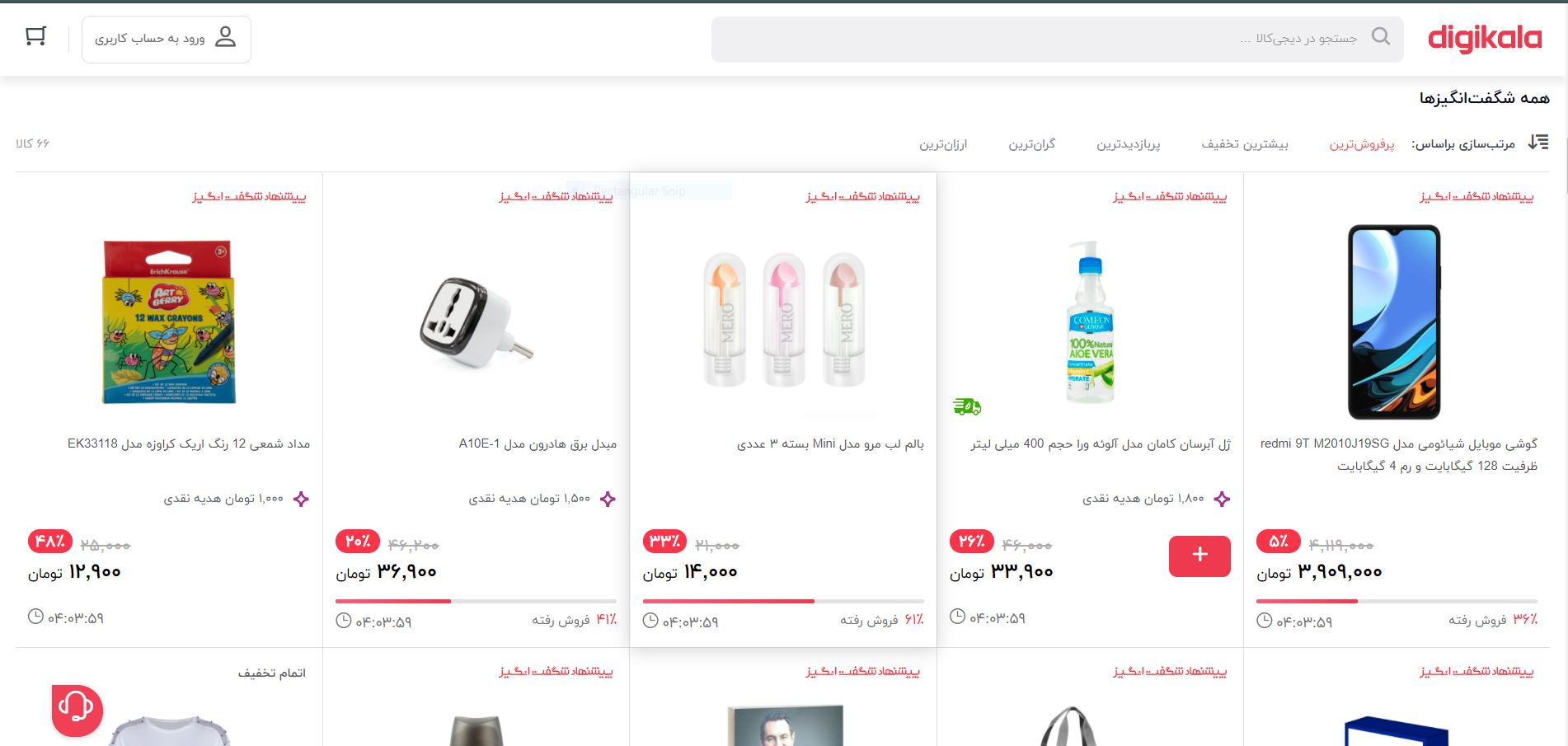
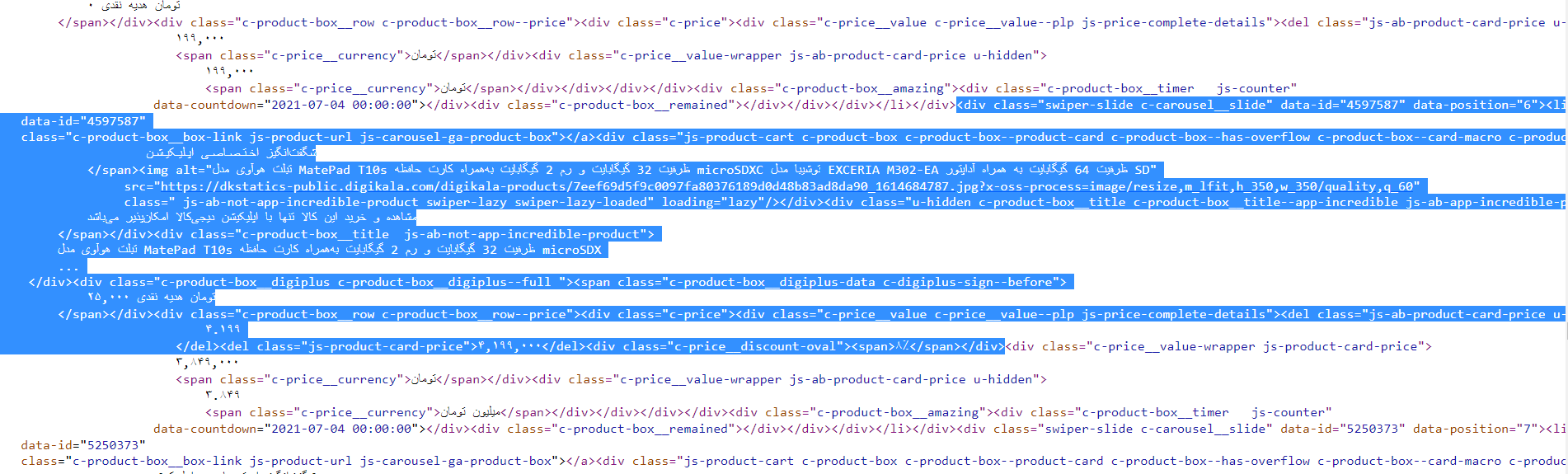
**شرح پروژه**

