TDI PROJECT FOR DATA ENGINEERING WEEK 3

API Data Collection, File Formats, and GitHub Collaboration

Introduction

Welcome to Week 3! This week, we will focus on collecting data from APIs using the `requests` library, saving the data into different file formats, and collaborating using GitHub. We will also explore version control strategies such as branching and team collaboration.

PART A: Video Links

Here are some helpful links that will introduce or reinforce the concepts you'll be applying this week:

Requests Library

File Types in DE

Args and Kwargs

Functions in Python

Pandas and Files

argparse Library

click Library

Logging in Python

Branching in Git

PART B: Assignment Tasks and Directions

This week's assignment will have you use the [SofaScore API](https://sofascore.p.rapidapi.com) to collect real-time data, store it in different file formats (Parquet and Excel), and collaborate on GitHub using branching strategies.

Task 1: GitHub Collaboration and Branching Strategy

Objective:
You'll implement Git for version control and collaborate on GitHub using branching strategies.

Steps:
- Create a dev branch for development, a staging branch for testing, and a prod branch for production-ready code. How does branching help you manage different stages of your codebase?
- Push your code to GitHub via the dev branch.

Task 2: Invite a Team Member to Collaborate

Objective:
You'll invite a peer to collaborate with you on this project and work on different branches.

Steps:
- Add a collaborator to your GitHub repository. What are the steps to give someone access to your repo?
- How do you and your collaborator each work on different branches without conflicts?
- The invited collaborator should test code in dev and accept a PR to staging and finally the inviter should test staging before accepting a PR to prod

Task 3: API Data Collection with `requests`

Objective:
You will be fetching live data from the SofaScore API and working with the `requests` library to collect it.

Steps:
- Create a Python function to fetch data from the SofaScore API using the `requests` library.
- What code would you write to make a GET request and authenticate with the API?
- How do you handle response codes and errors when fetching data from an API? Ensure your function has error handling for failed requests or incorrect API usage.

Example Task:
Create a Python function that fetches live sports data using the `requests` library. Ensure the data is stored in a structured format such as a list of dictionaries or a pandas DataFrame.

Task 4: Saving Data to Parquet and Excel Files

Objective:
You will now save the fetched data into Parquet and Excel files.

Steps:
- Write two functions to save the data from Task 1 into both `.parquet` and `.xlsx` file formats.
- What's the difference between Parquet and Excel file formats in terms of efficiency and data engineering?
- Why might Parquet be more suitable for large datasets compared to Excel?

Example Task:
Create two Python functions, one for saving the API data in a Parquet file using `pandas.to_parquet()`, and another to save it in an Excel file using `pandas.to_excel()`.

Task 5: API Data Updates

Objective:
The SofaScore API provides live, continuously updating data. You will set up a function to regularly fetch and update your dataset.

Steps:
- Write a function that fetches new data at regular intervals and appends it to your existing dataset.
- How can you automate regular data fetches using Python? Could you use something like `time.sleep()` or a scheduling library?

Example Task:
Create a loop that fetches data every hour and appends it to a Parquet file.

PART C - Submission Instructions

You are expected to submit your assignment by the end of the week. Submission will be done via Google Forms. You are encouraged to put your work online to help build your portfolio and show your learning

For Twitter Submission:
- Tag the official TDI page: @TDataImmersed
- Tag Annie @DabereNnamani
- Tag the project coordinator @The_Jonathaan
- Tag @python
- Tag @JOloganj
- Use the hashtag TDI

For LinkedIn Submission:
- Tag the TDI page: @TheDataImmersed
- Tag Annie @AnneNnamani
- Tag @josephologunja

For this assignment, you will be required to submit your work via GitHub.

PART D - Correction Class

Correction classes will be held every Saturday from 4 pm to 6 pm Nigerian Time on the TDI official Discord or a Google Meet link will be shared before the class.

Good luck with your assignment! We hope you find it both challenging and rewarding. If you have any questions, feel free to reach out to the mentors or the community on the TDI platform.