Web Scraping using LinkedIn - Steps, solving issues and customization

To implement **main_scraper.py**: This program scrapes the data from linkedin and write the results to a CSV file, namely results.py

1. Refer https://pypi.org/project/linkedin-jobs-scraper/
2. `pip3 install linkedin-jobs-scraper`
3. pip3 install requests
4. From [Downloads - ChromeDriver - WebDriver for Chrome](), install chromedriver
5. Change path in python file from `chrome_executable_path=None,` to `chrome_executable_path= “/wherever the exec file for chromedriver is located”`
6. Received error scraper:('Scraper failed to run in anonymous mode, authentication may be necessary for this environment. Please check the documentation on how to use an authenticated session.',) NoneType: None
7. Tried adding a proxy, but same problem repeated along with WARNING:li:scraper:(ProxyError(MaxRetryError("HTTPSConnectionPool(host='static-exp1.licdn.com', port=443): Max retries exceeded with url: /sc/h/d2cr7f7e79esv9eh2bs12t806 (Caused by ProxyError('Cannot connect to proxy.', NewConnectionError('<urllib3.connection.HTTPSConnection object at 0x7faa3d8551d0>: Failed to establish a new connection: [Errno 61] Connection refused')))")),)
8. Solution to the above problem is to give authentication. We need to remove proxies as proxies don't work with authentication. Implement authentication in following steps
   8.1. Login to LinkedIn using an account of your choice.
   8.2. Open Chrome developer tools:
   8.3. Go to tab `Application`, then from the left panel select `Storage` -> `Cookies` -> `https://www.linkedin.com`. In the main view locate the row with name `li_at` and copy content from the column `Value`.
   8.4. Set the environment variable `LI_AT_COOKIE` with the value obtained in step 3, then run your application as normal in terminal. Example:

```
LI_AT_COOKIE=<your li_at cookie value here> python
   your_app.py
```

9. You may experience the following rate limiting warning during execution: `[429] Too many requests. You should probably increase scraper "slow_mo" value or reduce concurrency.`

This means you are exceeding the number of requests per second allowed by the server (this is especially true when using authenticated sessions where the rate limits are much more strict). You can overcome this by:

9.1 Trying a higher value for `slow_mo` parameter (this will slow down scraper execution).

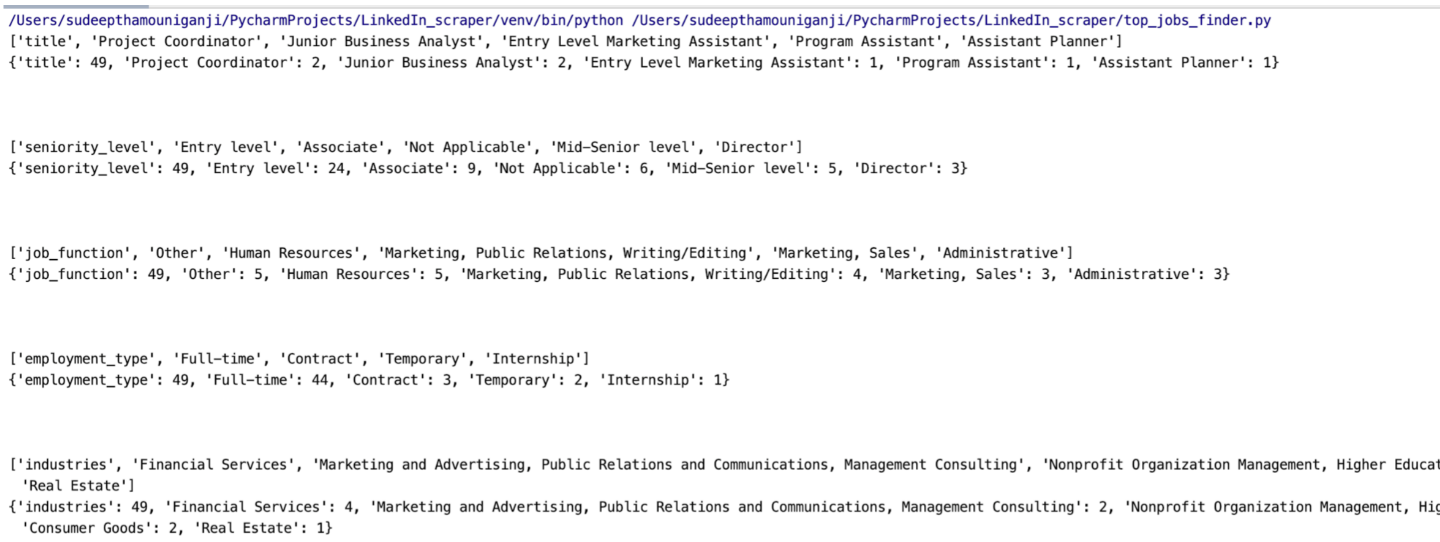
9.2 Reducing the value of `max_workers` to limit concurrency. I recommend to use no more than one worker in authenticated mode.

