Amirkabir University of Technology
(Tehran Polytechnic)

Report of the Web crawling and scraping

Professor:
Dr. M.Akbari (akbari.ma@aut.ac.ir)
Teaching Assistants:
A.Malekzade (malekzadeh@ieee.org) , Y.Ommi(yassi.ommi@gmail.com)

Student:
Mohammadreza Ardestani **(**ardestani.zr@gmail.com**)**

30, Dec, 2020

0) **Introduction**

- **0.1) How to Setup and run the code**
- **0.2) Houzz vs Lewis John?**
- **0.3) Output format**
- **0.4) Robots.txt**
- **0.5) Scrapy vs Selenium vs Beautiful Soup**

Phase 1) **Scraping Houzz**

First try: (Using bs4 for scraping Houzz website) Failed

Second try: (Using Scrapy) failed in runtime saving part

third try: (Using Scraping and additional functions) Successful

Phase 2) **Blogfa.com**

Phase 3) **IP rotation and personal data clearing**

Phase 4) **A little creativity**

+ Scraping Movie websites + BeautifulSoup + Pickle lib for pickling data

+ Implementing Scraping & Crawling with Octoparse toolkit

0) **Introduction**

0.1) How to Setup and run the code

In some cases you need to uncomment some lines(those that install libraries) and run other part's in order

This code is compatible with Colab and Python3

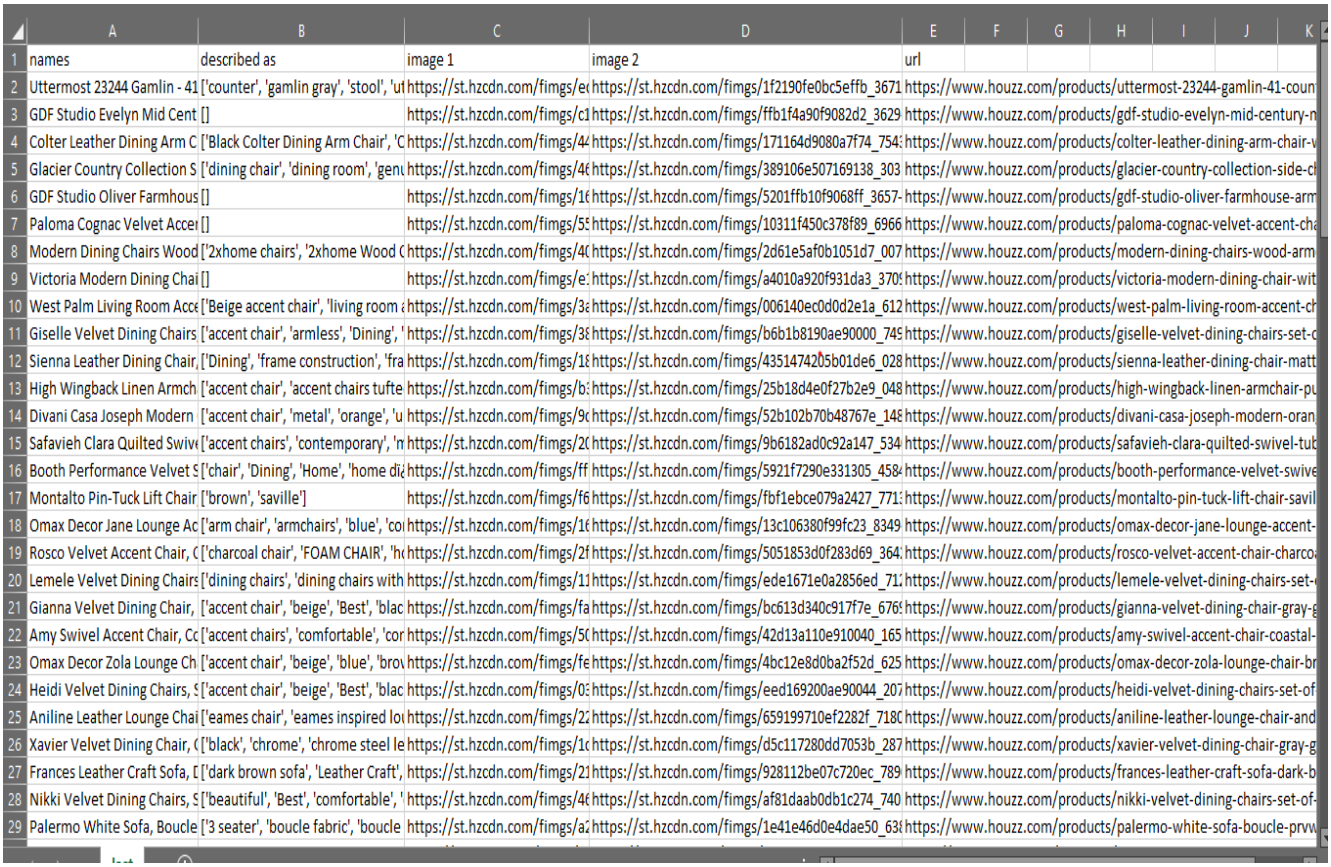
0.2) Houzz vs Lewis John?

