Cairo University
Faculty of Computers and Artificial Intelligence
Information Systems Department

Managing and Modelling Big Data [2021-2022]

Spark Task

Dataset – Wikimedia Project

The Wikimedia Foundation supports hundreds of thousands of people around the world in creating the largest 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 0-1am 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 a function for each of the queries that prints requested values. You must also create a document includes all of the results of each query.

1) Retrieve the first k records .(ex : first 10 records)
2) Compute the min, max, and average page size
3) 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)?
4) 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).
5) Determine the most frequently occurring page title in this dataset.

Cairo University
Faculty of Computers and Artificial Intelligence
Information Systems Department

Important Notes:

- This Task will be done in teams.
- Each team should be either 3 or 4 students.
- Task Delivery Deadline: **28 May 2022**
- Task grade is based on performance in the discussion.