COMP 370 Homework 6 – Data Collection

Assigned Oct 3, 2024
Due Oct 11, 2024 @ 11:59 PM

Conceptual Exercises

1. What is refactoring? Give three examples of refactoring techniques.
2. In a data science project, why does code naturally go through "phases" of messiness?
3. What are three techniques for creating more modular code?

Technical Exercises

In this homework, we'll work on doing some data collection using web APIs, along the lines of what we saw in class. We'll be using the News API available at newsapi.org.

Let's consider that we're working on a project called "newscover" where we're looking at the coverage that different topics get in the news.

Task 1: Setup access to News API

You'll need to create a developer account with newsapi.org (it won't cost anything). This will give you an API key that you can use to call the API. This will give you 100 API calls per day – this should be plenty for this assignment, though you should be economical and get started on the assignment early – just in case you run out of API calls on any given day.

Task 2: Build your newsapi utility

In your project, create a top-level python package called newscover. Inside that create the newsapi.py module. In this, create a function:

```
fetch_latest_news(api_key, news_keywords, lookback_days=10)
```

which queries the NewsAPI and returns a python list of english news articles (represented as dictionaries) containing those news keywords and published within the last <lookback_days>.

Test 3: Write unit tests

Write three unit tests (i.e., one TestCase class, two functions) that test your NewsAPI module. Put them in the newscover.tests.newsapi module.

- One test should ensure that fetch_latest_news fails when no news_keywords are provided.
- Another test should ensure that when lookback_days is set, it doesn't produce articles outside that timeframe
- The last test should ensure that fetch_latest_news fails when a keyword contains a non-alphabetic character.

Note that your api key should NOT be written into your test. It should be loaded from a file "test_secrets.json" that also sits in the newscover.tests package directory (you decide the format of this file – but make it self-documenting)

Task 4: Write a data collection tool

Write a CLI tool sitting in the newscover.collector module that has the following behavior

```
python -m newscover.collector -k <api_key> [-b <# days to lookback>] -i <input_file> -o <output_dir>
```

The input file is a json file containing a dictionary of named keyword lists. Like this

```
{ "trump_fiasco": ["trump", "trial"], "swift": ["taylor", "swift", "movie"] ]
```

For each keyword set with name N and keyword list X, the collector will execute a query for the keywords X and write the results to the <output_dir>/N.json.

Task 5: Removing redundant code

Rewrite the code below to remove the redundancy.

```
# get all the sales data by product type
book_sales_2022 = load("data/book_sales_2022.csv")
book_sales_2023 = load("data/book_sales_2023.csv")
book_sales_2024 = load("data/book_sales_2024.csv")

game_sales_2022 = load("data/game_sales_2022.csv")
game_sales_2023 = load("data/game_sales_2023.csv")
game_sales_2024 = load("data/game_sales_2024.csv")

# calculate the total sales for each year
total_sales_2022 = sum_sales(book_sales_2022, game_sales_2022)
total_sales_2023 = sum_sales(book_sales_2023, game_sales_2023)
total_sales_2024 = sum_sales(book_sales_2024, game_sales_2024)
```

Submission Instructions

- A file conceptual.md containing the answers to the conceptual exercises.
- Your newscover package directory and all *.py files
- lean5.py – the revised code for Task 5