At first Houzz (by now H) has rating and review for almost all of its products but Lewis John (by now L) has not review and rating for most of its products.

H is more simpler in terms of address of products but L is relatively complicated and is burst with a lot of branches and sub branches and it makes it even more difficult for Crawling.

In Robots.txt of H user have more permission for scraping products data than L. (I will explain Robots.txt later)

0.3) Output format



	A	B	C	D	E	F	G	H	I	J	K
1	names	described as	image 1	image 2	url						
2	Uttermost 23244 Gamlin - 41	['counter', 'gamlin gray', 'stool', 'ut	https://st.hzcdn.com/fimgs/e	https://st.hzcdn.com/fimgs/1f2190fe0bc5effb_3671	https://www.houzz.com/products/uttermost-23244-gamlin-41-coun						
3	GDF Studio Evelyn Mid Cent	[]	https://st.hzcdn.com/fimgs/c1	https://st.hzcdn.com/fimgs/ffb1f4a90f9082d2_3629	https://www.houzz.com/products/gdf-studio-evelyn-mid-century-n						
4	Colter Leather Dining Arm C	['Black Colter Dining Arm Chair', 'C	https://st.hzcdn.com/fimgs/4	https://st.hzcdn.com/fimgs/171164d9080a7f74_754	https://www.houzz.com/products/colter-leather-dining-arm-chair-v						
5	Glacier Country Collection S	['dining chair', 'dining room', 'gen	https://st.hzcdn.com/fimgs/4	https://st.hzcdn.com/fimgs/389106e507169138_303	https://www.houzz.com/products/glacier-country-collection-side-c						
6	GDF Studio Oliver Farmhous	[]	https://st.hzcdn.com/fimgs/1	https://st.hzcdn.com/fimgs/5201ffb10f9068ff_3657-	https://www.houzz.com/products/gdf-studio-oliver-farmhouse-arm						
7	Paloma Cognac Velvet Acce	[]	https://st.hzcdn.com/fimgs/5	https://st.hzcdn.com/fimgs/10311f450c378f89_6966	https://www.houzz.com/products/paloma-cognac-velvet-accent-cha						
8	Modern Dining Chairs Wood	['2xhome chairs', '2xhome Wood C	https://st.hzcdn.com/fimgs/4	https://st.hzcdn.com/fimgs/2d61e5af0b1051d7_007	https://www.houzz.com/products/modern-dining-chairs-wood-arm						
9	Victoria Modern Dining Chai	[]	https://st.hzcdn.com/fimgs/e	https://st.hzcdn.com/fimgs/a4010a920f931da3_370	https://www.houzz.com/products/victoria-modern-dining-chair-wit						
10	West Palm Living Room Acc	['Beige accent chair', 'living room a	https://st.hzcdn.com/fimgs/3	https://st.hzcdn.com/fimgs/006140ec0d0d2e1a_612	https://www.houzz.com/products/west-palm-living-room-accent-ch						
11	Giselle Velvet Dining Chairs	['accent chair', 'armless', 'Dining', '	https://st.hzcdn.com/fimgs/3	https://st.hzcdn.com/fimgs/b6b1b8190ae90000_749	https://www.houzz.com/products/giselle-velvet-dining-chairs-set-c						
12	Sienna Leather Dining Chair,	['Dining', 'frame construction', 'fra	https://st.hzcdn.com/fimgs/1	https://st.hzcdn.com/fimgs/435147405b01de6_028	https://www.houzz.com/products/sienna-leather-dining-chair-matt						
13	High Wingback Linen Armch	['accent chair', 'accent chairs tufte	https://st.hzcdn.com/fimgs/b	https://st.hzcdn.com/fimgs/25b18d4e0f27b2e9_048	https://www.houzz.com/products/high-wingback-linen-armchair-pu						
14	Divani Casa Joseph Modern	['accent chair', 'metal', 'orange', 'u	https://st.hzcdn.com/fimgs/9	https://st.hzcdn.com/fimgs/52b102b70b48767e_148	https://www.houzz.com/products/divani-casa-joseph-modern-oran						
15	Safavieh Clara Quilted Swiv	['accent chairs', 'contemporary', 'm	https://st.hzcdn.com/fimgs/2	https://st.hzcdn.com/fimgs/9b6182ad0c92a147_534	https://www.houzz.com/products/safavieh-clara-quilted-swivel-tub						
16	Booth Performance Velvet S	['chair', 'Dining', 'Home', 'home di	https://st.hzcdn.com/fimgs/ff	https://st.hzcdn.com/fimgs/5921f7290e331305_458	https://www.houzz.com/products/booth-performance-velvet-swive						
17	Montalto Pin-Tuck Lift Chair	['brown', 'saville']	https://st.hzcdn.com/fimgs/f6	https://st.hzcdn.com/fimgs/fbf1ebce079a2427_771	https://www.houzz.com/products/montalto-pin-tuck-lift-chair-savil						
18	Omax Decor Jane Lounge Ac	['arm chair', 'armchairs', 'blue', 'co	https://st.hzcdn.com/fimgs/1	https://st.hzcdn.com/fimgs/13c106380f99fc23_8349	https://www.houzz.com/products/omax-decor-jane-lounge-accent-						
19	Rosco Velvet Accent Chair, C	['charcoal chair', 'FOAM CHAIR', 'h	https://st.hzcdn.com/fimgs/2	https://st.hzcdn.com/fimgs/5051853d0f283d69_364	https://www.houzz.com/products/rosco-velvet-accent-chair-charco						
20	Lemele Velvet Dining Chairs	['dining chairs', 'dining chairs with	https://st.hzcdn.com/fimgs/1	https://st.hzcdn.com/fimgs/ede1671e0a2856ed_71	https://www.houzz.com/products/lemele-velvet-dining-chairs-set-						
21	Gianna Velvet Dining Chair,	['accent chair', 'beige', 'Best', 'blac	https://st.hzcdn.com/fimgs/fa	https://st.hzcdn.com/fimgs/bc613d340c917f7e_676	https://www.houzz.com/products/gianna-velvet-dining-chair-gray-g						
22	Amy Swivel Accent Chair, Cc	['accent chairs', 'comfortable', 'cor	https://st.hzcdn.com/fimgs/5	https://st.hzcdn.com/fimgs/42d13a110e910040_165	https://www.houzz.com/products/amy-swivel-accent-chair-coastal-						
23	Omax Decor Zola Lounge Ch	['accent chair', 'beige', 'blue', 'bro	https://st.hzcdn.com/fimgs/fe	https://st.hzcdn.com/fimgs/4bc12e8d0ba2f52d_625	https://www.houzz.com/products/omax-decor-zola-lounge-chair-br						
24	Heidi Velvet Dining Chairs, S	['accent chair', 'beige', 'Best', 'blac	https://st.hzcdn.com/fimgs/0	https://st.hzcdn.com/fimgs/eed169200ae90044_207	https://www.houzz.com/products/heidi-velvet-dining-chairs-set-of						
25	Aniline Leather Lounge Chai	['eames chair', 'eames inspired lo	https://st.hzcdn.com/fimgs/2	https://st.hzcdn.com/fimgs/659199710ef2282f_718	https://www.houzz.com/products/aniline-leather-lounge-chair-and						
26	Xavier Velvet Dining Chair, C	['black', 'chrome', 'chrome steel le	https://st.hzcdn.com/fimgs/1	https://st.hzcdn.com/fimgs/d5c117280dd7053b_287	https://www.houzz.com/products/xavier-velvet-dining-chair-gray-g						
27	Frances Leather Craft Sofa, D	['dark brown sofa', 'Leather Craft',	https://st.hzcdn.com/fimgs/2	https://st.hzcdn.com/fimgs/928112be07c720ec_789	https://www.houzz.com/products/frances-leather-craft-sofa-dark-b						
28	Nikki Velvet Dining Chairs, S	['beautiful', 'Best', 'comfortable', '	https://st.hzcdn.com/fimgs/4	https://st.hzcdn.com/fimgs/af81daab0db1c274_740	https://www.houzz.com/products/nikki-velvet-dining-chairs-set-of-						
29	Palermo White Sofa, Boucle	['3 seater', 'boucle fabric', 'boucle	https://st.hzcdn.com/fimgs/a	https://st.hzcdn.com/fimgs/1e41e46d0e4dae50_63	https://www.houzz.com/products/palermo-white-sofa-boucle-prvw						

