This AI Scraper Update Changes EVERYTHING!!

https://www.youtube.com/watch?v=-IAwW3pUEew

all right so you have asked about it so much to the point that I couldn't ignore it anymore I have added the pagination feature which is going to work hand in hand with the scraping for example if I want to scrape data from indeed for web developer all I need to do as always is get the URL here let's get the URL for web developer just copy it and then let's go back to our Universal web scraper let's put the URL in here let's choose the model let's say for example we're going to choose Gemini 1. flash to make everything for free let's enable the scraping now we can enable or disable the scraping and the pagination independently and here we are going to define the fields so for the fields let's say for example we're going to start by the job title then the company the location the salary the contract type you probably notice that this website is in French it can work on any language if this is English it's going to be the same thing your titles you can Define them in any language and the website can be in any language as well and lastly we are going to scrape the description and here we are going to also enable the pagination and we are going to leave this pagination text box empty because it is optional it can work without it so let's click on scrip and let's see what's going to happen so it is going to open Indeed as always it is going to start scraping and then from there it is basically going to give us the results here so let's put this in here and let's wait for the result and as you can see here we have our results these are basically all the columns we asked for and here even if if you are working remote which is St in French it is still able to detect it and then we have the salary in here so everything should be correct and when the salary does not exist it is not and as you can see here we have also the pagination so let's say for example if we are going to click on the first row we are going to probably have the second page because it started yeah it started from the second page we have already scraped the first page so it started from the second page if we click on let's say for example page this should be page exactly this is Page so as you can see here it has been able to get all the pages and also it has been able to get the data as well and if you notice here even though the the names of the pages does not correlate with this start argument that we have here that it has been able to detect even though that's not the case it has been able to detect the differ pages and now you're going to tell me okay you have been able ble to scrape the page and now we have multiple pages that we can work with can I scrape multiple Pages at the same time and the answer is yes we are going to be able to do so let's select for example these three pages let's copy them contrl C let's come here let's paste them and before doing that by the way we are going to come back and talk about it this is basically the pagination details how much it has spent and this is the total cost for the pagination and the scraping at the same time but don't worry we're going to come back to that and we are going to talk about that in details but here I have basically three URLs and I am going to come back here and click on scrape no need to disable the pagination when we have multiple URLs at the same time the pagination get disabled by default it is basically opening the websites and it's scraping the data and then we should have three different tables in here reflecting the three different URLs that we want to scrape and as you can see here it has been able to scrape the three different pages and if you for example

want this data in one table all you need to do is download CSV and then you are going to open this CSV and you're going to find all of your data already saved in here from the three different tables not just from one all right so before going on with the video and scraping from other websites we need to talk about the different things that I need to share with you in order for you to be able to use this scraper to the best of its abilities so the first thing that we need to talk about is where you can find the code and how you can reproduce this on your own machine I've done a similar thing in my last video but since this is a different code and since GitHub suspended me and have reinstated me now I don't really know if I trust them and believe me guys it's really hard to have some kind of version control without GitHub cuz I'm not used to gitlab or codg or any other platform so I need to share the code with you guys in a place that I am sure I can control therefore my website is the best place and you are going to have a separate page for this scraper and I'm going to share the link with you in the description below that's going to be the first thing that we are going to talk about about second thing we are going to talk about is how you can use the new pagination mode because you have to understand what I have done in my code the different clues that I am using in order to get the pagination so even if this scraper is not able to get the pagination you know exactly what to write inside of this prompt in case the scraper is not able to get you the pagination that you want then we are going to talk about the new feature of being able to scrape from multiple websites how does the scraper basically split between the URL and get the exact URLs to scrape later on and of course I am going to show you where does it store the data because I have added a new module that basically stores the data according to the name of the URL this way even if for example it fails at the last scraping you are still going to find the data for the first and the second URL that it has been able to scrape the last thing that we are going to talk about is of course the limitations of this Universal web scraper because there are some websites which was very hard to scrape where going to talk about them and then we are going to talk about which model performed the best in which case that's also something very important so according to your own data according to your own website which model should you go for and if you're ready to pay for your scraper or you want to go with a free approach so with that being said let's start the first part of where you can find the code so as I told you guys GitHub did reach back to me about the reinstatement request for my account and they basically said that they have suspected that there was someone has access to my account that's why they basically suspended the whole account and now they have partially lifted the Restriction but it is still restricted to a certain degree so if I share this project again on GitHub I don't know what's going to happen therefore the only insured way of sharing this project right now is going to be through this website and here you will find scrip Master 3.

