Spark[HW0]

Word count

- The typical “Hello, world!” app for Spark applications is known as word count. The map/reduce model is particularly well suited to applications like counting words in a document.
  Write a program based on these steps:
  1. read the ‘news.txt’ file and show the read result by collect operation. (10 pts)
  2. split each line into its words and show the first 2 lines results by take operation. (10 pts)
  3. write a flatmap function that returns (word, 1) for each word. Explain why we used flatmap but not map function. (20 pts)
  4. reduce step 3 results by the key which in this case is the word and return (word, #count). (30 pts)
  5. sort final result in descending order and save it in ‘wordCount.txt’. (30 pts)

Data

- Associated data file is news.txt

What to submit

- You should submit a zip file containing these files:
  1. step 1 and 2 screenshots.
  2. Step 3 explanation.
  3. Your code(.py or .jar or ...) file.
  4. WordCount.txt from step 5.

name your zip file like this:

spark[HW0]your name.zip

for example:

spark[HW0]reza hashemi.zip

SCREENSHOTS

- Your screenshots should be like sc1.png and sc2.png.(the whole terminal area is captured)

Aban 1400

Submit method

upload your zip file on dedicated section of course on https://elearn.ut.ac.ir

Deadline

1400-08-13 23:59

2021-11-04 23:59

Contact

shahab.ghodsi@ut.ac.ir

Tips

- You can use ‘[^\w]+’ regex to split a line into words, but it’s not the most accurate approach. Try find something better, your effort is appreciated by extra pts.

Attention

- Don’t share your homework with your classmates.
- Similar homeworks will be detected.