We have 5 column: name of product, description, image1 , image2 , ULR of corresponding page of those images.

We can use these images for Machine learning and image recognition.

0.4) Robots.txt

If you want to have an access to this file you should search " WebsiteURL/Robots.txt". This file announce the limitation of all type of users (humans, search engines, robots and other type of agents)
Robots.txt for H and L are saved with names "HouzzRobots.txt" and LewisRobots.txt" in the main folder of the project.

For more information check this websites: Link1 , Link2

0.5) Scrapy vs Selenium vs Beautiful Soup

In a nut shell:
if you are dealing with complex Scraping operation that requires huge speed and with low power consumption then Scrapy **would be a great choice.**

If you're new to programmer want to work with web scraping projects then you should go for *Beautiful Soup*. you can easily learn it and able to perform the operations very quickly up to a certain level of complexity.

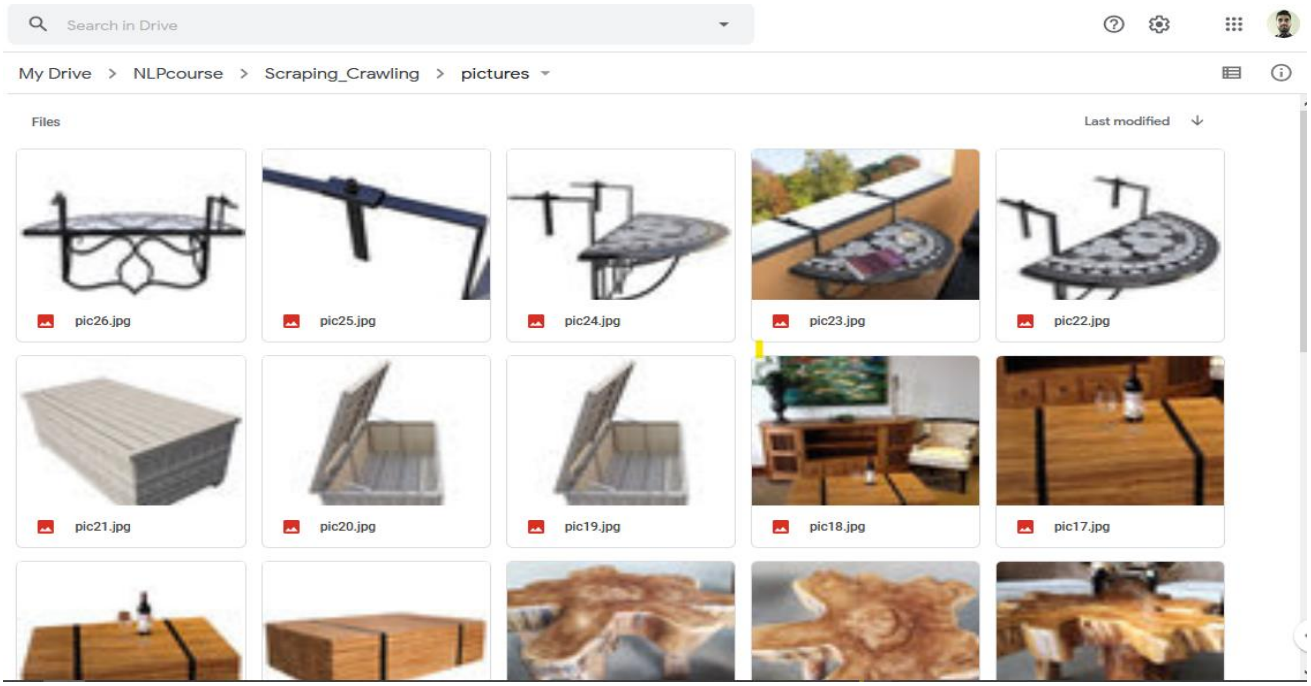
When you want to deal with Core JavaScript based web Applications and want to make browser automation with AJAX/PJAX Requests. then Selenium **would be a great choice.**

For more info, check this out

Phase 1) **Scraping Houzz**

First try: (Using bs4 for scraping Houzz website) Failed

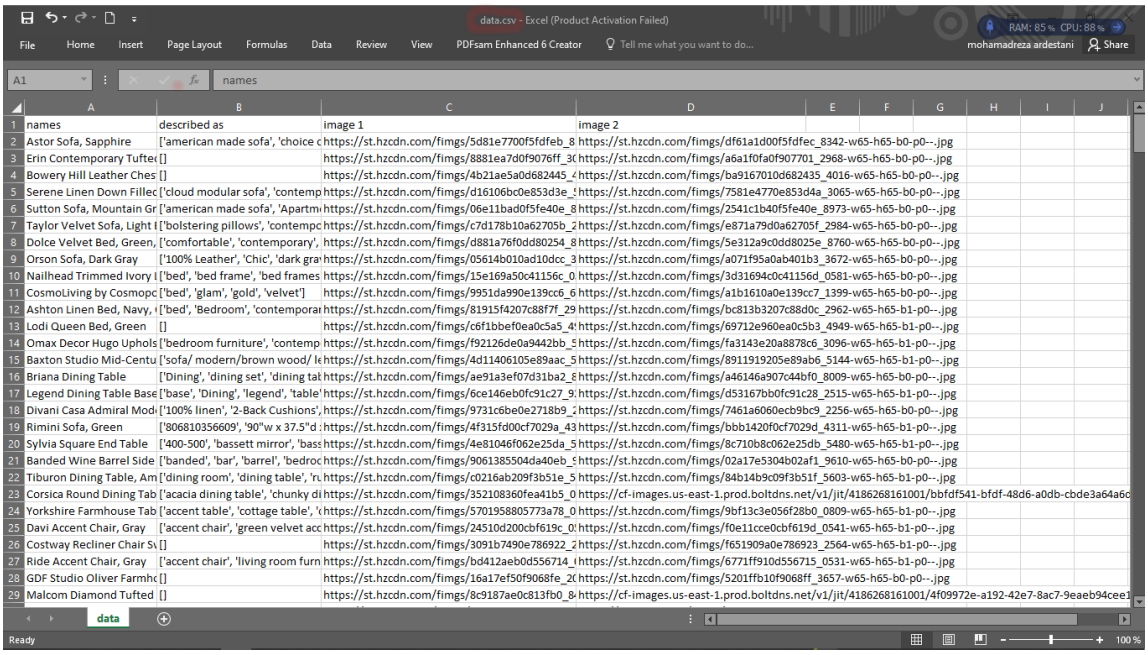
Collected images in the first try (but it's slow and we are going to second try)



