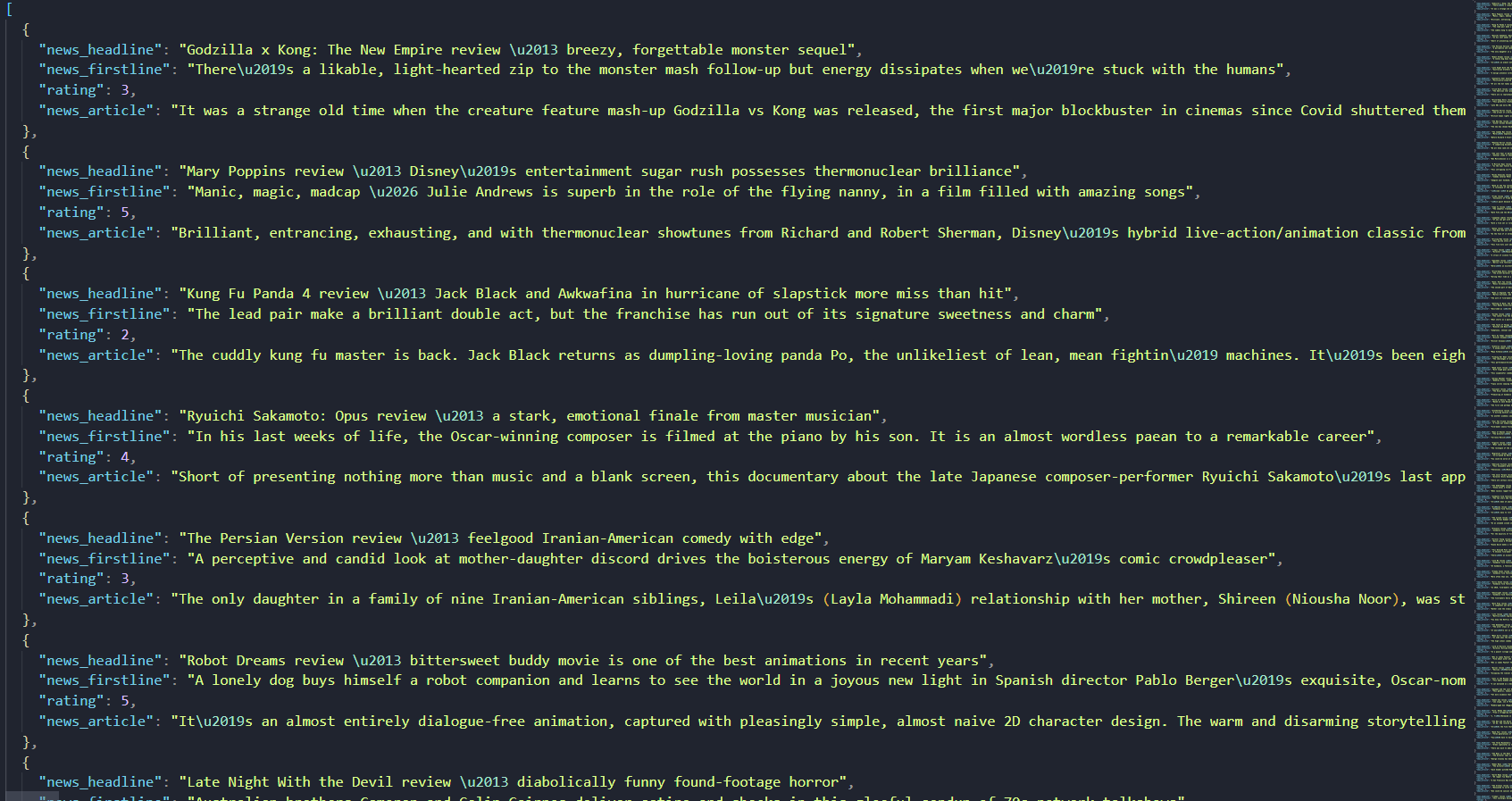
برای این تمرین داده ها را از سایت www.theguardian.com ، crawl کردیم. دادگان جمع آوری شده از بخش نقد فیلم ها میباشد. در ابتدا لینک های موجود در هر صفحه اصلی که لینک مرتبط به نقد هر فیلم میباشد را به دست می آوریم.



