Big Data Infrastructures 2018
Spark Homework

- Remember to first test your code by running Spark using the master local[*] option.
- Store the output of each task in HDFS, under your directory.

Task 1) Rewrite the word counter (object HelloSpark in the SampleCode.zip archive found on ILIAS) by using the reduceByKey function. Hint: reduce on a collection of pairs(<word>, 1).

Task 2) Write a hash-tag counter. That is, a program that takes as input the tweets data set "tweet_sample_raw_data.txt" found on ILIAS, and returns as output a list of pairs (hashtag, n.occurrences), sorted by number of occurrences, descending.

Task 3) Write a program that uses the same tweets data set "tweet_sample_raw_data.txt" on ILIAS and filters it by tweets from "San Francisco" and "Chicago, IL" cities, and returns their tweet text without stop words. You can find a list of English stop words at https://99webtools.com/blog/list-of-english-stop-words/

Deliverables:

Send your code to rana@exascale.info
Deadline: 1st of February 2019

Good luck!