Парсер Ozon.ru

Нужно написать парсер ozon.ru, который будет проходить по принципу робота с главной страницы по другим страницам и собирать некоторые данные. На выходе нужны json файлы с данными.

В первой итерации нужен парсер, который будет ходить только по категориям товаров (url содержит “/category/”).

Сложности

Ozon сильно защищается от парсинга. Постоянно выдается hCaptcha. При парсинге нужно использовать прокси и антикапчу. Парсер нужно будет делать с помощью JS парсера.

Входные данные

На вход подается список url ozon.ru. По умолчанию должна идти главная страница ozon.ru.

Паттерн url для прохода бота и добавления в очередь в формате регулярного выражения RegEx
по умолчанию:
https://www.ozon.ru/category/*

Но нужно исключить url вида (с УВ > 3 по числу слешей в url):
https://www.ozon.ru/category/platya-zhenskie-36397/a-a-awesome-apparel-by-ksenia-avakyan-82805272/

Паттерн url для парсинга (url с которых будут собираться данные) в формате регулярного выражения RegEx
по умолчанию:
https://www.ozon.ru/category/*

Глубина парсинга
Уровень вложенности целевых страниц для парсинга по отношению к стартовому множеству URL (по умолчанию к главной).

Задача

Парсер по принципу поискового робота ходит по внутренним ссылкам заданным входными параметрами и собирает различные данные. Со всех внутренних страниц

также продолжают собираться ссылки и добавляются в очередь для дальнейшего парсинга.

По умолчанию проход идет с главной страницы ozon.ru. Т.е. идет клик по меню бургер и получение оттуда стартового списка категорий.
xpath меню бургер:
//div[@id="stickyHeader"]//div[@data-widget="catalogMenu"]

Выходные данные

На выходе должен быть JSON файл:
Ozon Data [% datefile.format(format => '%d %m %Y') %].json

Для блоков парсинга указана область Xpath где нужно искать (но часто не полный xpath) и скриншот.
Для всех собираемых ссылок должны собираться как анкор ссылки, так и url ссылки.

Во всех текстовых блоках теоретически могут быть разрывы строк \r\n. Такие разрывы на выходе нужно заменять на текстовые “\R\N”. Если разрыв идет просто html тегом
 - на выходе так и оставляем html код.

Для каждой страницы парсинга каталогов должны собираться следующие данные.
url - исходный url страницы, для которого собираются данные

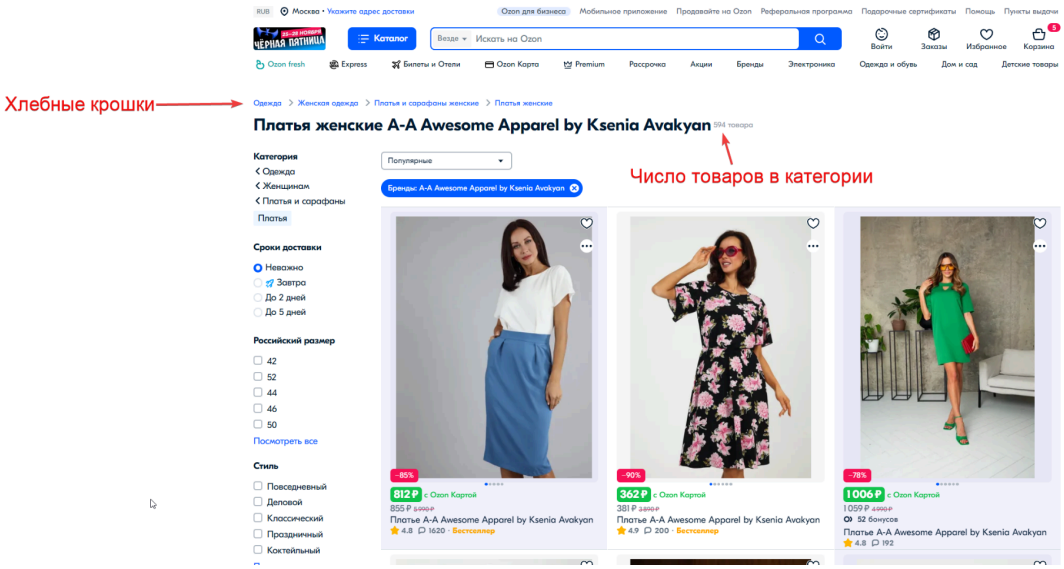
level - уровень вложенности url (по умолчанию по отношению к главной странице)

