# Super Computing for Big Data Assignment 1

Bo Wang
4479599
B.Wang-
6@student.tudelft.nl

## 1. OVERALL

This assignment contains two components, one for spark basics and one for spark streaming on twitter. Since spark was developed, it has become the most active big data projects and top-level apache project. From this assignment, I learned how to set up Spark environment, the basic operations of Spark and Spark Streaming.

As a Data Science students in grade one, I used to read this article: *"Three Reasons a Data Engineer Should Learn Scala"*[1]. This class also help me to know, to use and to be familiar with scala, wichi is really interesting.

## 2. PREPARATION

From the previous lab sessions of ET4310, we have learned the basic of Hadoop, mapreduce and data science ecosystem, which help us built a basic understing from the whole.For setup the Spark environment, we need to download spark and scala. As I'ms a Mac User, I chosse homebrew finished the task.

As I used to be familiar with Python, I wrote the first exercise with Python through PySpark. Then turned to Scala on the second exercise.

The whole assignment contains three files: report.pdf, exercise 1.md and exercise 2.md. I use Markdown recorded what I did and the markdown files is avaliable on my github page[2] which seems to be more prettier.

## 3. EXERCISE

### 3.1 Exercise 1

From exercise 1, I learned how to set up Spark environment, create a RDD(Resilient Distributed Dataset), take data from RDD , use loop to print the output, count the dataset, and filter dataset accoding some conditions. I used Python and PySpark in this exercise.

For more concreate process, please check exercise 1.md from the github link blow.

### 3.2 Exercise 2

From exercise 2, firstly, I learned how to use sbt build scala project. Secondly, I learned to create twitter credential to call twitter API. After registe an application on the website, we can get an apikey, apiScreat, accessToken, accessTokenSecret. With these keys we can connect to twitter web service. Thirdly, I learned the ways to handle streaming data from twitter, that is hashtag and retweet, and how to sort these data. I used Scala in this exercise.

For more concreate process, please check exercise 1.md from the github link blow.

## 4. CONCLUSION

Compared with Hadoop Mapreduce, Spark is more efficient, flexible and easier to use, particular the case in realtime data processing, such as twitter, facebook and linkedin. In the big data era, spark will play an increasingly important role in big data and cloud computing context.

For it's good performance, integration with big data ecosystem and functional paradigm, Scala is one of the best tool for big data processing.

[1] https://www.hakkalabs.co/articles/three-reasons-data-eng-learn-scala

[2] https://github.com/bwanglzu/SuperComputing4BigData/tree/master/assignment1

.