Data Science with Spark

Master 'Data Science'

X – ECE

Lab 3

Stream Processing and Messaging using Apache SPARK

2018/2019

Salim NAHLE

Organization:

- You can work on any Spark environment
- Only **PDF reports** are expected. The report shall contain the code of your application, explanations and necessary screenshots of the output.
- Please work in **pairs**! Each group (composed of 2 persons at most) shall submit one report. Do not forget to indicate your names in the report.
- The report shall be uploaded on the campus page before **Thursday 15/11/2018 midnight.**
- Late reports will be penalized (3 points/day)
- Reports shall be uploaded on the course's page on Moodle.

Lab Description:

Objective:

Create a Spark Streaming Application that plots out the popularity of hashtags on Twitter. This is done by streaming (live) tweets from Twitter. Then you need to extract the hashtags and display the 10 most popular (most frequent) among them.

3 figures are expected:

1. Figure 1 displays the 10 most popular hashtags in the last hour
2. Figure 2 displays the 10 most popular hashtags in the last 24 hours
3. Figure 3 displays the 10 most popular hashtags in the last week

Twitter Developer Account:

You need to create a Twitter Developer account to be able to stream the tweets.

https://developer.twitter.com/

Expected tools:

Using Spark Streaming (RDD-based or DataFrame-based) is mandatory.

You can also use other Big Data tools like Kafka, Flume, etc. for developing the application.

You can use Python matplotlib for plotting the tags' popularity.