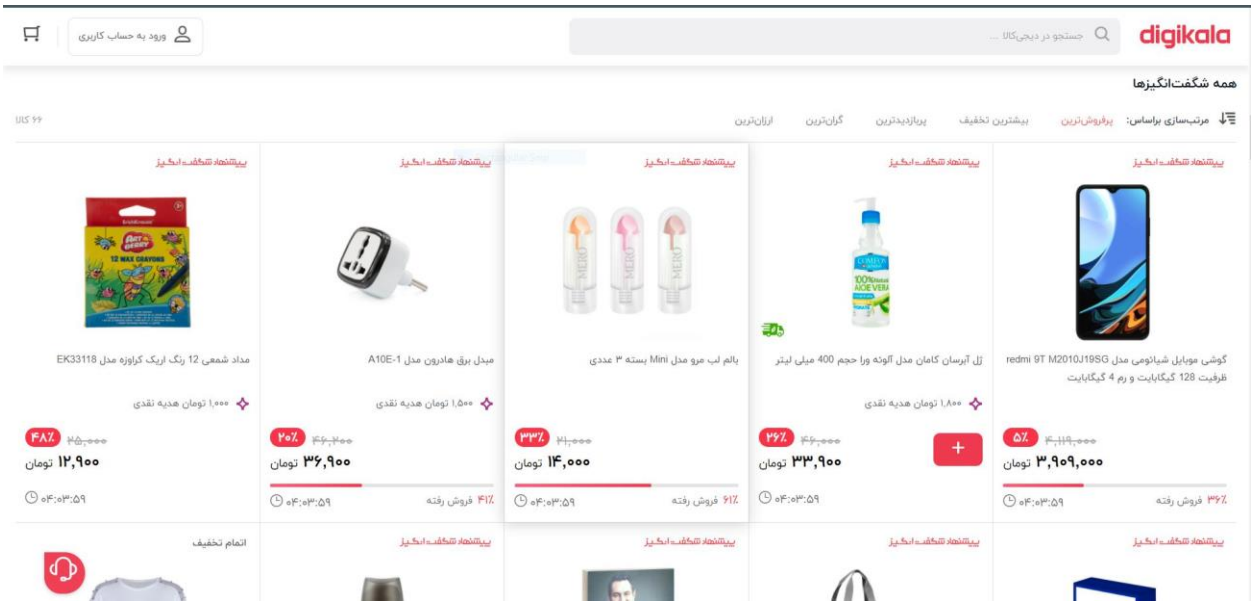
شرح پروژه

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

```
from bs4 import BeautifulSoup
from urllib.request import Request, urlopen
import pandas as pd
```

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

```
# https://www.digikala.com/incredible-offers/
originalUrl = "https://www.digikala.com"
urlLink = originalUrl + "/incredible-offers/"

names = []
offs = []
prices = []
noPages = 3
itemCount = 0
```

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

```
for i in range(1, noPages):

    url = urlLink + '?pageno=' + str(i)

    r = Request(url, headers={'User-Agent':'Mozilla/5.0'})
    webpage = urlopen(r).read()
    soup = BeautifulSoup(webpage, 'html.parser')
```

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

```
    for row in soup.findAll('div', attrs={'class':'c-product-list__item js-product-list-content'}):
        itemCount += 1
        rName = row.find('div', attrs={'class':'c-product-box__title js-ab-not-app-incredible-product'})
        rPrice = row.find('div', attrs={'class':'c-price__value-wrapper js-product-card-price'})
        rOff = row.find('div', attrs={'class':'c-price__discount-oval'})
```

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

```
def replaceMultiple(txt, dic):
    for i, j in dic.items():
        txt = txt.replace(i, j)
    return txt
```

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 None داده شوند

```
def checkActive(ro):
    if ro != None:
        return ro.text
    else:
        return None
```

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

```
def result(inp):
    global replaceDict
    out1 = checkActive(inp)
    if out1:
        return replaceMultiple(out1, replaceDict)
    else:
        return 'not available'
```

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

```
def extractText(txt, sth):
    tagStart = '<'+sth+'>'
    tagEnd = '</'+sth+'>'
    res1 = txt.find(tagStart)
    res2 = txt.find(tagEnd)
    if (res1 != -1):
        theIt = txt[res1+len(tagStart):res2]
        return theIt
    else:
        return None
```

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

```
        name = result(rName)
        price = result(rPrice)
        off = extractText(str(rOff), 'span')
        names.append(name)
        prices.append(price)
        offs.append(off)
```

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

```
dic = {'Name':names, 'Price':prices, 'Off':offs}
df = pd.DataFrame(dic)
```

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

```
df.to_csv('file1.csv', index=False, encoding='utf-8')
```

- 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.