title - содержимое тега <title> страницы

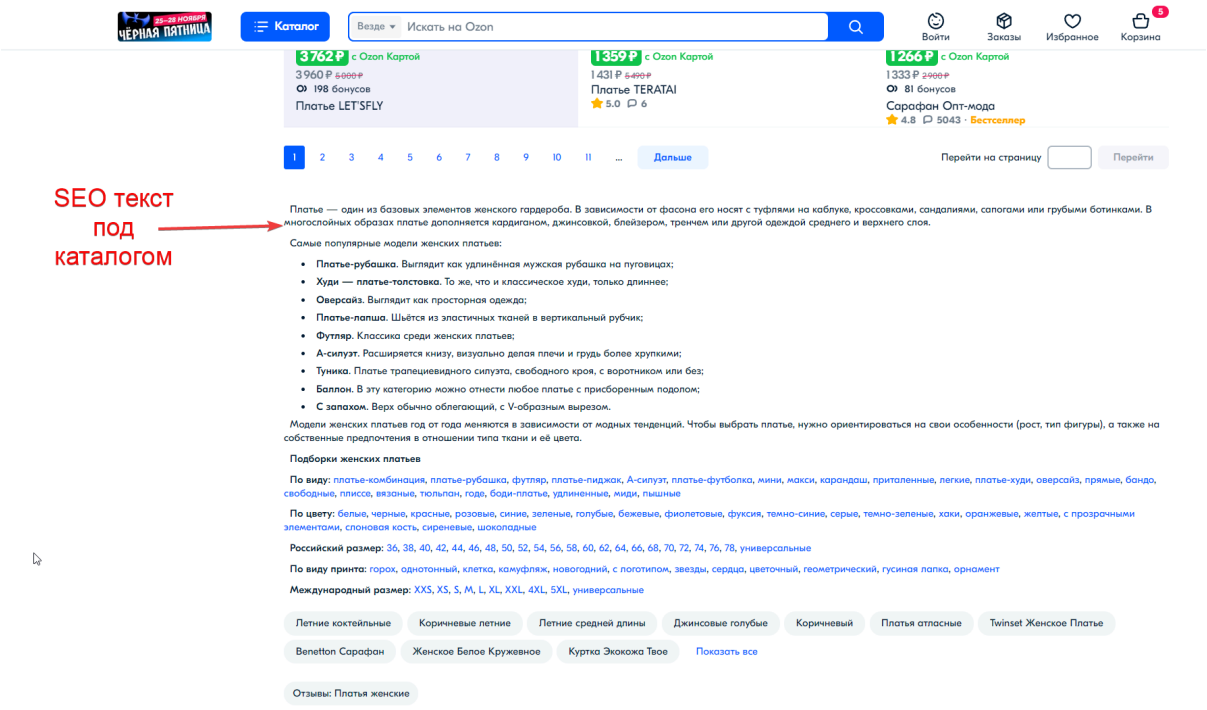
description - содержимое тега meta description страницы

h1 - содержиое тега H1. На ozon на каждой странице только 1 такой тег.

breadcrumbs - хлебные крошки страницы. Содержат url (в json обозначить u) и анкор (в json обозначить a).
xpath на исходном сайте:
//div[@data-widget="breadCrumbs"]