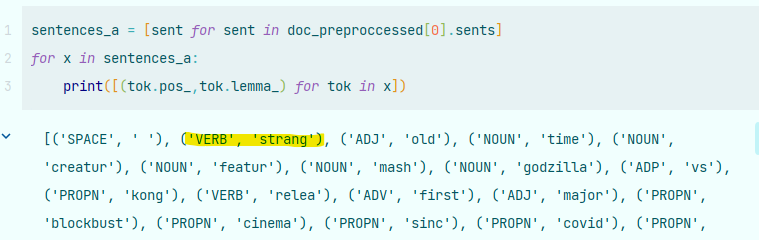
سپس از هر لینک اطلاعات را بدست می آوریم. اطلاعات ذخیره شده شامل هدلاین نقد ، امتیاز داده شده به فیلم و متن نقد و یک خط خلاصه نوشته شده میباشد.

اطلاعات کرال شده را در نهایت در یک فایل json ذخیره سازی میکنیم.



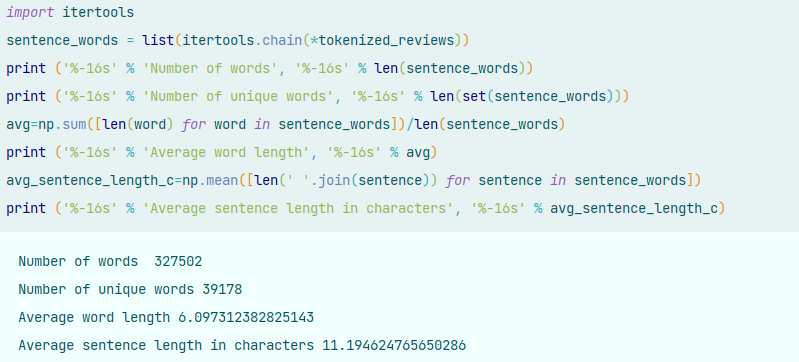
پس از اینکار به سراغ پردازش دادگان میرویم. در بخش اول تسک انتخاب شده به دست آوردن عبارات کلیدی و صفات استفاده شده در این نقد ها میباشد. مراحل مرتبط به کد ها در نوتبوک به شکل مناسب کامنت گذاری شده است.

برای اینکار باید پیش پردازش های گفته شده را انجام دهیم تا انجام اینکار راحت تر شود و اطلاعات با ارزش تر باقی بمانند. در ابتدا تمامی Punctuation ها را حذف میکنیم. سپس تمامی کلمات را Lowercase کرده و در نهایت ‘s های موجود در هر کلمه را حذف میکنیم. برای نشان دادن عملکرد از Porter stemmer استفاده کردیم. با استفاده از stemming اتفاقی که می افتد این است که پس از دادن دادگان به Spacy برای شناسایی صفات و عبارات کلیدی این مدل به علت stemming اتفاق افتاده دچار مشکل میشود و نمیتواند این کار را به شکل درست انجام دهد. به عنوان مثال کلمه strange در هنگام stemming به strang تبدیل میشود و آن را verb تشخیص میدهد که این اشتباه با توجه به اینکه تسک ما تشخیص بیشترین صفات به کار رفته در تحلیل فیلم ها میباشد غلط میباشد.





اطلاعات مرتبط با دیتای استفاده شده بعد از پیش پردازش گفته شده در بالا. این پیش پردازش شامل stemming نمیباشد.



با مقایسه داده ای که این پیش پردازش ها بر روی آن انجام شده در مقابل داده ای که فقط جملات توکنایز شده اند مشخص میشود با توجه به اینکه کلمات حذف شده تاثیری در تسک گفته شده ندارن، با حذف آن ها میتواند اطلاعات را سریعتر و به صورت موثرتری بدست آورد.

همانطور که در تصویر زیر مشخص است توانسته ایم تمامی Noun\_chuncks ها را بدست آوریم.



پس از آن بیشترین صفات استفاده شده توسط نویسندگان در این نقد ها را بدست آوردیم.



در بخش دوم با بررسی هدلاین های هر نقد با توجه به پترن موجود در آن سعی میکنیم که اسم فیلم را استخراج کنیم.

با توجه به اینکه به دنبال اسم فیلم هستیم پیش پردازش خاصی را مد نظر نداریم چون در صورت حذف کردن حتی stopwords اسم فیلم ممکن است متفاوت شود. تنها کاری که میکنیم توکنایز کردن جملات است. سپس با استفاده از regex پترن مدنظر را وارد کرده و سعی میکنیم اسم فیلم را بدست آوریم.

همانطور که در تصویر زیر مشخص است با استفاده از ابزار پردازش متن توانسته ایم با دقت مناسبی اسامی فیلم ها را از هدلاین نقد ها استخراج کنیم. این اسامی میتواند برای بهتر و کامل کردن دادگان استفاده شده و سبب شود که به ازای اسم فیلم ما نقد نوشته شده را نگه داری کنیم.