10. Make following changes in python file to write data to csv file.

```
def on_data(data: EventData):
  # print('[ON_DATA]', data.title, data.company, data.date, data.link, len(data.description))
  # print(data)
  with open('results.csv', mode='a+') as csv_file:
      fieldnames = ['job_id', 'link', 'apply_link', 'title', 'company', 'place', 'description', 'date',
'seniority_level', 'job_function','employment_type', 'industries']
      writer = csv.DictWriter(csv_file, fieldnames=fieldnames)
      #
      # writer.writeheader()
      writer.writerow({'job_id':data.job_id, 'link':data.link, 'apply_link':data.apply_link, 'title':data.title,
'company':data.company, 'place':data.place, 'description':data.description, 'date':data.date,
'seniority_level':data.seniority_level,
'job_function':data.job_function,'employment_type':data.employment_type, 'industries':data.industries})
```

Results_analyser.py:

This file reads the results from the results.csv file and analyses them to get useful insights like total count of different items based on different fields like job title, seniority level, job_function, employment_type, industries. It also finds the top 5 items in the different fields.



```
/Users/sudeepthamouniganji/PycharmProjects/LinkedIn_scraper/venv/bin/python /Users/sudeepthamouniganji/PycharmProjects/LinkedIn_scraper/top_jobs_finder.py
['title', 'Project Coordinator', 'Junior Business Analyst', 'Entry Level Marketing Assistant', 'Program Assistant', 'Assistant Planner']
{'title': 49, 'Project Coordinator': 2, 'Junior Business Analyst': 2, 'Entry Level Marketing Assistant': 1, 'Program Assistant': 1, 'Assistant Planner': 1}



['seniority_level', 'Entry level', 'Associate', 'Not Applicable', 'Mid-Senior level', 'Director']
{'seniority_level': 49, 'Entry level': 24, 'Associate': 9, 'Not Applicable': 6, 'Mid-Senior level': 5, 'Director': 3}



['job_function', 'Other', 'Human Resources', 'Marketing, Public Relations, Writing/Editing', 'Marketing, Sales', 'Administrative']
{'job_function': 49, 'Other': 5, 'Human Resources': 5, 'Marketing, Public Relations, Writing/Editing': 4, 'Marketing, Sales': 3, 'Administrative': 3}



['employment_type', 'Full-time', 'Contract', 'Temporary', 'Internship']
{'employment_type': 49, 'Full-time': 44, 'Contract': 3, 'Temporary': 2, 'Internship': 1}



['industries', 'Financial Services', 'Marketing and Advertising, Public Relations and Communications, Management Consulting', 'Nonprofit Organization Management, Higher Educat
 'Real Estate']
{'industries': 49, 'Financial Services': 4, 'Marketing and Advertising, Public Relations and Communications, Management Consulting': 2, 'Nonprofit Organization Management, Hig
 'Consumer Goods': 2, 'Real Estate': 1}
```

My views:

- Major problem is LinkedIn doesn't support scraping much. You can still add your authentication details and go into developer mode and scrape some data slowly.
- Make sure you don't make many requests in a short time. If that happens and error occurs, increase slow_mo value in program.
- The program contains different queries, 1st one is general query without filters.
- The 2nd query uses no filter and gives 50 jobs across US in any company and any job. The results.csv is currently based on running this query. *Currently, since the code is in production mode, I have kept the limit very low, here 50, But based on need, we should increase the limit in order to get more accurate insights from the results.csv while using Query 2.*
- The remaining queries uses different filters for different job roles.
- The data is being achieved well and the filters are working. Mostly the roles are software/technical jobs in general query.
- I observed in the developer tools that sometimes the request is getting blocked. In this case, wait and try after sometime.

It is possible to customize queries with the following filters:

- RELEVANCE:
    - `RELEVANT`
    - `RECENT`
- TIME:
    - `DAY`
    - `WEEK`
    - `MONTH`
    - `ANY`
- TYPE:
    - `FULL_TIME`
    - `PART_TIME`
    - `TEMPORARY`
    - `CONTRACT`
- EXPERIENCE LEVEL:
    - `INTERNSHIP`
    - `ENTRY_LEVEL`

- `ASSOCIATE`
- `MID_SENIOR`
- `DIRECTOR`

Proposed solution:

- Create a linkedin account for key learning. Automate the search to take top jobs for each country and search those jobs at the rate of one hour per 190 countries as we need the results to update only once in a week.

Other General Problem:

- Privacy issues for LinkedIn. Refer link: prohibition of scraping software