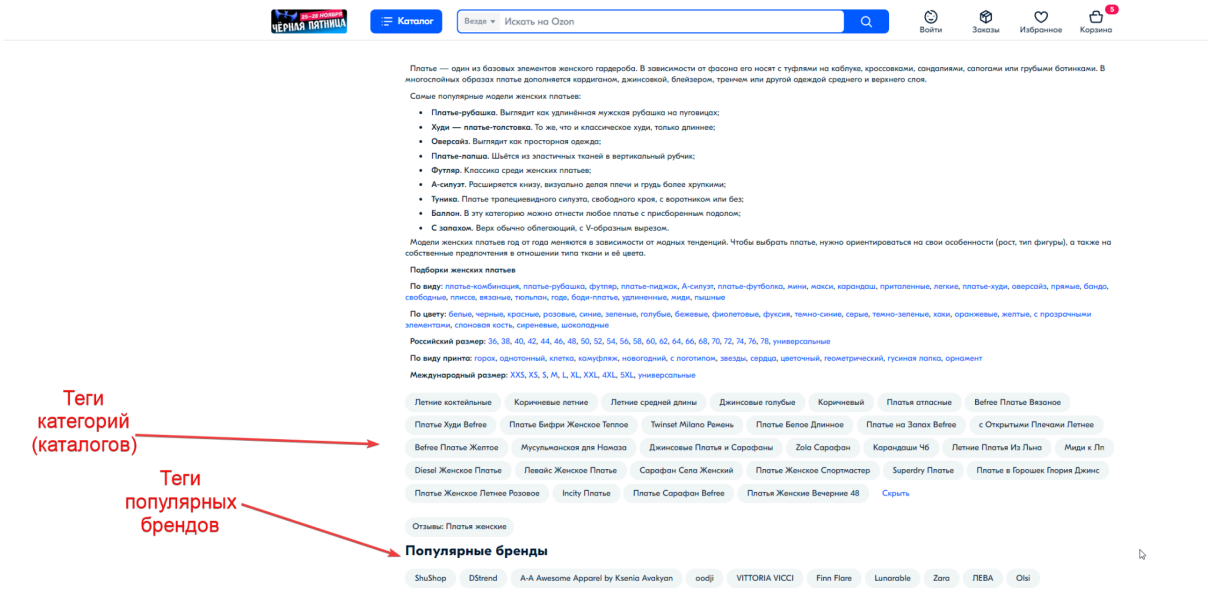
catalogBottomSEOText - HTML код с текстом под каталогом товаров. Может содержать блоки ссылок. В выходном JSON текст должен находится в элементе catalogBottomSEOText.text, а ссылки в массивах catalogBottomSEOText.links (для url и анкоров отдельные элементы u, a).
xpath на исходном сайте:
//div[@data-widget="semanticText"]
//div[@data-widget="semanticText"]//a[contains(@href, "/category/")]



catalogTags - Ссылки на теги (на озон это тоже категории). Включает в себя как ссылки над каталогом товаров, так и под ним. Также включает ссылки на популярные бренды

(в json отдельный массив). Перед парсингом нужно будет имитировать клик на “Показать все” под каталогом товаров, чтобы развернулись все теги.
В выходном Json должен содержать в себе вложенные массивы тегов категорий (catalogTags.category.u и catalogTags.category.a) и тегов популярных брендов (catalogTags.popularBrands.u и catalogTags.popularBrands.a)

xpath на исходном сайте:
//div[@data-widget="tagList"]//a[contains(@href, "/category/")]

