Codeforces Problem Scraper

Assignment 1

AI chatbot for CP

December 21, 2024

Overview

In this assignment, you will build a web scraper to extract content from Codeforces. The project follows a systematic approach:

1. **Analysis Phase:**
   Analyze how Codeforces structures web pages so that you know exactly where to look for the problem statements, test cases, and editorials. Implement the core scraping function in Python using BeautifulSoup and Selenium.

2. **Core Implementation:**
   It will scrape problem statements along with metadata information like tags and time/memory constraints, preserve mathematical formulas, and format code pieces appropriately. You'll further extend this to handle editorial content, ensuring proper formatting of explanations and solution code.

3. **Data Organization:**
   Your scraper will store problems and editorials in a structured format. Problem statements are saved as text files and metadata in JSON format. The implementation includes proper error handling and rate limiting to respect Codeforce's servers.

We have provided a clear project structure and milestones to guide you through building this tool and we hope it serves it purpose.

The Libraries

Free feel to use others but during our implementation, these were enough.

```
from bs4 import BeautifulSoup
import requests
from selenium.webdriver import Chrome
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.chrome.service import Service
import os
import json
```

Recommended Steps

Setup

1. Create a project directory structure.
2. Create configuration for data storage paths.

Problem Scraper Implementation

1. Implement problem data extraction:
   - Problem title, statement, tags
2. Implement data storage:
   - Store problem statement in text format
   - Store metadata in JSON format

Editorial Scraper Implementation

1. Implement editorial extraction:
   - Editorial content identification
   - Code block handling & LaTeXpreservation
2. Implement content processing:
   - Handle section headers

Submission Guidelines

1. Fork the reference repository (`https://github.com/YogitShankar/ChatbotGDG/tree/main`)
2. Implement your changes and submit your implementation as a pull request
3. Include:
   - Python file for the scrapper
   - Simple documentation supporting it in markdown format
   - Sample scraped data

Reference Implementation

The base implementation of this will be uploaded at after the deadline: `https://github.com/YogitShankar/ChatbotGDG/tree/main`

Notes

Remember to respect Codeforces' terms of service and implement appropriate delays between requests to avoid overwhelming their servers. Good luck with your implementation!