هدف از پروژه web scraping صفحه فروش شگفت انگیز دیجیکالا، سرعت بخشیدن به روند خرید و فروش و کسب آسان تر اطلاعات هست. در این پروژه از زبان برنامه نویسی پایتون و پکیج های pandas, beautifulSoup4, urllib بهره گرفته شده است. ابتدا به بررسی صفحه اصلی پرداخته و سپس کد توضیح داده میشود.



<https://www.digikala.com/incredible-offers/>

در این صفحه قسمت های متعددی وجود دارند که بخش بالا، قسمت مد نظر ما هست. که همه کالا ها با تخفیف مذکور لیست شده اند. در سورس صفحه در این قسمت اطلاعات زیر موجود هستند:



که اگر مرتب شوند دیده میشود که داده ها در تگ های متناسب قسمت بندی شده اند.

نام ها در تگ div class={c-product-box\_\_title js-ab-not-app-incredible-product}

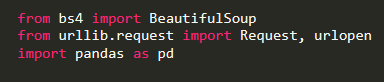
قیمت ها در تگ div class={c-price\_\_value-wrapper js-product-card-price}

تخفیف ها در تگ div class={c-price\_\_discount-oval}

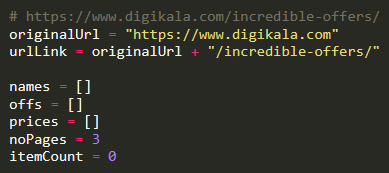
با استخراج داده های این تگ ها داده مورد نیاز خود را بدست میآوریم.

**در توضیح کد**

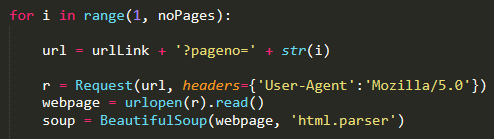
ابتدا کتابخانه های مورد نیاز را وارد میکنیم



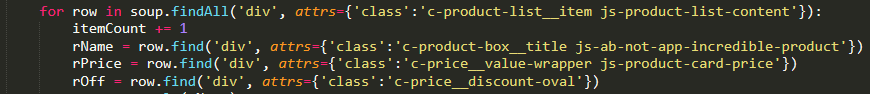
سپس آدرس سایت و برای داده های خود 3 لیست در نظر میگیریم. همچنین برای تعداد صفحات مورد بررسی و تعداد کالا ها متغیر هایی در نظر میگیریم