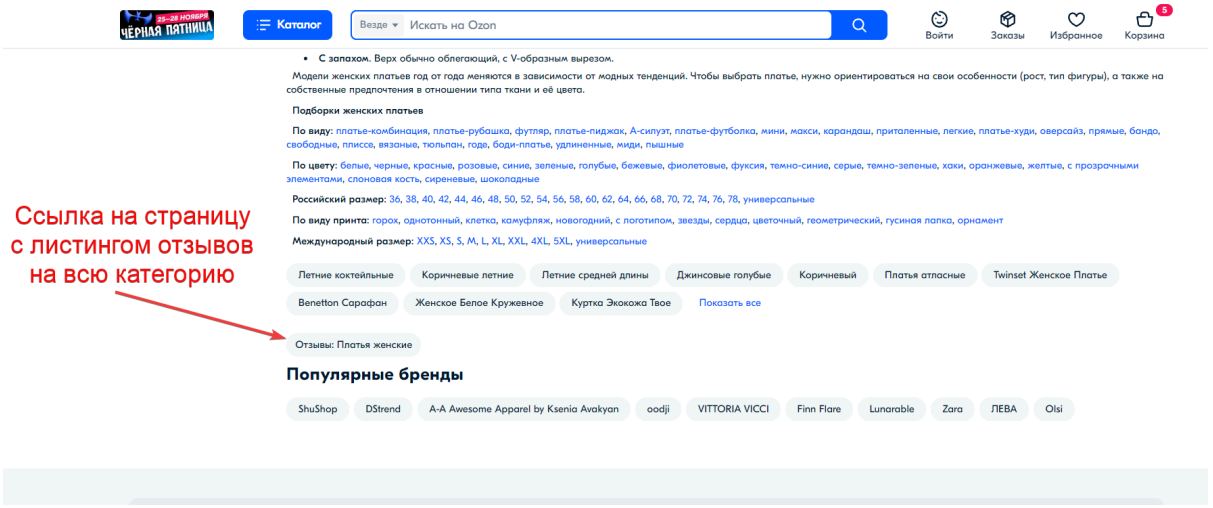


countItems - число товаров в каталоге

xpath на исходном сайте:
//div[@data-widget="resultsHeader"]//div[@class="je0"]

catalogReviews - ссылка на листинг с отзывами на всю категорию.

xpath на исходном сайте:
//div[@data-widget="tagList"][@class="se0 s0e s1e"]//a



Пример выходных данных

Приведем пример выходных данных для url:
https://www.ozon.ru/category/platya-zhenskie-36397/

Ограничимся для примера 2мя элементами в каждом массиве (теги и др.)

JSON в примере ниже похоже не совсем валидный, на выходе должен быть валидный!

{
url: "https://www.ozon.ru/category/platya-zhenskie-36397/",
level: 3,
title: "Платья женские купить в интернет-магазине OZON",
description: "Платья для женщин в интернет-магазине OZON по выгодным ценам от 600 рублей! ✔Характеристики ✔Стоимость ✔Фото ✔Огромный ассортимент каталога ✔Отзывы покупателей! Бесплатная доставка 🚚 по всей России!",
h1: "Платья женские",

breadCrumbs: [
{
a: "Одежда",
u: "https://www.ozon.ru/category/odezhda-obuv-i-aksessuary-7500/",
},
{
a: "Женская одежда",
u: "https://www.ozon.ru/category/zhenskaya-odezhda-7501/",
},
{
a: "Платья и сарафаны женские",
u: "https://www.ozon.ru/category/platya-zhenskie-7502/",
},
],

catalogBottomSEOText: [
{
text: "<p>Платье является одним из базовых элементов женского гардероба. В зависимости от фасона его носят с туфлями на каблуке, кроссовками, сандалиями, сапогами или грубыми ботинками. В многослойных образах платье дополняется кардиганом, джинсовкой, блейзером, тренчем или другой одеждой среднего и верхнего слоя....",
links: [
{
a: "платье-комбинация",
u: "https://www.ozon.ru/category/platya-kombinatsii/",
},
{

a: "платье-рубашка",
u: "https://www.ozon.ru/category/platya-rubashki/",
},
],
],

catalogTags: {
category: [
{
a: "Летние коктейльные",
u: "https://www.ozon.ru/category/platya-letnie-kokteylnye/",
},
{
a: "Коричневые летние",
u: "https://www.ozon.ru/category/platya-korichnevye-letnie/",
},
],

popularBrands: [
{
a: "ShuShop",
u: "https://www.ozon.ru/category/platya-zhenskie-36397/shushop-87305977/",
},
{
a: "DStrend",
u: "https://www.ozon.ru/category/platya-zhenskie-36397/dstrend-100099587/",
},
],
},

countItems: 446393,

catalogReviews: "https://www.ozon.ru/category/platya-zhenskie-36397/review/",

catalogReviews: [
{
a: "Отзывы: Платья женские",
u: "https://www.ozon.ru/category/platya-zhenskie-36397/review/",
},
],
}