Second try: (Using Scrapy) failed in runtime saving part

<u>Data.csv file is for second try.</u>

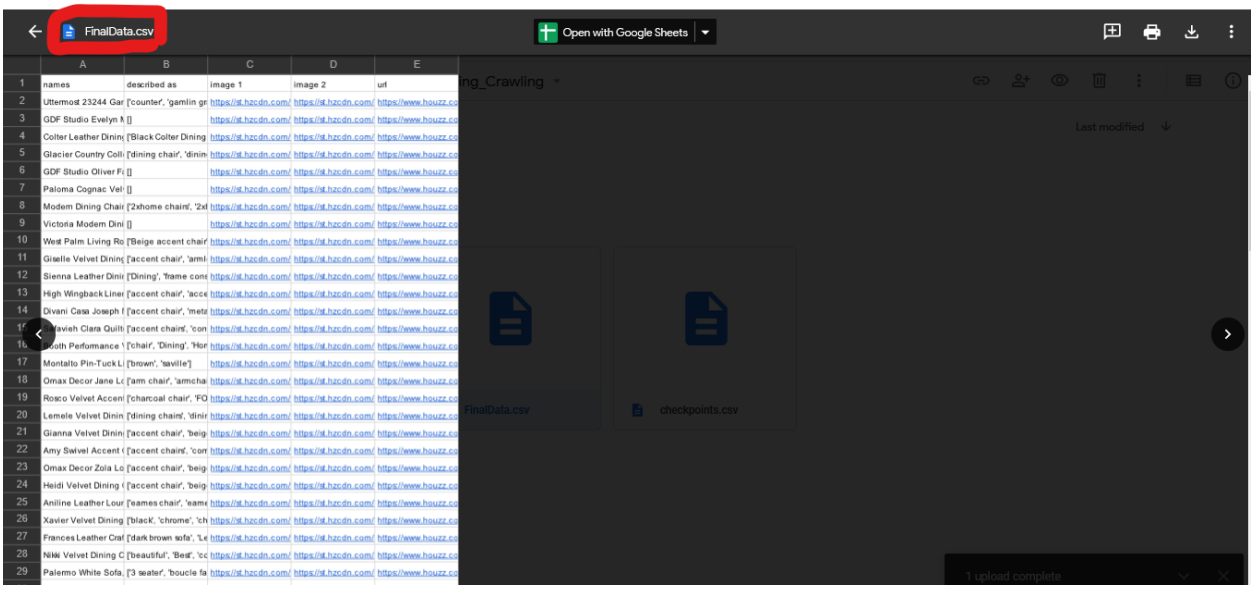
It works perfectly but the only issue is we save data at the end of scraping.