and inside of here you are going to find the code I still haven't updated it but the moment I will share this video you will find everything updated with the video in here so that's where you can basically copy paste the code in my last video I have told you exactly how to configure if you're not a developer yourself you only need a vs code a python configured inside and then a couple of files that you need to create and then you need to copy paste the code and you need to create a folder inside of the code that is called output that's all you need to do and you can basically run this on your machine using the command streamlit run streamlit app.

py if you still don't know how to do it just go back to my last video there I basically have followed step by step just try to apply the same thing and you should be good to go now let's

go back to the most important part which is the pagination and how does it work behind the scenes so here if you go to my code you will see that I have created a new module pagination detector and inside of this module I have a couple functions the most important about all of them is the detect the pagination elements you will find my prompt in order to get the pagination from any URL so the first thing that the universal web scraper look for is if you're your url already have some kind of indication about the pages it will be so much easier for it to detect the pagination from the page so let's say for example I am in eBay and I have this URL as you can see here this URL does not really have any type of page equal or page equal inside of it meaning that you basically have to understand the structure in order for you to be able to do it easily now you can absolutely give this to the web scraper and it's still probably going to work but if you want to guarantee that it is absolutely going to work just go to the second page inside of here and you are going to see that we have a new query parameter or URL parameter that we basically give to the page in order to indicate that we want to go to the second page for our search for Samsung Smart TVs so if you give this URL you have a higher chance of getting the right pagination that's one the second thing if there is no pagination indication inside of here what it does is that it will search inside of here to try to find a pattern of the URLs so here it will try to find the pattern for the URLs if it does find the pattern of the URLs it will be so easy to do and that would be just perfect but if it doesn't for example like here we have eBay page three then we are going to have eBay page Etc the problem is if it doesn't find these URLs already in the page like for example inside of indeed here let's inspect the element you're going to see that there's not a valid URL so the scraper have to understand how to intelligently create the URL so we have been able to get the passation from here even without any kind of indication but if we have to give it indication in the case of indeed we would say that we only have partial URLs please complete the URL from what I just give you in the field URL so this way you will be sure that it is going to give you the whole URL now in a different use case let's say for example we are scraping data out of scrap me.

live here we can see that we have 2 4 and it's going to 4748 if we go and inspect the data we will find that we only have URLs that span from to and from to if we go here we will find page four but if we go here it's going to be page number dots and then here we are going to find page so the universal web scraper have to understand that this page number dots means that there are so many pages that exist but are not mentioned on this page so it has to complete all the pages between and and if we go back here to the scraper let's disable the scraping we only want to do the pagination and let's give it this URL copy it and paste it inside of here and click on scrape so here it should give us a table with all the pages not just the pages that it has found in the HTML and as you can see here it added all the pages between and all of these are our pages and in this case it is worth to mention that sometimes with Gemini flash especially with Gemini flash it just basically just give us even more results that we don't even ask for so here we have more than we have pages that do not exist if I click on this this does not exist and I only notice this with Gemini flash if for example I choose GPT for minu and I click on scrape so GPT mini will give me page one and also will give me all the pages until page it will not give me more pages so it is more correct than Gemini flash but the difference is Gemini flash is actually for free so if you want to use Gemini flash you can just s please stop at page number cuz you know that the last page and that indication will hopefully stop it at page so these are the cases that I have basically dealt with but there are cases where you can't find any queer URL and there are no page URLs inside of the website in this

