابتدا تابع crawl را تعریف کرده که با گرفتن MPA (G, PG, PG-13, R) یک Request به صفحه مربوط به آن زده و اطلاعات آن را دریافت میکنیم. (در هر صفحه 50 فیلم وجود دارد به همین دلیل Index مربوط به شروع صفحه را در هر Request 50 تا زیاد میکنیم تا تمام فیلم های مربوط به آن رده را دریافت کنیم.) نمونه لینک برای رده سنی PG-13 :

[https://www.imdb.com/search/title/?title\_type=feature,tv\_movie,tv\_special,documentary,short,tv\_short&certificates=US%3A**PG-13**&start=1](https://www.imdb.com/search/title/?title_type=feature,tv_movie,tv_special,documentary,short,tv_short&certificates=US%3APG-13&start=1)

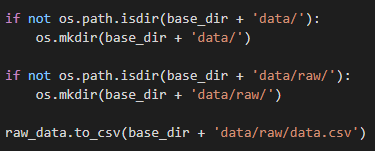
سپس با استفاده از ابزار BeautifulSoap اطلاعات درون صفحه را استخراج کرده و Title، Plot و MPA آن را درون لیست raw\_list اضافه میکنیم.



در ادامه با صدا زدن تابع crawl برای تمام MPA ها، raw\_list را آپدیت میکنیم و با تبدیل raw\_list به Pandas DataFrame، 5 رکورد اول آن را نمایش میدهیم.(در ابتدای هر Plot یک ‘\n’ آمده که آن را نیز در این مرحله حذف میکنیم.)



در آخر با ساخت پوشه های مورد نیاز DataFrame ساخته شده را در قالب csv ذخیره میکنیم.

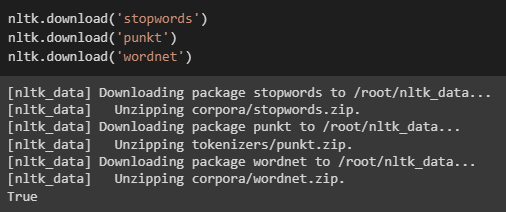


# پیش پردازش

در ابتدا داده خام ذخیره شده در مرحله قبل را میخوانیم.

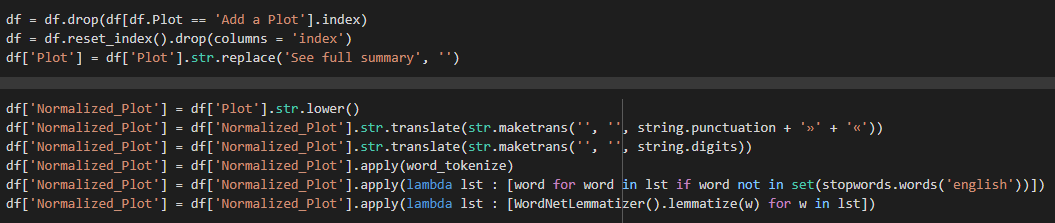


سپس داده های مورد نیاز کتابخانه nltk را دریافت کرده که در ادامه از آن ها استفاده خواهیم کرد.

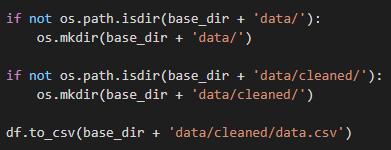


بعضی از فیلم های موجود در IMDb خلاصه فیلم ندارند که آن ها را از داده ها حذف میکنیم. همچنین بعضی از خلاصه ها به دلیل بلند بودن در صفحه ها نیامده اند و عبارت ‘See full summary’ در انتهای Plot آن ها نوشته شده که این عبارت را از خلاصه ها حذف میکنیم.

در ادامه Normalization های مرسوم را با استفاده از nltk و کتابخانه string در Python روی داده ها اعمال میکنیم. ابتدا تمام کلمات Plot ها را Lowercase کرده، سپس علامت گذاری ها و اعداد را حذف میکنیم. در نهایت Plot ها را Tokenized کرده، Stop Word ها را حذف کرده و عمل Lemmatization را روی آن انجام میدهیم و در ستون Normalized\_Plot ذخیره میکنیم.

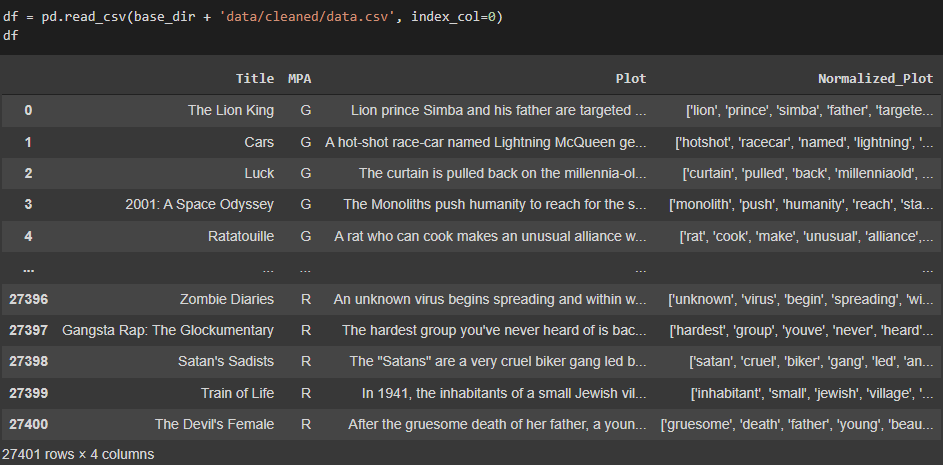


در آخر با ساخت پوشه های مورد نیاز داده های تمیز شده را در قالب csv ذخیره میکنیم.

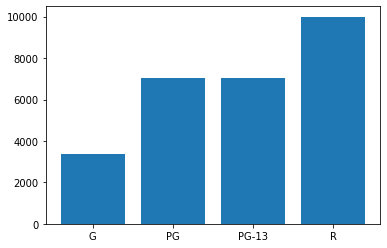


# آمار

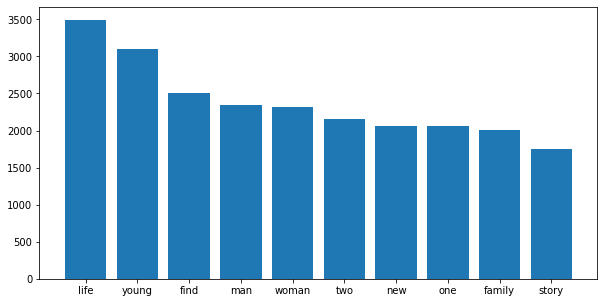
ابتدا داده های تمیز شده در مرحله قبل را میخوانیم.



سپس تعداد داده های مربوط به هر MPA را درون یک Bar Plot نمایش میدهیم که در زیر مشاهده میکنید. همان طور که پیداست تعداد فیلم های با درجه سنی G در IMDb از بقیه کمتر و بیشترین فیلم به درجه سنی R تعلق دارد.



در ادامه با استفاده از تابع Counter در کتابخانه collections تعداد کلمات موجود در Dataset را شمرده و در Bat Plot زیر نمایش داده ایم.



در آخر تعداد جمله ها، تعداد کل کلمات بعد از Normalize کردن داده ها و تعداد کلمات یکتا را میتوانید مشاهده کنید.