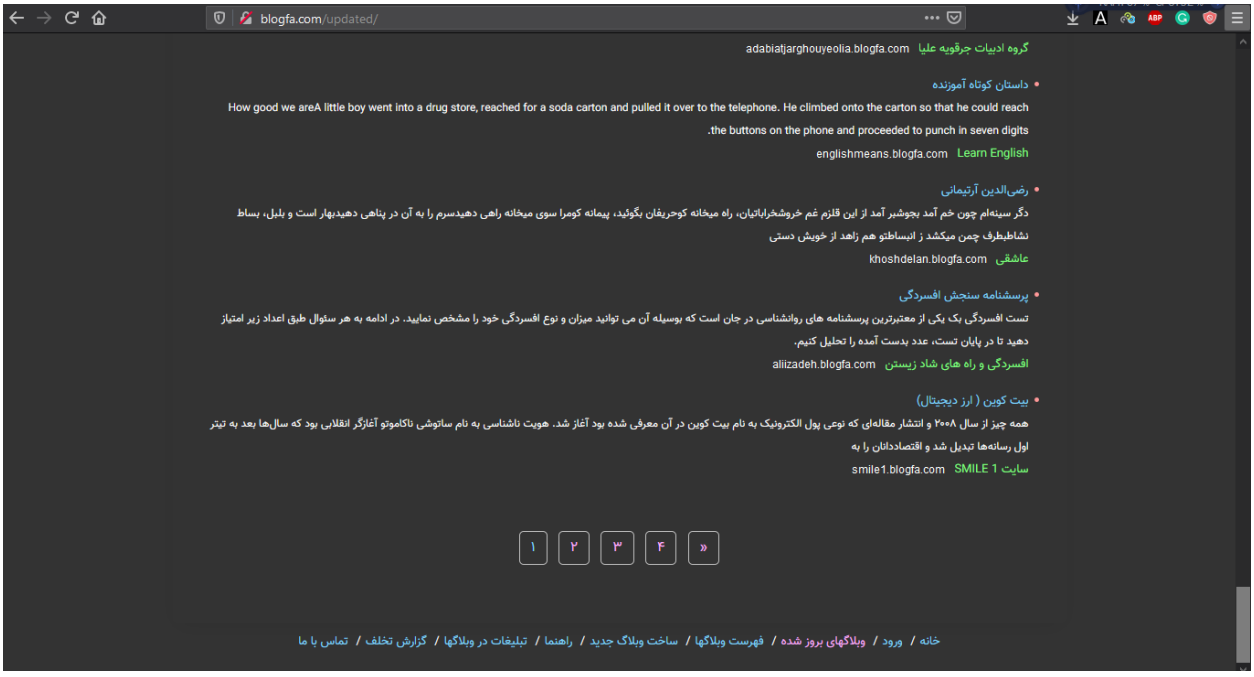
	A	B	C	D
1	names	described as	image 1	image 2
2	Astor Sofa, Sapphire	['american made sofa', 'choice c	https://st.hzcdn.com/fimgs/5d81e7700f5fdfeb_8	https://st.hzcdn.com/fimgs/df61a1d00f5fdfec_8342-w65-h65-b0-p0--.jpg
3	Erin Contemporary Tufte	[]	https://st.hzcdn.com/fimgs/8881ea7d0f9076ff_3(	https://st.hzcdn.com/fimgs/a6a1f0fa0f907701_2968-w65-h65-b0-p0--.jpg
4	Bowery Hill Leather Ches	[]	https://st.hzcdn.com/fimgs/4b21ae5a0d682445_4	https://st.hzcdn.com/fimgs/ba9167010d682435_4016-w65-h65-b0-p0--.jpg
5	Serene Linen Down Fillec	['cloud modular sofa', 'contemp	https://st.hzcdn.com/fimgs/d16106bc0e853d3e_{	https://st.hzcdn.com/fimgs/7581e4770e853d4a_3065-w65-h65-b0-p0--.jpg
6	Sutton Sofa, Mountain Gr	['american made sofa', 'Apartm	https://st.hzcdn.com/fimgs/06e11bad0f5fe40e_8	https://st.hzcdn.com/fimgs/2541c1b40f5fe40e_8973-w65-h65-b0-p0--.jpg
7	Taylor Velvet Sofa, Light I	['bolstering pillows', 'contempc	https://st.hzcdn.com/fimgs/c7d178b10a62705b_2	https://st.hzcdn.com/fimgs/e871a79d0a62705f_2984-w65-h65-b0-p0--.jpg
8	Dolce Velvet Bed, Green,	['comfortable', 'contemporary',	https://st.hzcdn.com/fimgs/d881a76f0dd80254_8	https://st.hzcdn.com/fimgs/5e312a9c0dd8025e_8760-w65-h65-b0-p0--.jpg
9	Orson Sofa, Dark Gray	['100% Leather', 'Chic', 'dark gra	https://st.hzcdn.com/fimgs/05614b010ad10dcc_3	https://st.hzcdn.com/fimgs/a071f95a0ab401b3_3672-w65-h65-b0-p0--.jpg
10	Nailhead Trimmed Ivory I	['bed', 'bed frame', 'bed frames	https://st.hzcdn.com/fimgs/15e169a50c41156c_0	https://st.hzcdn.com/fimgs/3d31694c0c41156d_0581-w65-h65-b0-p0--.jpg
11	CosmoLiving by Cosmopc	['bed', 'glam', 'gold', 'velvet']	https://st.hzcdn.com/fimgs/9951da990e139cc6_6	https://st.hzcdn.com/fimgs/a1b1610a0e139cc7_1399-w65-h65-b0-p0--.jpg
12	Ashton Linen Bed, Navy, (	['bed', 'Bedroom', 'contemporar	https://st.hzcdn.com/fimgs/81915f4207c88f7f_29	https://st.hzcdn.com/fimgs/bc813b3207c88d0c_2962-w65-h65-b1-p0--.jpg
13	Lodi Queen Bed, Green	[]	https://st.hzcdn.com/fimgs/c6f1bbef0ea0c5a5_4(	https://st.hzcdn.com/fimgs/69712e960ea0c5b3_4949-w65-h65-b1-p0--.jpg
14	Omax Decor Hugo Uphols	['bedroom furniture', 'contemp	https://st.hzcdn.com/fimgs/f92126de0a9442bb_5	https://st.hzcdn.com/fimgs/fa3143e20a8878c6_3096-w65-h65-b1-p0--.jpg
15	Baxton Studio Mid-Centu	['sofa/ modern/brown wood/ le	https://st.hzcdn.com/fimgs/4d11406105e89aac_5	https://st.hzcdn.com/fimgs/8911919205e89ab6_5144-w65-h65-b1-p0--.jpg
