Notes : June 25, 2021

Project : **[Download files, Extract Data and Store data](https://www.freelancer.com/projects/web-scraping/Download-files-Extract-Data-Store" \t "_self)**

6/25/2021 Project Status

Project status I think is Green. Have successfully extracted the latest submissions for all tickers requested. I have processed all the 10K’s and 10Q’s for the tickers you provided back to 2011.

I packaged up the code and data and will send in a separate message.

I wrote up some documentation on how to use the Python scripts.

I also found that the amended filings on the SEC website don’t seem to have a fully populated set of data for the period they are restated. Would like to get your opinion as to deal with that.

Project Issues : / Questions

| # | Question | Contact | Date Of Question | Responses | Status |
| --- | --- | --- | --- | --- | --- |
| 1 | The NLY iXBRL file contains 2,144 Facts, your CSV file has 2,071. Any idea where the difference is coming from? | Brian B | 6/16/2021 | Jeff Jones : I filtered some of the text fields out, as I thought you were focusing on leverage calculations-. I have removed those filters and the Facts number in the XBRL viewer files now matches the CSV files. | Pending closure – waiting on comments from Brian.  6/25 – No further issues reported. Closing this. |
| 2 | In some of the files they have a format field to designate $millions, can that added be added as column? For example in the NLY file, CashAndCashEquivalentsAtCarryingValue is 222 but shows up in the text as 22.2. | Brian B | 6/16/2021 | Jeff Jones 6/16 : The CashAndCashEquivalentsAtCarryingValue is stated as 22.2 in a footnote where it is described as being in millions. This is an example of where I think there might be some issues with the XBRL viewer or perhaps I am missing something. The value in the XML file is 22200000 and the ‘decimals’ field is -5. This would result in a value of 222 (which is what is in the produced CSV). However, in the footnote on the report it states CashAndCashEquivalentsAtCarryingValue 22.2 as being in millions. The ‘decimals’ value in the XML file should be -6. | Pending closure – waiting on comments from Brian. |
|  |  |  | 6/17/2021 | Jeff Jones 6/17 : This is resolved. I read through the SEC XBRL specification and found that the ‘decimals’ attribute only defines the level of rounding a number has been subjected to. It is not supposed to be used to calculate the actual value. If the ‘decimals’ attribute is ‘INF’, then the number has not been rounded. As most numbers in a 10-Q have ‘decimals’ attribute between -6 through +6, the rounding will result in totals not footing exactly for things like Total Assets, etc. I added the level of rounding for a number to the CSV extract to indicate whether a number is exact (‘decimals’=’INF’) or rounded(‘decimals=some positive or negative number). In the case that Brian reported, the CashAndCashEquivalentsAtCarryingValue field has a value of 22200000, which is $22,200,000. This is the correct value, but it has been rounded to the nearest 100,000 (ie. ‘decimals’=-5). This means that CashAndCashEquivalentsAtCarryingValue could have an actual value of 22,150,000 – 22,249,999. | 6/25 – No further questions on this topic. Closing this issue. |
| 3 | Do you mind adding the ticker as a column? | Brian B | 6/16/2021 | Jeff Jones 6/16: This is complete. It is the first column in each file. | Pending closure – waiting on comments from Brian.  6/25 – No further questions on this topic. This is being closed. |
| 4 | As I start merging some datasets it will make my life easier. In the instance where we go back a quarter, how would you handle not restoring a lot of the same values? | Brian B | 6/16/2021 | Jeff Jones 6/16 : If you could explain what you are referring to, that would help. If a field is published on a report on a certain date, the report date and the ‘as of’ date of the field would be used as eys to store it (and subsequently retrieve it. So, if the CashAndCashEquivalentsAtCarryingValue for 12/31/2020 is published on the 12/31/2020 and the 3/31/2021 10Q’s, they are stored as separate pieces of data (so you can see if they have restated it). If that isn’t what you need, happy to change it (right now, it is just writing to the CSV file so no change really required. | Waiting on comments from Brian.  6/25 – No further questions or comments on this topic. Closing this for now. |
| 5 | In the AAIC file, I see a few values of “Suppressed”. Any idea what that means? | Brian B | 6/16/2021 | Jeff Jones : There are some text fields which include a lot of additional formatting information like fonts and colors. I took the liberty of suppressing them as they have special characters which are not part of the standard ASCII character set (so you can’t print them or view them in a CSV file). We should discuss if you need those. | Waiting on comments from Brian as to whether suppressing these was okay.  6/25 – I didn’t get any further questions or comments, so I am closing it for now. We can circle back if required. |
