Cairo University
Faculty of Computers and Artificial Intelligence
Managing and Modelling Big Data (2022/2023)

Assignment 2

Dataset – Wikimedia Project:

The Wikimedia Foundation supports hundreds of thousands of people around the world in creating thelargest free knowledge projects in history. The work of volunteers helps millions of people around the globe discover information, contribute knowledge, and share it with others no matter their bandwidth.

In this task you are going to explore the page views of Wikimedia projects. Download the page view statistics generated between 0-1 am on Jan 1, 2016 from here.

Each line, delimited by a white space, contains the statistics for one Wikimedia page. The schema looks as follows:

Field	**Meaning**
Project code	The project identifier for each page.
Page title	A string containing the title of the page.
Page hits	Number of requests on the specific hour.
Page size	Size of the page

Develop spark application in any programming language that implements the below functions ***once using map-reduce paradigm in spark and once using spark loops*** and compare their performance in terms of time.
You must also create a document includes all the results of each query:

1) Compute the min, max, and average page size.
2) Determine the number of page titles that start with the article “The”. How many of those page titles are not part of the English project (Pages that are part of the English project have “en” as first field)?
3) Determine the number of unique terms appearing in the page titles. Note that in page titles, terms are delimited by “_” instead of a white space. You can use any number of normalization steps (e.g. lowercasing, removal of non-alphanumeric characters).
4) Extract each title and the number of times it was repeated.
5) Combine between data of pages with the same title and save each pair of pages data in order to display them.

Cairo University
Faculty of Computers and Artificial Intelligence
Managing and Modelling Big Data (2022/2023)

Important Notes:

- This is a group assignment of 4 members (at most) and the members should be from the same group/lab.
- All team members should work and fully understand everything in the assignment even if you distributed the questions, you should understand your colleague's questions.
- The due date is on Saturday, 20th of May. No late submission is allowed. No submission through e-mails.
- Do not share your code with anyone, so that no other student would take your files and submit it under their names.
- Any cheating will be graded ZERO for both teams.
- Each team should discuss the assignment with his/her lab TA. Any team member who misses attending the discussion will take ZERO.