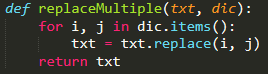
وارد حلقه برای تعداد صفحات میشویم، برای هر صفحه ابتدا لینک کامل را ساخته، سپس به آن سایت و سرور درخواست فرستاده و جواب را میخوانیم و به کمک کتابخانه bs4 اطلاعات صفحه را پردازش میکنیم



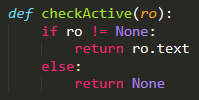
بعد از دریافت page source، مطابق تگ های توضیح داده شده در قسمت قبل، بخش کالا های خود را جدا کرده و در آن بخش ها داده های مورد نیاز خود را جدا میکنیم، همچنین به ازای هر کالا مقدار متغیر شمارش کالا یکی اضافه میشود



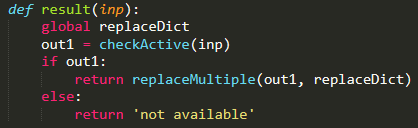
داده اسم کالا در سورس صفحه شامل فاصله های خالی و اضافی هست. این مشکل در تابع replaceMultiple رفع شده است، بدین صورت که یک dict برای تعویض فاصله های اضافی با فضای خالی در نظر گرفته شده، که در این تابع به ترتیب این تعویض ها انجام میشوند



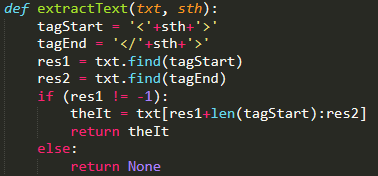
همچنین در صورت اتمام تخفیف، داده قیمت مقدار None به خود میگیرد که برای رفع آن مشکل از تابع checkActive استفاده شده. در این تابع شرطی بررسی میشود که اگر کالا موجود بود، قیمت آن و اگر نبود خروجی Noneداده شوند



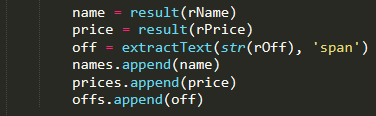
برای راحتی نوشتار، این دو عملیات در تابع result باهم ادغام شده اند



داده تخفیف نیز در div مخصوص خود، در قسمت span قرار دارد. با توجه به مشکلات استخراج داده span، برای بدست آوردن مقدار تخفیف، از عملیات روی متن استفاده شده است، بدین صورت که متن بین تگ های span و /span خارج میشود. این عملیات در تابع extractText پیاده سازی شده است



پس از آماده سازی اولیه داده های نام و قیمت و تخفیف، آنها را در لیست های مربوطه اضافه میکنیم



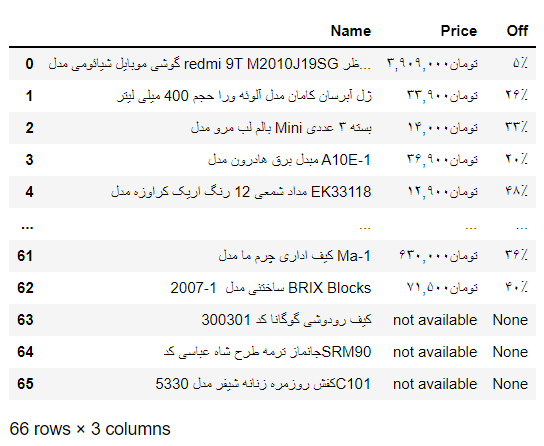
پس از بررسی همه سایت ها(که در این مورد صفحات 1 و 2 موجود بودند) داده ها را در یک dict ذخیره میکنیم و سپس آن را به فرمت DataFrame از پکیج pandas در میآوریم. هدف از این کار راحتی و امکانات کار با داده در pandas هست



سپس میتوانیم داده ها را به فرمت csv تبدیل کنیم تا در سایر برنامه ها نیز قابلیت خواندن داشته باشد، لازم به ذکر است چون داده ها به زبان فارسی هستند، از utf-8 encoding استفاده میکنیم



* داده هایی که اتمام موجودی باشند، قیمت و تخفیف نخواهند داشت
* داده هایی که تخفیف آنها به پایان رسیده باشد، در ستون تخفیف مقداری ندارند.



همچنین برای راحتی بیشتر، در قسمت اضافی پروژه، داده های برست آمده در یک صفحه وب نیز ذخیره میشوند



