Working Prototype Known Problems Report

PathFinder, PathFinders, May 26, 2021

List of functions not working correctly

- **Scraping**
    - **Scraper.__init__():**
        -
    - **Scraper.scrape_linkedin():**
        - **Very rarely, a page will have an error when the scraper tries to interact with the "show more" button on the job description. This does not cause the operation of the program to stop. The user will be prompted to check liggs in the command line, this refers to 'formatted.txt' if the -o option is selected when the script is run. This is caused by the button existing in the HTML, but**
        - **There are rare circumstances under which the error handling for scraping a job description in Scraper.scrape_job() will fail, ie. exhaust it's number of retries. This could happen for a few reasons:**
            - **The public IP of the system the program is run on is not whitelisted on the proxy service's API.**
            - **The webdriver is failing (ie. not loading the proper page, or loading an error page)for an unknown reason.**
            - **Multiple proxies IPs that are banned are rotated into each other in sequence, leading to the number of retries not properly avoiding said bans.**
            - **The user's internet is failing.**

          Note that this error is fairly uncommon.
- **Scraper.get_page(url):**
    - **ERR_TUNNEL_CONNECTION_FAILED is a webdriver failure that happens rarely when a URL is accessed with Selenium's get_page() function. There is code in place to handle this error, however we can not cause it intentionally, and it does not occur at a frequency that allows us to test the error handling code. If this error does occur, and handling does not catch it, Selenium's WebDriverException will be thrown.**
- **Crawler.scrape_jobs():**
    - **The very first page the web scraper scrapes will have 0 jobs scraped. Every subsequent scrape will perform as normal. This error has been investigated thoroughly, but no root cause has been found. This bug has a limited, if not negligible, effect on the functionality of the product.**

- If the script has ended in the middle of execution either due to an error or user interruption, while data is scraped from job posts, but before it has been sent to the database, on subsequent runs of the program, during the first 0 page scrape, said data will be processed and sent to the database. This is unintended behavior, but does not affect functionality. In fact, this proves to be a useful feature to make sure all scraped data is formatted and sent to the database.

- Database
    - It is possible for a situation to arise where all values given are null and basically a useless insertion is made into the database.
- Data formatting
    - DataFormatter.extract_tech_terms():
        - There are terms in the keywords text file that are potentially buggy, and will be taken out of context and added to the database. When found, we add these keywords to a list of blacklisted terms, but some do exist. A good example would be the term ‘User’ showing up in a small portion of job posts as a skill.
    - DataFormatter.extract_yoe():
        - Due to the formatting of a job post, incorrect years of experience might be extracted. Any job post with an unrealistic number for Years of Experience are rejected.
- Web app
    - When displaying fewer than 20 results in the chart, there will be a significantly darker line from the center of the chart to its top.
        - This could potentially be fixed by redoing the chart transition code to properly deal with changing amounts of data rather than filling the chart with empty data so that the chart would always have 20 data entries.