16	Briana Dining Table	['Dining', 'dining set', 'dining tal	https://st.hzcdn.com/fimgs/ae91a3ef07d31ba2_8	https://st.hzcdn.com/fimgs/a46146a907c44bf0_8009-w65-h65-b0-p0--.jpg
17	Legend Dining Table Base	['base', 'Dining', 'legend', 'table'	https://st.hzcdn.com/fimgs/6ce146eb0fc91c27_9	https://st.hzcdn.com/fimgs/d53167bb0fc91c28_2515-w65-h65-b1-p0--.jpg
18	Divani Casa Admiral Mod	['100% linen', '2-Back Cushions',	https://st.hzcdn.com/fimgs/9731c6be0e2718b9_2	https://st.hzcdn.com/fimgs/7461a6060ecb9bc9_2256-w65-h65-b0-p0--.jpg
19	Rimini Sofa, Green	['806810356609', '90"w x 37.5"d	https://st.hzcdn.com/fimgs/4f315fd00cf7029a_43	https://st.hzcdn.com/fimgs/bbb1420f0cf7029d_4311-w65-h65-b1-p0--.jpg
20	Sylvia Square End Table	['400-500', 'bassett mirror', 'bass	https://st.hzcdn.com/fimgs/4e81046f062e25da_5	https://st.hzcdn.com/fimgs/8c710b8c062e25db_5480-w65-h65-b1-p0--.jpg
21	Banded Wine Barrel Side	['banded', 'bar', 'barrel', 'bedroc	https://st.hzcdn.com/fimgs/9061385504da40eb_{	https://st.hzcdn.com/fimgs/02a17e5304b02af1_9610-w65-h65-b0-p0--.jpg
22	Tiburon Dining Table, Am	['dining room', 'dining table', 'ru	https://st.hzcdn.com/fimgs/c0216ab209f3b51e_5	https://st.hzcdn.com/fimgs/84b14b9c09f3b51f_5603-w65-h65-b1-p0--.jpg
23	Corsica Round Dining Tab	['acacia dining table', 'chunky di	https://st.hzcdn.com/fimgs/352108360fea41b5_0	https://cf-images.us-east-1.prod.boltdns.net/v1/jit/4186268161001/bbfdf541-bfdf-48d6-a0db-cbde3a64a6c
24	Yorkshire Farmhouse Tab	['accent table', 'cottage table', '	https://st.hzcdn.com/fimgs/570195880577 3a78_0	https://st.hzcdn.com/fimgs/9bf13c3e056f28b0_0809-w65-h65-b1-p0--.jpg
25	Davi Accent Chair, Gray	['accent chair', 'green velvet acc	https://st.hzcdn.com/fimgs/24510d200cbf619c_0	https://st.hzcdn.com/fimgs/f0e11cce0cbf619d_0541-w65-h65-b1-p0--.jpg
26	Costway Recliner Chair S	[]	https://st.hzcdn.com/fimgs/3091b7490e786922_2	https://st.hzcdn.com/fimgs/f651909a0e786923_2564-w65-h65-b1-p0--.jpg
27	Ride Accent Chair, Gray	['accent chair', 'living room furn	https://st.hzcdn.com/fimgs/bd412aeb0d556714_(	https://st.hzcdn.com/fimgs/6771ff910d556715_0531-w65-h65-b1-p0--.jpg
28	GDF Studio Oliver Farmhc	[]	https://st.hzcdn.com/fimgs/16a17ef50f9068fe_2(	https://st.hzcdn.com/fimgs/5201ffb10f9068ff_3657-w65-h65-b0-p0--.jpg
29	Malcom Diamond Tufted	[]	https://st.hzcdn.com/fimgs/8c9187ae0c813fb0_8	https://cf-images.us-east-1.prod.boltdns.net/v1/jit/4186268161001/4f09972e-a192-42e7-8ac7-9eaeb94cee1

