Pageview Analyzer

Assignment 10 | CIS 206

1. Complete one or more of the following tutorials:
    - RegexOne: Learn Regular Expressions
    - Regular-Expressions: Tutorial
    - Ryan's Tutorials: Regular Expressions Tutorial
    - RexEgg: Regex Tutorial

2. Review Wikitech: Analytics/Archive/Data/Pagecounts-all-sites. Use Wikimedia Dumps to download two files of hourly log data for all Wikimedia projects. The filename of each log file is in the format `pageviews-yyyymmdd-hh0000`. The format for each log file is:

   ```
   domain_code page_title count_views total_response_size
   ```

   Wikiversity's domain_code is en.v. Sample data files are available at Sample Data 1 and Sample Data 2. Write a program that reads all pageview log files in the current directory and uses RegEx groups to parse the data. For en.v records only, create a dictionary using page_title as the key and the sum of count_views as the value.

3. After reading all files and summing count_views, display the top 100 pages and corresponding count_views sorted in descending order by count_views, and alphabetically in the case of a tie.

4. The format for a page_title is `Title[/Subpage...]`. The title without subpages may be considered the overall learning project. Iterate over the dictionary and use RegEx to separate titles from subpages. Create a separate dictionary with a key for for each learning project and the sum of the page and its subpage count_views as the value. Display the top 100 learning projects and corresponding count_views sorted in descending order by count_views, and alphabetically in the case of a tie.

5. For each of the above, use separate functions for each type of processing. Reuse functions where possible, such as in sorting and searching. Avoid using global variables by passing parameters and returning results. Include appropriate data validation and parameter validation. Add program and function documentation, consistent with the documentation standards for your selected programming language.