| 6 | When going back to prior quarters, will the supplemental Schema and presentation/calculation files also need to be downloaded? | Brian B | 6/16/2021 | Jeff Jones 6/16 : I did not download them as I thought everything you needed was in the main XBRL instance (xml file). However, if the schema’s changed drastically in some of the older 10Q’s, I might have to. | Waiting on comments from Brian as to whether he thinks we need those additional files.  6/25 – I did not see any requirement for the additional files. I am closing this for now, unless Brian sees a need (which can be addressed as required). |
| 7 | Looking at some of these tickers, some of them didn’t file the full XML package going back to 2010. If you ran one of the old .htm files that have the xml tags, would it handle it? | Brian B | 6/16/2021 | Jeff Jones 6/16 : I can try that tomorrow. I will let you know. | 6/17 The XML parsing code did not process the HTML (non-BRL) files. Parsing HTML will require another project. Another option is to investigate other data sources other than the SEC.  6/25 – The HTML data points really do not have any field definitions or tags that would be in common with the data coming from the XML files. If there are a small number of data points needed from an HTML page, I can extract them but it would require some additional costs. |
| 8 | If not, I’m fine going back as far as the XML packages are provided. As a next step, can we grab one particular company and pull the historical Qs and scrape them? | Brian B | 6/16/2021 | Jeff Jones 6/17 : I ran NLY 10k’s and 10q’s back to 12/31/2019 (that’s as far as XML was available for them). I also ran AGNC back to 3/31/2019 (again, that as far back as their XML was available).  The files processed successfully. The Fact Counts for each file matched that published by the SEC on their website. | Waiting on comments / feedback.  6/25 – I was able to get XML files back to 2011 for almost all the companies on this list. I am closing this for now, unless there is a need to get specific files. |
| 9 | One small thing, i think the filed called 20200930\_NLY... shows a report\_date of 20190930 | Brian B | 6/21/2021 |  | 6/25 – This was resolved when I downloaded all the files again from the SEC. |
| 10 | I was looking at some of the stockholders equity breakouts, for instance the row of data that is NLY 20201231 20201231 us-gaap:AdditionalPaidInCapitalMember StockRepurchasedAndRetiredDuringPeriodValue 209094000  is there anything in the files that would be able to identify that as "Buyback of common stock" ? | Brian B | 6/21/2021 | Regarding your question on share buybacks, there is a schedule of changes to equity. The company AAIC has one in their 3/31/21 10Q. There are tags which identify common share buybacks.  us-gaap:CommonClassAMember","StockRepurchaseProgramRem ainingNumberOfSharesAuthorizedToBeRepurchased"  I see that for NLY, the board approved up to 1.5b common share buy back. There isn't anything where that is quantified in any schedule with a tag on it. | 6/25 – Brian indicated he found what he needed on this when he had a fresh set of eyes on the data the next morning.  I am considering this closed. |
| 11 | For now, can you make the output for each company go into a single file? | Brian B | 6/21/2021 |  | 6/25 – The last set of data sent to Brian had all data combined for each ticker. |

**Requirements Section (In progress):**

The tool needs to download files from SEC, extract data and store the data in a format that will be to analyze data over time series. This will be for ~20 companies (NLY, AGNC, DX...) and each of their 10-Q reports going back until 2010. The files are in a special reporting language called XBRL which is displayed as .XML. There are various references on YouTube and GitHub keyword (SEC Edgar or XBRL). The structure of the data can be tricky as each company reports differently and some companies have changed their reporting.

Attached:

A sample file that can be downloaded that contains the .XML.

The screen grab comes from this link [login to view URL]

Example of the sort of what i'm trying to do is analyze the changes of the of the Estimated Fair Value column over time for each row of data.

Please review the attached data (another company example can be provided) and read/research the topic a little before providing your quote. In your quote tell me something particular about the data/project that will confirm you are qualified. \*\*Bonus points if you think you can the Non-XBRL tables extracted.

Price is flexible if you can proof that you have firm grasp on what needs to be done and the expected output.