Exercises with Spark RDDs

These are a set of exercises to practice with Spark RDDs. Some of them use data from the Wikimedia projects pagecounts dataset. You can access the files from this dataset from this url: https://dumps.wikimedia.org/other/pagecounts-raw/ or use the data provided in the attached folder.

The dataset is organized in many files. For quick tests you can use a single file, to make execution faster. Here is a quick link to a sample file:

https://dumps.wikimedia.org/other/pagecounts-raw/2010/2010-08/pagecounts-20100806-030000.gz

It contains data from 6th August, 2010, from 3:00 to 3:59. While doing these exercises, place the dataset files you downloaded in a local or HDFS directory that is accessible from your Spark Context. Most of these exercises can be executed in the Spark Shell. It is recommended, however, that you create an auxiliary file to save your code, and to be used as a Spark script. I recommend you use jupyter for now.

Write Spark code, in python, to perform the following tasks:

1. Load the Wikimedia dataset into a Spark RDD called pagecounts. Print the first 10 lines of the RDD to make sure the data is accessed correctly. Try to make the output easy to read.
2. Count the number of lines in pagecounts and store it in a variable.
3. Create a new RDD called enPages. This RDD should contain only the lines of pagecounts that refer to the Wikipedia in English project. Check the dataset description web page (above) to learn about the file structure (Hint: The project code for the Wikipedia in English is “en”).
4. Although the input data is organized in rows and columns, the enPages RDD does not have a corresponding structure. Program a transformation for this RDD that splits the file lines into separated fields, and store those in tuples. Make sure that tuple fields that are numbers are of class Long. Name the resulting RDD enPagesTuples.
5. Order pages in enPagesTuples by descending page size. Show the top 5 largest pages.
6. Locate what is the most visited page of Wikipedia in English in your dataset and print its name and number of visits.
7. Calculate the necessary data to create an histogram of page visits. The histogram should have 20 bins (BONUS: Put the histogram code into a function that receives the input RDD and number of bins as parameters. The function should not return the bin sizes, but instead print the histogram with ascii characters).