third try: (Using Scraping and additional functions) Successful FinalData.csv is for the third try

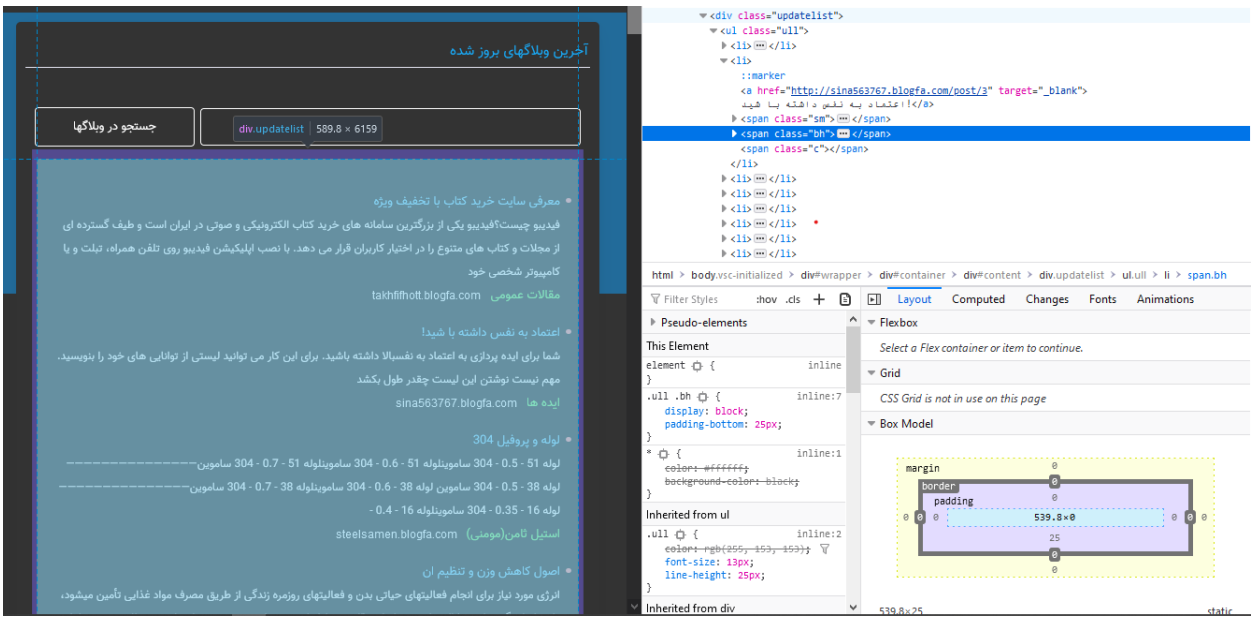
In this try we store the data at running time ☺ .



