Assignment

Url checker

Overview

This assignment is designed to trigger your creativity and to see how you organize your code. There is no right or wrong answer; we are looking for the optimal solution when it comes to performance, modularity and ease of use.

Your task consists in writing a python script that takes as primary argument a text file containing a list or urls (see attached), tries to get the webpage of each site and counts the number of javascript HTML tags in the page (<script = ...> </script> counts for one).

The script has the following requirements:

- It needs to have persistence on disk (csv, database, ...). We want to be able to inspect the results quickly by loading the data in pandas for instance. That way we can get quick statistics like the average number of HTML tags across all pages.
- It should be easy to stop and resume the execution in the middle of the script. This also comes back to the previous idea of the persistence on disk.
- The script needs to log any errors and why it failed getting the webpage. Networks and web servers are sometimes unreliable therefore we need an easy way of retrying the url that failed.

Based on the previous information, please provide a small script to tackle this issue. All the code in one python file should do unless you feel like creating a package.

While the list of urls attached is short we have to imagine this script could be used to process thousands of urls. Anything you can implement to make it as fast as possible would be great.

Good luck!