

Homework 3

MSDS 694, Diane Woodbridge

Description

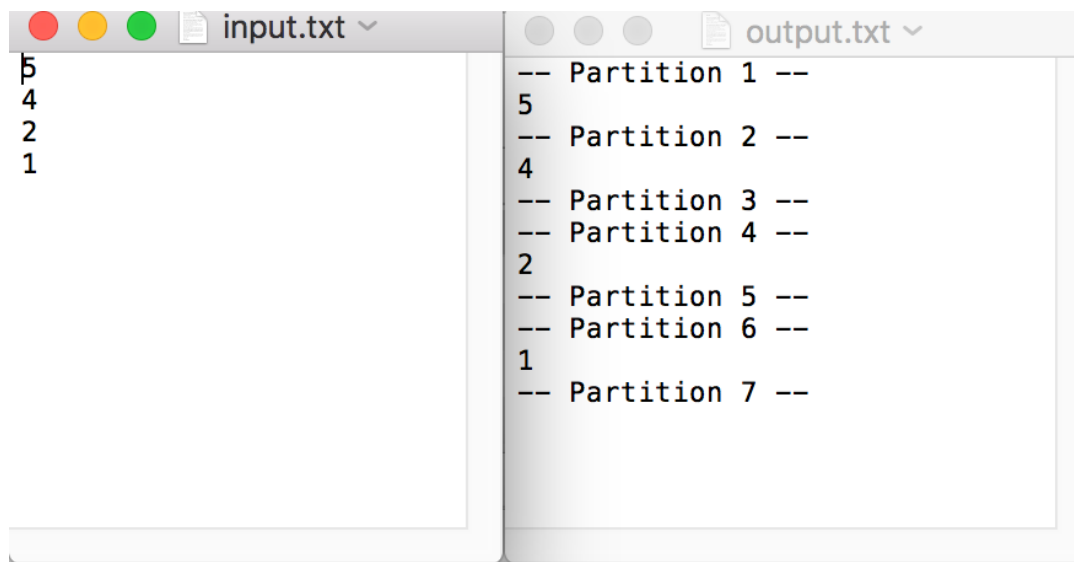
Given the *input_file* and *app_name* in “user_definition.py”, complete the pyspark script “hw3.py” to:

1. Load data into *partition_ct*, the number of partitions to load the input file.
2. Print the partition number (“-- Partition N --”) and data in each respective partition in separate lines.
Please see the output.txt for the detailed formatting.
Once you run diff output.txt input_n/output.txt, it should not return any messages(4pt).
3. Follows pep 8 standards – Should not have any errors when running pycodestyle (2pt).

Submit the hw3.py file (**ONLY**) - the name of your file should be hw3_LastName_Firstname.py on Canvas. Make sure it runs in **Python 3.6 and pyspark 2.3**.

We provide two example input files (input_1/input.txt, input_2/USF_Mission.txt) and corresponding output.txt. **Please do not hardcode input_file, output_file and partition_ct (They are given in the user_definition.py).**

If you run spark-submit hw3_Woodbridge_Diane.py, the output should be **similar to the following. (The number of lines and format should be the same.)**



The screenshot shows a code editor with two files open: input.txt and output.txt. The input.txt file contains the following text:

```
5
4
2
1
```

The output.txt file contains the following text:

```
-- Partition 1 --
5
-- Partition 2 --
4
-- Partition 3 --
-- Partition 4 --
2
-- Partition 5 --
-- Partition 6 --
1
-- Partition 7 --
```