Phase 2) **Blogfa.com**



Blogfa shows recent updated blogs in /undated/ directory. So we use this link. Also blog fa only has 5 pages for showing these result.
http://blogfa.com/updated/?p=2 for instance this is page number two.



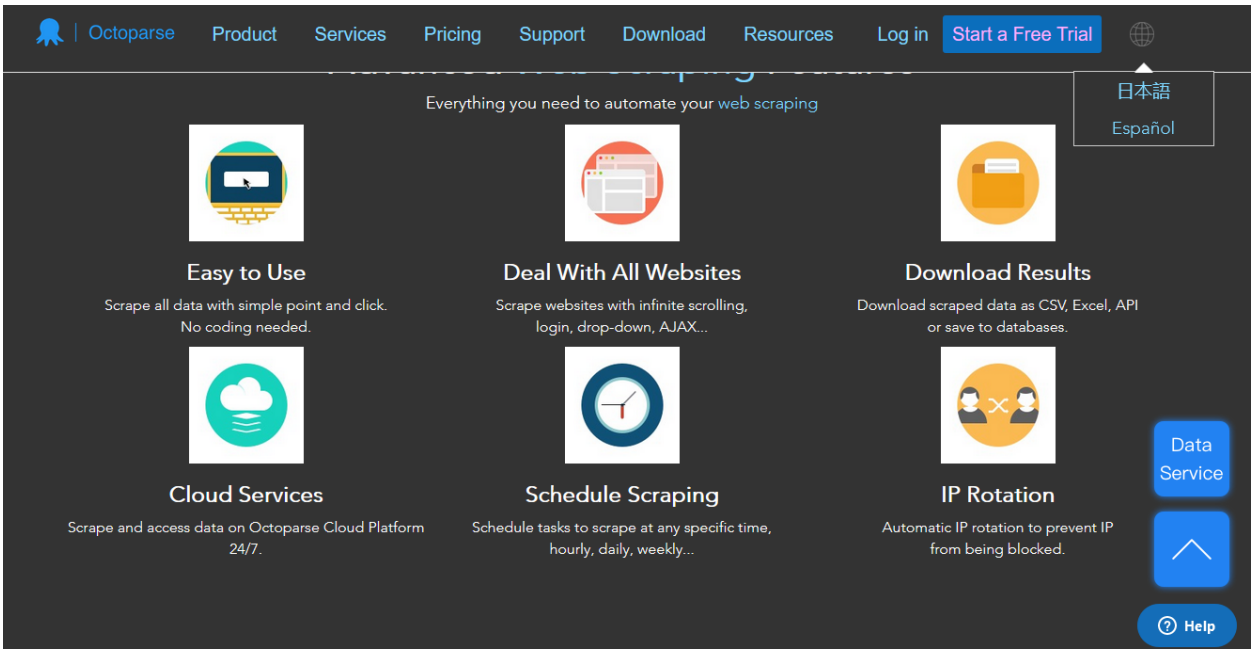
By manually using website's interface source code I figured out how I can fetch URL of all blogs in each page.

Phase 3) **IP rotation and personal data clearing**
My reference for this part is this link. They have done a great job in this era. I've thoroughly explained this part in the code file.

Phase 4) **A little creativity**
+ Scraping Movie websites + BeautifulSoup + Pickle lib for pickling data
I went through one website and I Scraped transcripts of some American stand-up comedians and I used pickle library for storing data in an efficient way.
The purpose of this part is analyzing why they are funny and how much their works is based on making and using Paradox. (And furthermore , why some our mind likes comedy)

<< More detail in the code file .>>

+ Implementing Scraping & Crawling with Octoparse toolkit



This website is highly reliable for scraping and crawling. In this framework we can set parameters without unnecessary coding for gathering data.

Sincerely yours and Thanks for your time and attentions.