	Name	Price	Off
0	...ظر redmi 9T M2010J19SG گوشی موبایل شیائومی مدل	۳,۹۰۹,۰۰۰تومان	۵٪
1	ژل آبرسان کامان مدل آلوئه ورا حجم 400 میلی لیتر	۳۳,۹۰۰تومان	۲۶٪
2	بسته ۳ عددی Mini بالم لب مرو مدل	۱۴,۰۰۰تومان	۳۳٪
3	A10E-1 مبدل برق هادرون مدل	۳۶,۹۰۰تومان	۲۰٪
4	EK33118 مداد شمعی 12 رنگ اریک کراوزه مدل	۱۲,۹۰۰تومان	۴۸٪
...	...	...	...
61	Ma-1 کیف اداری چرم ما مدل	۶۳۰,۰۰۰تومان	۳۶٪
62	BRIX Blocks ساختنی مدل 2007-1	۷۱,۵۰۰تومان	۴۰٪
63	کیف رودوشی گوگانا کد 300301	not available	None
64	SRM90جانماز ترمه طرح شاه عباسی کد	not available	None
65	C101کفش روزمره زنانه شیفر مدل 5330	not available	None

66 rows × 3 columns

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

```
df.to_html('file2.html')
```

	Name	Price	Off
0	ظرفیت 128 گیگابایت و رم 4 گیگابایت redmi 9T M2010J19SG گوشی موبایل شیائومی مدل	۳,۹۰۹,۰۰۰تومان	۵٪
1	ژل آبرسان کامان مدل آلوئه ورا حجم 400 میلی لیتر	۳۳,۹۰۰تومان	۲۶٪
2	بسته ۳ عددی Mini بالم لب مرو مدل	۱۴,۰۰۰تومان	۳۳٪
3	A10E-1 مبدل برق هادرون مدل	۳۶,۹۰۰تومان	۲۰٪
4	EK33118 مداد شمعی 12 رنگ اریک کراوزه مدل	۱۲,۹۰۰تومان	۴۸٪
5	بسته 500 عددی A+ فولیو پرایم مدل A4 کاغذ	۶۶,۵۰۰تومان	۱۹٪
6	FCLT3337 ساک لوازم کودک و نوزاد فوروارد کد	۱۶۹,۰۰۰تومان	۷۴٪
7	کتاب اثر مرکب اثر دارن هاردی انتشارات باران خرد	۱۹,۶۰۰تومان	۸۰٪
8	حجم 150 میلی لیتر Bvlgari اسپری ضد تعریق مردانه پرستیژ مدل	۱۲,۹۰۰تومان	۱۹٪
9	تی شرت زنانه کد 120	۶۰,۰۰۰تومان	None
10	حجم 85 میلی لیتر Warrior ادو پرفیوم مردانه اسکلاره مدل	۱۵۵,۹۵۰تومان	None
11	کتاب پنج قدم فاصله اثر ریچل لیپینکات نشر میلکان	۳۶,۸۰۰تومان	۳۲٪
12	مجموعه 6 عددی A211 جوراب مردانه مدل نویدکد	۷۴,۵۰۰تومان	۵۷٪
13	FCLT3060 کیف ابزار فوروارد مدل	۴۹,۰۰۰تومان	۵۹٪
14	ظرفیت 10000 میلی آمپر ساعت PB01 شارژر همراه رانیک پرایم مدل	۱۹۹,۰۰۰تومان	۴۵٪
15	FCLT402-2 کوله پشتی 65 لیتری فوروارد مدل	۳۲۹,۹۰۰تومان	۶۱٪
16	روان نویس طرح یونیکورن کد 9911	۲۹,۰۰۰تومان	۲۸٪
17	سرویس خواب ژینورا مدل هانا یکنفره 2 تکه	۴۰۰,۰۰۰تومان	None