Cloud Computing | Word Frequency Analysis using Docker Assignment

Objective

The objective of this assignment is to familiarize students with text processing, stop word removal, and word frequency analysis using Python. Additionally, students will gain experience in creating a Docker image and running a container to perform the task.

Instructions

1. **Dataset**

Using the following dataset from Kaggle: [Random paragraphs | Kaggle].

Instruct them to download the dataset and save it as a text file named "random_paragraphs.txt".

2. **Python Script:**

Write a Python script to perform the following tasks:

- Read the contents of the "random_paragraphs.txt" file.
- Remove stop words from the text using a suitable library or manually created stop word list.
- Count the frequency of each word in the processed text.
- Display the word frequency count to the console.

Note: Students can use libraries like NLTK or spaCy for stop word removal.

3. **Dockerfile:**

Create a Dockerfile to package your Python script and dependencies into a Docker image. The Dockerfile should include the necessary instructions to build the image.

Submission Criteria

Follow these steps to submit your assignment:

- Create a public GitHub repository.
- Upload the Dockerfile, the "random_paragraphs.txt" file, and the Python script to the repository.
- Ensure that the repository is **public** and **accessible**.
- Obtain the repository link.

Submission Form

Fill out the following form: [Assignment Submission Form].

You must provide your name, ID, and the link to your GitHub repository in the form.

The submission deadline is Wednesday 24th, 11:59 PM.

Grading Criteria

- Proper implementation of the Python script to read the text file, remove stop words, and count word frequency.
- Correct usage of a Dockerfile to create an image and run a container.
- Successful submission of the Dockerfile, "random_paragraphs.txt" file, and Python script to the GitHub repository.
- Submission of the GitHub repository link through the provided form.