NextHikes

Overview

Business Need

You work at Nexthikes as a Data Scientist and A software company has given you a project on collecting articles from Python. You have to scrap the website data and check the word frequency in that article.

Your job is to do web scraping and word frequency.

Web Scraping, also known as web harvesting or web data extraction, is a type of data scraping used to gather information from websites. It allows you to retrieve information from a website that you can't access through an API or other means. Python provides several libraries that make web scraping easier.

One of the most popular libraries for web scraping in Python is Beautiful Soup1. It is used to extract information from HTML and XML files. Beautiful Soup creates a parse tree from the page source code, which can be used to navigate, search, or modify the data.

Another essential Library is Requests: It is used to make HTTP requests to a specific URL and retrieve the response. Requests provide built-in functionalities for managing both the request and response.

Ultimate Guide to Web Scraping with Python Part 1: Requests and BeautifulSoup – LearnDataSciI

Finding the Frequency of words in the test or string is important to know the weightage of words in the articles. Calculate the frequency of each word in the given string - GeeksforGeeks

Tasks:

Task 1: Web Scraping: Use libraries like requests and BeautifulSoup to scrape data from a website [Welcome to Python. org]-week 1 & week 2

And save data in a text file.

Task 2: Word Frequency Create a program that reads a created text file from task 1 and counts the frequency of each word.   -week 3

Learning Outcomes

Technical Skills:

1. Checking the Word Count and Word Frequency [Core Python],
2. Web Scraping library [requests and BeautifulSoup]

Badges

Each week, one user will be awarded one of the badges below for the best performance in the category below.

In addition to being the badge holder for that badge, each badge winner will get +20 points to the overall score.

This approach aims to support and reward expertise in different parts of the Core Python and Basic Data Science.

There will also be a mark that will be added to the most innovative approach.

Interim Submissions

To be categorized into the different weeks till 2nd week

- Your employer wants a quick meeting after you've done a first quick pass of the data and wants to know whether further investigation is useful. To achieve this, summarize your findings from Task 1 and Task 2
- Link to your GitHub code that includes your Python scripts.

Feedback

You may not receive detailed comments on your interim submission but will receive a grade.

Final Submission **(20 days)**

- Python Scripts are to be saved in the Project 1 folder.
- Link to your GitHub code that should include your required code.

Feedback

You will receive comments/feedback in addition to a grade.