Overview

Name: flag_abandon.py

Purpose: This tool processes an Excel file containing GitHub repository information to compute and add new columns that indicate whether each repository is "abandoned” (TRUE/FALSE). Abandoned is defined as no commits within a user-specified number of days and is calculated based on the "Last Commit" **(LC)** timestamp column for each repository.

Requirements

1. **Python 3.x.** and **pandas** are required to run this script. However, if you have installed the Data Scraping Tool **(DST)** successfully, no other package installation is needed.
2. If you are running the DST successfully, you will need to run this script in that same environment so it can access the packages.

Accessing the Tool

Please clone the following repo into your filesystem.
https://github.com/annika-14/Predicting-Abandonment
This contains updated DST scripts with error handling and LC timestamp data collection. It also contains the flag_abandon script.

Usage

```
python3 [path_to_flag_abandon.py] [path_to_excel_file]
```

1. The input file must be a valid Excel file (it doesn’t have to have an .xlsx extension)
2. The Excel file inputted must contain a column named Last Commit containing integer Unix timestamps for each repository's last commit.
    - Ideally, the file used is an output Excel sheet of the DST which automatically calculates this info (if you use the version found in the above repo)
3. All file paths must be valid and accessible from the current working directory

4. This command must be run in an environment that has all necessary packages

User Input

After running the command, you will be prompted to enter the beginning day, end day, and step value. The script will display the range of days and ask for confirmation. If confirmed, the script will proceed. If not, you can re-enter the values until they are correct.

The days inputted must be a valid range (i.e., both positive integers with the end day greater than or equal to the start day). The tool will flag each repo as abandoned or not abandoned based on how long it has been since the LC timestamp, compared to each day inputted.

An example run is below, with complete execution as well. User input is represented in **bold** while all else is the script execution.

***

```
you@your-computer-prompt main % python3 ./flag_abandon.py input.xlsx

Enter the beginning day (positive integer): 30
Enter the end day (positive integer, equal or larger to beginning day): 120
Enter the step value (positive integer greater or equal to 1): 30
The steps from 30 to 120 by 30 are [30, 60, 90, 120]
Are these values okay? (y/n): y
30
60
90
120
Process data saved to input.xlsx
```

***

Results

1. The script will modify the Excel file to add new columns for each specified day. It will be named “Abandoned within _ days” where each day is a new column.
2. Each new column will indicate whether the repository has been abandoned (no commits) within the specified number of days with TRUE for abandoned and FALSE for not abandoned
3. To view the results, open the original Excel file used as input and it will have been updated with the additional columns.
4. If the column value is 30 days, for example, it will calculate whether or not the timestamp of the LC was within 30 days ago or not.
5. It measures numbers of days ago from the moment the script is run each time.