case we basically cannot find any indication and probably the universal web scraper will fail because there's no query in here and there are no patterns of pagination inside the page okay so now that we have talked about pagination let's go and talk about scraping mode for multiple URLs at the same time so to be able to scrape multiple URLs I thought about having multiple separators that we can have here either a semicolon or a comma or any other separator but the problem is we can still have a comma inside of a URL so for really complex URLs you can probably have a comma it's not really recommended but you can still have it same thing for semicolons the thing that we know is forbidden to have in a URL is a space so I think this is the perfect separator if you want to scrape data from two URLs for example all you need to do is get the first URL then copy a URL space put it in here and then this should interpret these two as two different URLs that it would scrape the data from now let's actually scrape the data from these two URLs just so that I can explain something you know what let's actually use a good example let's actually use eBay for example let's go with this URL this is eBay and we want to scrape data for adjustable height desks and let's see if it's going to be able to scrape data out of this website let's basically go to the first page here we're going to get this first URL we are going to put it in here and then let's go to the second URL and also get it and put it right after the first URL of course it should be separated by a space this enable scraping and all what we want to scrape in this case is going to be the name and the price so let's actually launch the scrap in and as you can see here it basically gave me the results there are a lot of details that I actually want to talk about but let's start with the fact that why did I opt for basically two different tables instead of combining them well the first reason is to see how much data you got from the first URL and then how much data you got from the second URL so here clearly you can see that from the first URL you only got Ros but the second URL we got so much more data so there's already a problem with the scraping the second reason is when we want to scrape data from two different websites that have nothing to do with each other and we still want to scrape the same data for example the name and the price from eBay and let's say Amazon we can still actually scrape from two different websites and have the two tables basically independent and if we want to download each table we can just come here and choose to download that specific table as a CSV if you want to download everything combined we can just go here and download CSV so that is why I have divided both tables now let's go back and talk about why did we have only rows when we know for sure that we have so much more products in here I think there is a better job to be done on the promp in maybe there's also some agentic workflow that we need to introduce in order to create bque the results now what I have noticed is that Gemini flash is better at extracting the maximum amount of results from the data that we have so here let's choose Gemini Flash and let's launch the same scraping and let's see what's going to happen so as you can see here Gemini flash usually gives you all the results so that is the best thing about Gemini flash honestly it is the best at extracting the maximum amount of results so this is one of the tips and tricks that I actually found out while using these applications so here we have 57,000 tokens in the input 3,890 tokens and even if we paid we did not pay anything because Gemini flesh is free for the calls per minute 1,500 call per day therefore this is actually for free but if you go over the 126,000 free tokens I have noticed that Gemini Flash starts giving problems so as long as you're not at 127,000 tokens your scraping will be correct okay that is very good good there is one last thing I need to talk about in here which is if you have been paying attention this alert that I have here scraping completed results saved in output eBay and then

the date of when we started the scraping so here if I go to my project and then I go to Output you will see that it will basically create a folder per scraping session meaning that if you're scraping one website you will only have one scraping if you scrape three URLs inside of it you will have the scraping of three URLs and inside of these folders for example the latest one that we have just done you are going to find the row data for the first URL the row data for the second URL the Json for the first and second URL and our Excel sheets for the first and second URL so you are going to find the whole entire data and if for some reason your scraping fails you will still going to find the row data and you're still going to find the Json and the Excel sheets for the ones that have not failed so this is a way to fall back and it's very good because it gives you exactly the name of the URL that you have extracted inside of here no matter how that URL is going to change this folder will be always named after the URL so you don't have to worry about that and there's no action from your side all right so we have seen the pagination we have seen the scraping multiple URLs the tips and tricks that I have been using now let's talk about the limitations so there are some websites where whatever you do once you open the websites for the first time and you're not signed in or anything it will basically block you for example u.

com whatever you do if you open this website for the first time it will still give you a capture so unless you find a way to do this capture it is not going to give you access so to be able to circumvent this and because this is not a scraper at scale this is a productivity tool you will need to create an account inside of that website and once you create the account you will easily get inside the website without having to verify any type of capture the second limitation is with the websites that simply have a number of tokens inside of their websites that is basically just ridiculous an example examp if this is Aliexpress if you go and choose whatever if you choose hair dryer Kettle whatever it is going to be an incredible amount of tokens in here it is simply out of this word I think it was 200,000 tokens so this amount of data is so hard to be processed let's take for example this URL let's copy it let's put it inside of our Universal web scraper let's choose a model like GPT mini let's keep it at name and price if we click on scrape it's it's going to start scraping and I suspect that it's going to find an error so here we have an error and if we go back here of course this is saying scraping completed results saved because the row data has been saved but if we go back to the back end you will see that this model maximum context length is 128,000 tokens however your message is 234,000 tokens it's an absolutely crazy number I think Amazon is 60,000 tokens maximum 880,000 tokens other websites like e I think it's at 30,000 tokens but AliExpress is just added this word the amount of data that you have here is just incredible and I would say that even a traditional scraper would find a harder time scraping data out of here but it's still going to work better because you basically know exactly where you're going but again the trade-off is that it's a not Universal web scraper you cannot just expect it to work on any other website so as I said in my last videos this is basically a group project I am just taking feedback from you guys there's one other feature that has been really asked for which is Docker and I think I will be adding that next but other than that if you guys want to see other features added to this Universal web scraper just tell me in the comments the comments that have the most interaction would be of course prioritized much like I have done in the past with that being said thank you guys so much for watching don't forget to like And subscribe it does really make a difference it does help and see you guys next time peace